WESTFÄLISCHE
WILHELMS-UNIVERSITÄT
MÜNSTER

High-Quality Web Information Provisioning and Quality-Based Data Pricing

Florian Stahl

approach market
cost model relation case relevant use
context knowledge general analysis criteria versioning
well based goods different curation marketplace
system
information instance
defined developed attributes provided
individual described price documents
curator first
Nevertheless given figure
prices possible using provide overall number pay regarding comparison database seen thesis contrast
example make providers Furthermore pricing order quality search
following Curation LandSAR problem means
factors content presented likely source
online result score used item marketplaces referred
models process Internet query according green user whether
part function available high-quality customers
domain introduced work relational important
similar services form needs actual
products sources point good
utility group WiPo
considered cases commonly fact tourism
allows product way Petri technical
focus provider discussed service
data
new business
scale research
value time set
applied Web
engines

WESTFÄLISCHE
WILHELMS-UNIVERSITÄT
MÜNSTER

DBIS Group
Databases &
Information Systems

High-Quality Web Information Provisioning and Quality-Based Data Pricing

Inauguraldissertation zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften durch die
Wirtschaftswissenschaftliche Fakultät der
Westfälischen Wilhelms-Universität Münster

Vorgelegt von

Florian Stahl, MSc
aus Gießen

Juni 2015

Dekanin	Prof. Dr. Theresia Theurl
Erster Gutachter	Prof. Dr. Gottfried Vossen
Zweiter Gutachter	Prof. Dr. Dr. h.c. Dr. h.c. Jörg Becker, Prof. h.c. (NRU - HSE, Moskau)
Mündliche Prüfung	07.07.2015

Florian Stahl

High-Quality Web Information Provisioning and Quality-Based Data Pricing

MV WISSENSCHAFT

WESTFÄLISCHE
WILHELMS-UNIVERSITÄT
MÜNSTER

Wissenschaftliche Schriften der WWU Münster

Reihe IV

Band 9

ulb
Universitäts- und
Landesbibliothek Münster

Florian Stahl

High-Quality Web Information Provisioning and Quality-Based Data Pricing

MV WISSENSCHAFT

Wissenschaftliche Schriften der WWU Münster
herausgegeben von der Universitäts- und Landesbibliothek Münster
http://www.ulb.uni-muenster.de

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Dieses Buch steht gleichzeitig in einer elektronischen Version über den Publikations- und Archivierungsserver der WWU Münster zur Verfügung.
http://www.ulb.uni-muenster.de/wissenschaftliche-schriften

Florian Stahl
„High-Quality Web Information Provisioning and Quality-Based Data Pricing“
Wissenschaftliche Schriften der WWU Münster, Reihe IV, Band 9

ISBN 978-3-8405-0133-3 (Druckausgabe)
URN urn:nbn:de:hbz:6-67249643318 (elektronische Version)

direkt zur Online-Version:

Satz / Titelbild: Florian Stahl
Umschlag: MV-Verlag
Druck und Bindung: MV-Verlag

Geleitwort

Im Zeitalter von Big Data, dem IT-Schlagwort der letzten Jahre, kommt dem angemessenen Umgang mit Daten und daraus gewinnbarer Information immer höhere Bedeutung zu. Dies ist zwar im Grunde so seit der „Erfindung“ und Verbreitung des Web als Dienst, der auf dem Internet basiert; allerdings wurden in den letzten Jahren durch die Digitalisierung von fast allem und jedem und die Anbindung von nahezu allen denkbaren Artefakten an das Web hier immer neue Dimensionen erreicht. Als Privatperson hat man sich an den daraus resultierenden „Komfort“ gewöhnt und ist sich gleichzeitig der zunehmenden Orwellschen Überwachung und Durchleuchtung bewusst. Als Unternehmen ist man daran interessiert, Daten unterschiedlichster Quellen in einen gemeinsamen Kontext zu bringen, um dann aus deren Analyse neue Dienst- oder Produktangebote zu schaffen und den Kontakt zum Kunden (weiter) zu individualisieren. Die aus dieser Situation resultierenden Forschungsfragen sind von enormer Breite und nahezu täglich werden neue gestellt, aber es gibt natürlich große Unterschiede in deren Relevanz. Florian Stahl greift in seiner Dissertation zwei solche Fragen heraus, die auf den ersten Blick vielleicht wenig miteinander zu tun haben, die sich aber auf den zweiten nicht nur als fundamental, sondern sogar als folgerichtig zusammenhängend herausstellen. Bei der ersten geht es um die Frage der anwendungs- bzw. situationsbezogenen Bereitstellung qualitativ hochwertiger Information, und zwar nicht (nur) durch umfangreiches Suchen, eventuell über mehrere Suchmaschinen hinweg, sondern durch einen Prozess, in welchem eine Kuratierung von Daten bzw. Suchergebnissen eine zentrale Rolle spielt. Idealerweise ist das Ergebnis eines solchen Prozesses „transportabel“ bzw. mobil in dem Sinne, dass man es mit sich herumtragen kann und dass es daher sogar offline funktioniert, denn hierfür gibt es nach wie vor zahlreiche Anwendungen. Die hier gegebene Antwort lautet WiPo bzw. „Web in your Pocket“, welches ausführlich vorgestellt wird. Bei der zweiten Frage geht es um den Handel mit Daten: Wenn man schon Daten hochqualitativ aufbereitet, dann kann man sich das Ergebnis eigentlich auch vergüten lassen. Diese Überlegung führt auf eine aktuelle Entwicklung – Datenmarktplätze – und die Frage, wie man auf einem solchen Marktplatz Preisgestaltung betreibt und diese idealerweise sogar für Käufer und Verkäufer fair gestaltet. Grundsätzlich werden auf einem Marktplatz Daten unterschiedlichster Qualität vertrieben, aber intuitiv bestimmt sich der Preis über die (vom Verkäufer gebotene) Qualität sowie über den (für den Käufer vorhandenen) Wert. Florian Stahl fokussiert seine Betrachtungen auf strukturierte Daten und ebnet damit den Weg für einen Einsatz relationaler Datenbankkonzepte; die Idee ist, versionierbare Sichten (einer relationalen Datenbank bzw. einer daraus errechneten Universalrelation) anzubieten, deren Qualität je nach Preisvorstellung des Kunden nach oben oder unten skaliert wird; der Verkäuferpreis bleibt dabei jeweils verborgen, bis der Kunde den Zuschlag erhält. Neu an seiner Vorgehensweise ist die Tatsache, dass er eine ganze Reihe von Kriterien in die Berechnung von Qualität einfließen lässt, die sich individuell gewichten lassen: Manche Kunden werden an vielen Daten interessiert sein, andere an möglichst vollständigen, wieder andere an gut dokumentierten usw.; hat ein Kunde dann die für ihn relevante Gewichtung festgelegt, kann daraus ein Gesamt-Qualitäts-Score errechnet werden. Neben der Kundenperspektive wird auch die Anbieterseite betrachtet, und es wird gezeigt, dass das Problem der Ermittlung eines für Anbieter und Kunden fairen Preises für Daten bestimmter Qualität als ein Multiple-Choice-Knapsack-Problem (MCKP) aufgefasst werden kann. Damit lassen sich approximative, für das Pricing adaptierte Algorithmen formulieren, die so beschaffen sind, dass Datenqualität sich (erwartungsgemäß) umgekehrt proportional zu algorithmischer Laufzeit verhält: Je mehr Qualität ich wünsche, desto höher wird der Berechnungsaufwand. In der Arbeit von Florian Stahl wird ein umfangreiches Instrumentarium aus BWL-, Electronic Business- und Wirtschaftsinformatik-Konzepten mit Modellen und Techniken der Informatik in geschickter und eindrucksvoller Weise kombiniert. Dies zeigt sich zum einen am WiPo-Konzept mit seinen diversen Einsatzszenarien und seiner prototypischen Realisierung und zum anderen an der Erkenntnis, faire Marktplatz-Preisgestaltung für strukturierte Daten in der Terminologie relationaler Datenbanken ausdrücken und die Problemlösung auf eine Version des Knapsack-Problems zurückführen zu können. Die Arbeit stellt damit eine für einen an hoch qualitativer Informationsverarbeitung interessierten Leser äußerst interessante Lektüre dar, und ich wünsche ihr eine breite Leserschaft.

Münster, im Juli 2015 Prof. Dr. Gottfried Vossen

Acknowledgements

First and foremost, I would like to thank my doctoral adviser Prof. Dr Gottfried Vossen. I very much appreciate his support and guidance throughout my research and the process of writing this thesis. Moreover, I am grateful for the numerous opportunities he gave me to present my research at national and international conferences as well as to visit New Zealand twice to work on the Web in your Pocket (WiPo) project. I would also like to thank Prof. Dr Dr h.c. Dr h.c. Jörg Becker, Prof. h.c. (NRU - HSE, Moscow) for his willingness to act as a second reviewer of this thesis, and Prof. Dr Thorsten Wiesel for acting as a third examiner during the defence of this thesis.

In academia, best results can be achieved when working collaboratively. Therefore, I would like to acknowledge my co-authors who have always supported me in my endeavours and helped me develop new ideas. In addition to Gottfried Vossen, I thank Assoc. Prof. Dr Stuart Dillon and Dr Karyn Rastrick with whom I worked on the idea of WiPo. Also, there are a number of students who supported this work by implementing the WiPo architecture, most notably Adrian Godde, Bastian Hagedorn, Bastian Köpcke, and Martin Rehberger. Prof. Dr Alexander Löser and Alexander Muschalle from Berlin, I thank for their collaboration in the initial phase of my work on data marketplaces. Also, I thank Fabian Schomm and Lara Vomfell for their joint work on the data marketplace surveys.

During my time at DBIS group, Dr Gunnar Thies, Dr Till Haselmann, and Dr Christian Forster set good examples in completing a doctorate and supported me during the time of writing my thesis. Particularly, Dr Till Haselmann supported me morally, TEX-nically, and provided invaluable feedback regarding my work.

Preantepenultimately, I thank my colleagues at DBIS group: Dr Jens Lechtenbörger, Iman Kamehkhosh, David Fekete, Nicolas Pflanzl, and Fabian Schomm, who were always happy to discuss my research and gave me useful feedback regarding my preliminary results. For commendable technical support, I thank Ralf Farke, and for appreciable administrative support, I thank Claudia Werkmeister and her successor Melanie Wietkamp.

Antepenultimately, my dear friends Paul and Simon shall be acknowledged for having done a jolly good job in ensuring that what I have written is actually English. Besides, Paul has provided invaluable feedback regarding the content of my work.

Penultimately, I would like to thank my family for their infinite support and encouragement throughout the entire time of my PhD studies and the same holds true for my girlfriend's family. Ultimately, my deep-felt and most sincere thanks I would like to express to my girlfriend Anna-Lena who – besides proofreading the first draft of this work – has shown an exemplary patience with me during the final stage of this thesis and given me all the love I could have wished for.

Münster, July 2015 Florian Stahl

Contents

List of Figures

List of Tables

List of Algorithms

1. Introduction

Information is one of the key elements of the Internet and has been described as the fuel of an information economy [HC02]. It has gained so much importance that *information* can nowadays be considered a third production factor besides *capital* and *labour* [Nor11; RK96; Mac62]. This increased used of Information Technology (IT) is accompanied by a socio-cultural shift, leading to the dawn of the so-called information age [Nor11], in which economy and society are heavily influenced by automated information processing. As both society and economy are currently in transition, this shift will eventually result in an information economy [Wei00], the first mentions of which date back to the 1960s [Mac62]. This goes along with the observation that after the commoditisation of hardware and software, it is now data[1] – often used synonymously with information [PLW02] – that is valuable [MO06]. The shift towards an information economy is often attributed to the technological innovations and organisational changes the Internet and the Web have brought along [Cas10]. In contrast to previous changes in economy, such as the industrial revolution, the currently ongoing information revolution does not only affect firms and their way of production but also the way goods are traded on markets [Wei00].

While on the one hand information and data become more valuable to companies, the amount of data available online, on the other hand, is growing at an enormous pace [HT03; LT06; Ram11; Eco10; Koo13]. This has led to a point at which the Web can be considered the largest collection of data and information ever created [BR10]. Therefore, in contrast to many other resources, the main challenge in handling information is not its scarcity but the plethora of information available [SV12]; and not finding the right information may be costly for companies [FS01]. Two main issues arise from that: firstly, it is a major challenge to find the most relevant information for a given problem and, secondly, the quality of this data is often questionable which leads to the problems of finding and purchasing high-quality data. Next, these two challenges will be de-

[1] In this work, the term data, similarly to the mass noun information, will be treated as a singular whenever data is referred to as a concept. However, if specifically the plural of datum is referred to – mainly in the more technical chapters –, the plural forms will be used. This is in line with contemporary usage of the term [Oxfedb].

scribed in-depth. Consequently, the aim of this thesis is derived in Section 1.1. Thereafter, Section 1.2 presents an outline of this thesis.

1.1. Aim of this Thesis

As information has been discovered as a productive resource, the quality thereof is of higher importance than ever before [OLC11; NR00]. Accordingly, this is a matter of great interest to many research communities. This is in particular true for the business and information systems research community, e. g., [TWRB14; WLPS98; OLC11; WS96], but also the database research community in which the topic is inherently rooted, e. g., [NR00; Nau02; Tan14; Wei99; BS06; Bas90; JV97]. However, no precise definition of the term *information quality* exists. Most commonly, it is defined as *fitness for use*, which has an inherent user focus [WS96], i. e., the eventual quality of a given piece of information is rated by the ability of users to utilise it.

Furthermore, there is an ambiguity regarding the terms *information quality* and *data quality*. It will become evident later in this work that *data* (raw numbers and strings) builds the basis of *information*, which can be defined as *data with a meaning* or as *processed data* [PLW02]. Therefore, it is only logical that data quality can be considered to be concerned with more technical aspects, whereas information quality is more concerned with non-technical aspects [MWLZ09]. Nevertheless, given the interrelation between data and information, the quality of both can be viewed as two aspects of the same concept. Thus, many authors refer to both when using the nomenclature data quality, or use the terms interchangeably, e. g., [MWLZ09; PLW02]. While this is reasonable in many cases because of the inherent interconnection, for the purpose of this work, it is important to keep a clear separation of the two terms and according quality concepts.

The first part of this thesis will investigate high-quality information provisioning, where quality refers first and foremost to the fact that information is relevant and correct in a given scenario. In this regard, it can be argued that finding information online has tremendously improved over the last decades, starting with directories and reaching its current height with algorithmic search engines – the de-facto standard today. However, it has been argued that for numerous cases the currently existing search technology is not sufficient [SV12; DSV12; SV13; Cer10; Dop09]. In particular for niche domains, such as rare medical conditions, it can be very difficult to find high-quality information. This is even more complicated if a number of sources are to be integrated, as in the following query: “name all European universities that offer an information systems degree in cities larger than 40,000 inhabitants and within 50 km to an airport”.

Throughout, the history of mankind, high-quality information has been connected with humans who selected and organised it. This task of choosing relevant information and organising it in digital repositories is now known as *digital curation*. Given the human focus in the definition of data information and quality, it is self-evident that only humans can decide what is really relevant to them, thus, finding high-quality information in the plethora of information available has long been the task of information specialists [Bacoo]. While average users can commonly be satisfied with average search results provided by standard search engines, this is not acceptable in any niche domain – such as the treatment for rare medical conditions or the planning and conduction of search and rescue operations.

In the light of this, it can be argued that high-quality information provisioning through a standardised interface as a single point of truth is a vastly unsolved problem. Thus, the first part of this thesis will focus on:

> *Develop a software artefact that answers a user’s domain-specific informational queries levering curation to satisfy their high-quality information needs.*

In this context, the design science paradigm by Peffers et al. [PTRC07] will be followed.

Addressing the same issue, namely that high-quality information and data can be the proverbial needle in a haystack, data providers and data marketplaces have emerged. In contrast to the provisioning of information just described, where a context is always provided, the data sold by these vendors is commonly not tailored to a specific use case. Thus, the term data is more appropriate than information. As a consequence, in this latter case data has to be processed by the customer in order for it to become meaningful. Common providers of data include statistical offices, financial data providers, or weather data providers.

Recently, data providers have started selling their data but also data-related products such as analysis using so-called *data marketplaces*. Data marketplaces act as intermediaries between providers and users (customers) of data. Whereas providers benefit as they reach a larger audience, customers have the advantage of dealing with a single data marketplace rather than numerous providers. Additionally, many data marketplaces also create data by crawling the Web and providing analysis of this data.

The topic of trading data has become of so much interest that it even resulted in an art project, in which one could pay with personal FACEBOOK[2] data in a supermarket [Nor14]. Although this art project was focused entirely on personal data, it illustrates well the wide recognition of data as a tradable good and its inherent value.

Nevertheless, even though there is undoubtedly a market for data and data-related services as well as the recognition that data has a value [Mil12b; BTF11; TKRB11], there is little understanding of where this value stems from [BHS11; Mil12b]. Similar to the observation that data quality can best be gauged by the eventual consumer, it can be argued that its value is different to various people [SV99; SF08]. The combination of both subjective *quality* attribution and subjective *value* attribution make the matter even more complicated. Given this complex structure, it is not surprising that until now, little guidance on how to price data goods has been provided. Thus, the research aim of the second part is:

> *Provide a fair pricing scheme to be utilised by data providers on data marketplaces that allows for quality-based versioning and according price discrimination for custom-tailored relational data goods.*

It is obvious that both research areas complement each other. Having differentiated between data quality and information quality as two aspects of the same concept; this concept could be coined *quality of an information good.* The first research aim addresses satisfying high-quality information needs tailored to specific niche domains. In contrast, the second research aim addresses data quality as a means of fair pricing of data.

1.2. Structure of this Thesis

This thesis is structured in seven chapters. Following this introduction, which has outlined the setting as well as the aim of this thesis, Chapter 2 presents the formal basics needed to fully understand the remainder of this work. Next, Chapter 3 lays the foundation for the main Web information provisioning part of this work, namely the Web in your Pocket (WiPo) approach, which is extensively discussed in Chapter 4. Subsequently, the second main part of this work, namely data marketplaces and data pricing, is introduced in Chapter 5, before Chapter 6 presents the data pricing in considerable detail. Finally, this work is concluded

[2] http://www.facebook.com, accessed: 2015-05-31.

in Chapter 7 which summarises the major aspects of this work. This structure is illustrated in Figure 1.1. Subsequently, the contents of each chapter will be outlined in more detail.

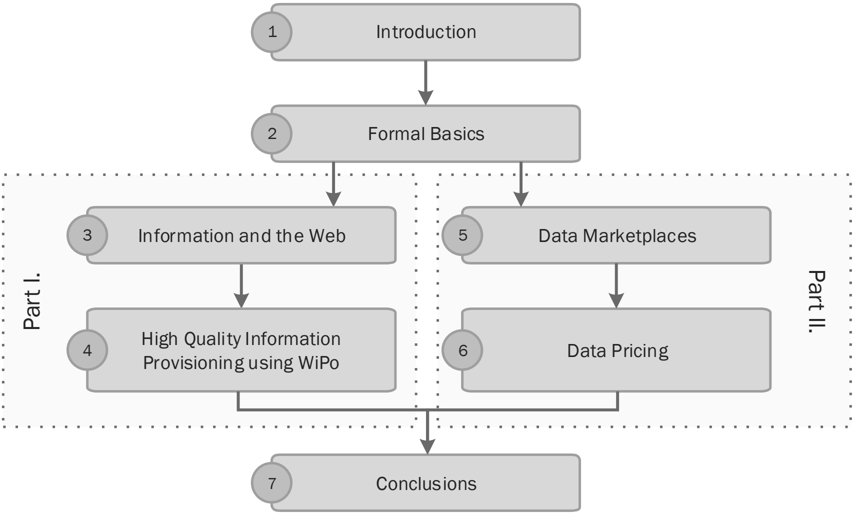


Figure 1.1.: Structure of this Thesis.

In order to express the findings of this work, two formal modelling techniques will be introduced in Chapter 2. In the context of information provisioning through WiPo, Petri Nets will be used as a means of modelling; they are introduced and described in Section 2.1. Furthermore, to formally define and work with data marketplaces, the relational data model and relational algebra will be needed. Consequently, they are introduced in Section 2.2.

Starting with Chapter 3, the information provisioning part of this work begins. The chapter demonstrates briefly how mankind has always striven for advances in information gathering and processing. This is followed by a definition of information in the context of this work in Section 3.1. Then, the developments in information provisioning over the last quarter of a century – the age of the World Wide Web (WWW) – are outlined in Section 3.2. Section 3.3 presents current challenges in Web search. Thereafter, the role of humans in the process of high-quality information delivery is outlined in Section 3.4. On this basis, limitations of current search technology are highlighted and the first aim of this thesis, namely the development of a curation-based (i. e., manually supported) search process is outlined. Design science as a method is justified in Section 3.5.

In this way, Chapter 3 paves the way for Chapter 4, the first main chapter of this work, which outlines the WiPo approach (Section 4.1) and demonstrates how the concept has been implemented by presenting the architecture of a prototypic implementation (Section 4.2). The subsequent Section 4.3 discusses four sample scenarios in which WiPo can be beneficial by developing and applying a comparison model. The chapter concludes with an outline of potential extensions to WiPo and future research directions.

Similar to the information provisioning part, the data marketplaces part of this thesis follows a two-step approach. First, Chapter 5 outlines aspects of value creation of data and data marketplaces by recapitulating some basic microeconomics in Section 5.1. Thereafter, the actual topic of data marketplaces is approached in a theoretical manner, defining data marketplaces according to the relevant economic literature in Section 5.2. Next, data marketplaces are approached from a practical point of view in Section 5.3, including a description gained from prior interview studies. Subsequently, pricing of information goods is discussed in Section 5.4.

Having laid the foundations for a more in-depth discussion of data marketplaces and data pricing, Chapter 6 begins with a review of previous work on pricing of data on data marketplaces in Section 6.1. Section 6.2 then establishes the focus of this part to be quality-based pricing. Consequently, data quality will be extensively discussed and a quality scoring model for data marketplaces will be developed in Section 6.4. Subsequently, a quality-based pricing model will be derived based on the Multiple-Choice Knapsack Problem in Section 6.5. Both previously mentioned sections contain an extensive example to ease the understanding of the elaborations. Eventually Section 6.6 concludes the data marketplaces part of this work.

Finally, Chapter 7 concludes this thesis by summarising the main contributions in Section 7.1, and providing an outlook on future developments in Section 7.2.

2. Formal Basics

In this work two main formal modelling techniques will be used. In the context of information provisioning through WiPo, Petri Nets will be used as a means of modelling, which will be explained in Section 2.1. To this end, firstly, the basic concept behind Petri Nets is elaborated on. Subsequently, this basic model is extended by more advanced data models, more advanced arcs, and the introduction of an inclusive *or* concept in Sections 2.1.1, 2.1.2, and 2.1.3, respectively.

Additionally, to formally define and work with data marketplaces, relational algebra will be needed and shall, thus, be explained in Section 2.2. More precisely, Section 2.2.1 introduces the relational model and defines some basic concepts such as relations and databases. Subsequently, operations that can be done on a relation or a database, using relational algebra, are introduced in Section 2.2.2. Finally, Section 2.2.3 illustrates the relational algebra by means of examples.

2.1. Petri Nets as a Means of Modelling Processes

Petri Nets were originally developed by Petri [Pet62] as a means to describe communication with automata. Since then, Petri Nets have been developed further and applied to different scenarios, including process modelling in general and more particular business process modelling [AH04; SVOK11; Obe96; AS11; LO01; LO03; Aal98].

This section will briefly introduce Petri Nets as a means of modelling processes. It builds on a number of sources which shall be briefly mentioned beforehand. A good introduction to Petri Nets in general can be found for instance in [Rei10; SVOK11] or more comprehensively, even if slightly dated, in the two volumes [RR98a; RR98b]. Regarding business process modelling with Petri Nets, [AS11] and [SVOK11] are representative for two main streams. While van der Aalst and Stahl [AS11] use coloured Petri Nets, Schönthaler et al. [SVOK11] favour Extensible Markup Language (XML) Nets.

Both are extensions of the original Petri Nets, here described according to [RE98]. Petri Nets are a strictly alternating sequence of *places* (depicted as circles), *transitions* (originally depicted as connection between to places, now mostly depicted as rectangles), and *arcs* connecting the former two. The simplest form of this type of net is depicted in Figure 2.1. It has to be pointed out that arcs are commonly directed, resulting in two categories of places: one serving as input to a transition and the other serving as an output of a transition.

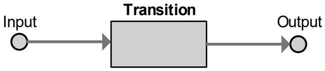


Figure 2.1.: Simplest Form of a Petri Net.

Furthermore, it should be noted that more than one place can serve as an input or an output, respectively. However, it does not have to be the same number of inputs and outputs but can be different numbers; an example of this is given in Figure 2.2.

Initially, places could hold exactly one *token* (depicted as black dots in a place). Over time however, this basic notation was extended and places were allowed to hold more than one token in so-called Place/Transition Systems, explained for instance in [DR98].

In order for a transition to be activated, commonly referred to as *fire*, all places going into a transition have to hold at least one token. Then, the transition removes a token from each inbound place and produces a token in each outbound place. This basic concept is depicted in Figure 2.2. At this point, it should be clarified that Petri Nets are commonly displayed without initial tokens. If they do contain tokens, they are referred to as systems rather than nets [Len03; Obe96]. During this work the term net will be used throughout for simplicity. Furthermore, the focus of this work will be on the overall process flow rather than on the dynamics of the systems.

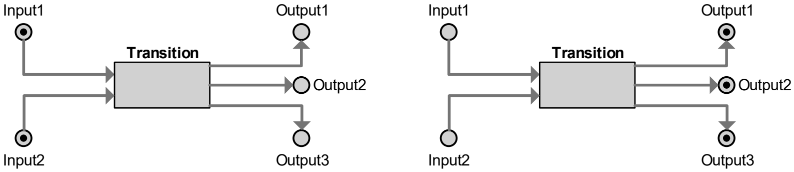
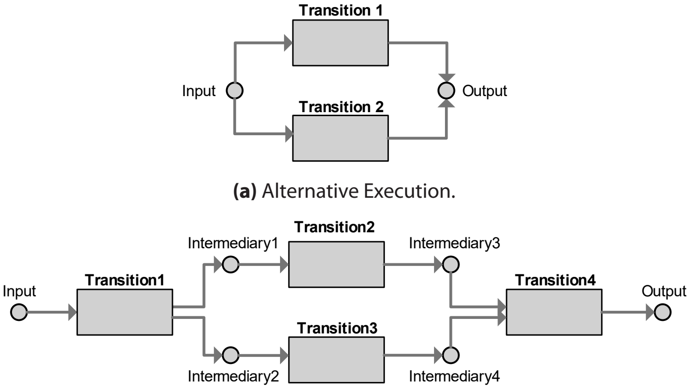


Figure 2.2.: Petri Net Before (left) and After (right) the Transition Fired.

The fact that a transition can have more than one input and output place allows for Petri Nets to branch. Depending on whether the branching follows a place or a transition, different situations are possible. The easiest case is a sequential order of transitions and places as shown in Figure 2.1. In contrast, Figure 2.3 shows branching of a Petri Net with alternative execution (Transition 1 and 2 (a)) and parallel execution (Transitions 2 and 3 (b)).



(a) Alternative Execution.

(b) Parallel Execution.

Figure 2.3.: Branched Petri Net with Alternative Execution (a) and Parallel Execution (b).

Since branching can lead to rather complex net structures, the concept of hierarchical substitution, explained for instance in [SVOK11; AS11], will be used in this work. Basically, this notation allows it to refine a transition in its own sub-net. This allows for the development of a clear and concise super-net with a number of refinements. Here, activities that can be refined will be referred to as abstraction.

While basically using the same concept, both approaches [SVOK11] and [AS11] differ in their graphical representation. While van der Aalst and Stahl [AS11] annotate an abstraction with the letters HS (for hierarchical substitution) to indicate that a particular transition has a refinement, Schönthaler et al. [SVOK11] prefer a stack of transitions as representation. In this work, the stack representation will be used to keep textual annotations to a minimum in order to increase readability. Furthermore, it should be pointed out that all places preceding or following the abstraction in question are also to be present in the refinement. Figure 2.4 shows the graphical representation of a hierarchical substitution applied here.

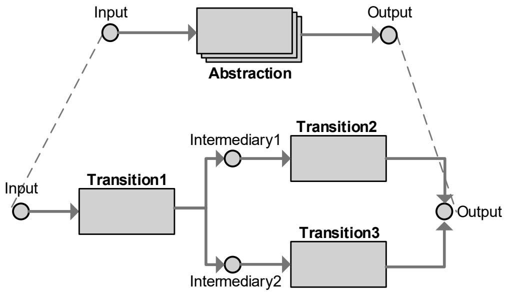


Figure 2.4.: Hierarchical Substitution Depicted as in [SVOK11].

2.1.1. Advanced Data Models and Petri Nets

In order to differentiate between different types of tokens more advanced data models were introduced to Petri Nets. One manifestation of this is the Coloured Petri Net, employing coloured tokens which were introduced (according to [GL81; GL79]) by Lautenbach. The term colour is misleading, as it is used synonymously to value [AS11]. A good overview of this concept is given by Jensen [Jen89; Jen98] or by van der Aalst and Stahl [AS11] who applied Coloured Petri Nets to workflow modelling. In this model colour is used to describe the manifestation of a number of attributes.

A similar, yet more structured, approach with regard to object structure is Predicate/ Transition nets [GL81; GL79], where places model properties of, or relations between, individuals. This can be interpreted as a relation schema [Len03]. Nevertheless, Jensen [Jen89] clarifies that both, Coloured Petri Nets and Predicate/Transition nets, are rather dialects of the same language.

Jaeschke and Schek [JS82] stated that strict adherence to first-normal-form-relations is not always possible when modelling business objects as this does neither allow for multiple values in a column nor for nesting. As a solution, they proposed an algebra for Non-First-Normal-Form (NF^2) relations. Similarly, Oberweis [Obe96] found that when Petri Nets are to be used for workflow modelling, Predicate/Transition net are not sufficient and developed NF^2-Relation/Transition nets which also support hierarchical data structures while keeping the advantages of Predicate/Transition Nets.

Eventually, this stream of Petri Net research evolved into Petri Nets using XML for data modelling. These were introduced by Lenz and Oberweis [LO01;

LO03], comprehensively defined and described in [Len03], and also used by Schönthaler et al. [SVOK11]. In this type of net, places hold XML-documents that follow a common XML-schema. Transitions are connected to places and operate on objects contained within. As in basic Petri Nets documents following the in-coming schema are consumed (or read which is in contrast to classical Petri Nets). Then, transitions do modify the input and produce output documents which must adhere to the outgoing schema(s). Arcs are directed to indicate whether an object is consumed/read or created.

Given their strength and their ability to model documents right into the process, XML-based nets are the ideal choice for the description of the WiPo process. In particular, filters, which are introduced in the next subsection, simplify modelling the WiPo process. Nevertheless, anticipating Chapter 4, the WiPo prototype was developed using JavaScript Object Notation (JSON)[1] data objects, because of JSON's flexibility and small overheads. JSON is a lightweight data-interchange format targeted at both humans and computers based on a subset of the JavaScript language [jsoed]. It is now defined in its own standard [ecm13] by Ecma International[2]. Compared to XML it has fewer overheads and has been referred to as a "fat-free alternative to XML" in [Cro06]. Given the fact that research has been conducted into how to translate one into the other [Wan11; Lee11] and in this seemingly unpublished article [NF13] even on schema level, the assumption that insights from XML-based Petri Nets can be applied to JSON-based nets is not far-fetched. That said, it is not the aim of this thesis to develop a new type of Petri Net; however, data modelling is supposed to be carried out using JSON, as it has been used in the prototypical implementation. All this being said, XML-nets are by no means the only way of modelling the WiPo process.

2.1.2. Advanced Arcs in Petri Nets

Whilst original Petri Nets had only directed arcs between places and transitions, here, the approach of Schönthaler et al. [SVOK11] will be followed, who present three types of arcs; the original directed arc, arcs without direction known as reading arcs, and bi-directional arcs referred to as updating arcs.

Reading arcs, also known as test arcs, were first introduced for Coloured Petri Nets [CH93]. Basically, this arc grants a non-exclusive reading privilege to the adjacent transition, allowing a transition to fire without removing the incoming token from its place. Originally, test arcs where depicted as a connector

[1] http://json.org/, accessed: 2015-05-31.

[2] http://www.ecma-international.org/, accessed: 2015-05-31.

with a crossbar at both ends. In XML-nets regular arc are used and the reading is triggered by specific filters attached to the arcs [Len03]. Schönthaler et al. [SVOK11] use a simple connector without an arrow or crossbar. This has the advantage that they are easily recognisable and do not over-complicate the drawings. Figure 2.5 shows the original and the simplified version of reading arcs. The latter will be used here.

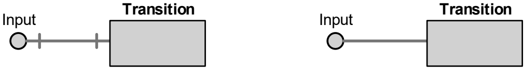


Figure 2.5.: Reading Arcs Original (left) and Simplified (right).

Seemingly, update arcs only appear in the surroundings of Horus software GmbH (HORUS)[3], providing the Petri-Net-based modelling tool HORUS. Update arcs are, for instance, discussed in [SVOK11] which describes the Horus method, an approach to modelling business processes using the HORUS tool. In HORUS, and for the purpose of this work, update arcs are a simplification of an incoming and outgoing arc to the same transition and place. Thus, an updated connection may be used to remove, modify, and replace a token from a place. This means that while being updated the token cannot be accessed by another transition. This is depicted in Figure 2.6.

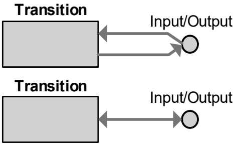


Figure 2.6.: Petri Net with Update Arc (bottom) and an Equivalent Representation Using Two Arcs (top).

2.1.3. Advanced Branching: Introducing OR to Petri Nets

In case of parallel execution, see Figure 2.3(b), splitting and joining of path is done by transitions, while in the alternative case, see Figure 2.3(a), this is achieved by connecting more than one arc to a single place. However, this has

[3] http://www.horus.biz/, accessed: 2015-05-31.

the disadvantage that transitions may compete for ingoing tokens which cannot be solved deterministically for basic Petri Nets [RE98; AS11]. Coloured Petri Nets provide so-called guards which help to determine when a transition actually fires based on the colour (attributes) of a token [AS11]. In XML-nets this is achieved by labelling arcs to indicate what properties a document has to fulfil in order to be consumed. These labels are also referred to as filters. Basically, they specify which documents to use. For instance, a transition could require objects to have a non-empty description field and only consume those. While guards and filters may reduce the risk of conflicted states, they do not per se guarantee conflict-freeness. However, given that there should be a business rule when modelling workflows, conflicts can be avoided by choosing filters adequately.

For Coloured Petri Nets and Predicate/Transition Nets joining in a place is not a problem because they simply add tokens. However, joining in a place may well be a problem in XML-nets because a transition may just modify an outgoing place. For instance, a description field in a place could be written by two transitions causing one to override the other. That said, following the argument for input-places, modelling appropriate filters and meaningful processes, the risk of conflicts can be kept to a minimum.

Originally, Petri Nets do not have a graphical representation for the concept of *OR*. As mentioned before, a transition fires when all ingoing place bear tokens and produces a token for each outgoing place. Schönthaler et al. [SVOK11] describe an *OR* concept which they treat as an exclusive *or* (*XOR*). In this work both *OR* and *XOR* will be used. Both will be depicted throughout this work as shown in Figure 2.7. It should be said that for simplicity reasons this was done for incoming and outgoing places in one figure but they do not have to occur together.

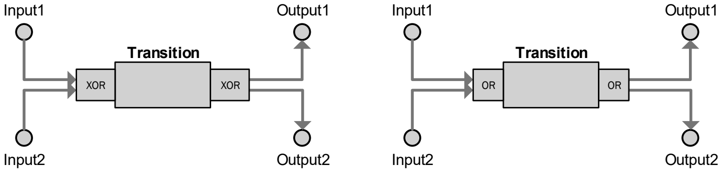


Figure 2.7.: Petri Net Using XOR (left) and OR (right).

Figure 2.8 shows how an *XOR* can be implemented in full using four auxiliary transitions. Whilst for an incoming *XOR* this can be realised without any special semantics because through the join of paths only one can be executed, it is more difficult for an outgoing *XOR*. In this case the *XOR* functionality has to be implemented using appropriate mutually exclusive filters for *Help3* and *Help4*. It is intuitively clear that depicting a Petri Net using *XOR* simplifies the graphical representation. Without it each Petri Net would be extended for each *XOR* by *n* transitions, where *n* is the number of places going into or coming out of the individual *XOR* and one place (the joining place).

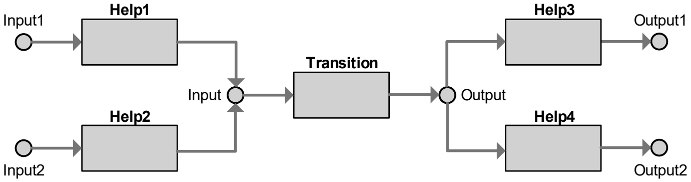


Figure 2.8.: Petri Net Using XOR in Full.

In [Keu14] which is concerned with bi-directional mapping of process models and text, it was already hinted that an inclusive *or* (*OR*) is missing in [SVOK11]. For the purpose of this work it is necessary to introduce *OR* as an alternative to *XOR*. Similar to *XOR*, it would of course be possible to model *OR* using places and transitions, but this expansion of the graphical representation is unnecessary. The simplified version is also depicted in Figure 2.7.

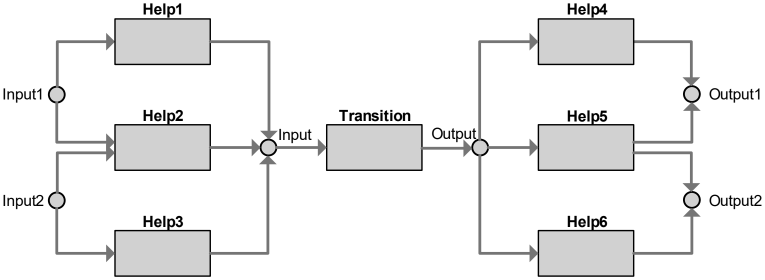


Figure 2.9.: Petri Net Using OR in Full.

Figure 2.9 shows how *OR* can be implemented in full using six auxiliary transitions. In contrast to the *XOR*, this is a little more complicated. Here, for the input side, mutually exclusive filters have to be used which determine whether only *Input1* is to be used (transition *Help1*), only *Input2* is to be used (transition *Help3*), or both (transition *Help2*). Furthermore, in the *OR* case, the outgoing case is symmetrical to the incoming case. Depicting *ORs* in full would inflate the nets even more. Similar to the *XOR* case, for every *OR* one additional place is needed. In contrast to it though, $2^n - 1$ additional transition would be needed, n again being the number of places going into or out of the individual *OR* and $2^n - 1$ being all possible combinations without the empty set.

2.2. The Relational Model and Relational Algebra

In order to understand the modelling of data marketplaces as well as pricing in this context, as discussed in Chapter 6, some basic definitions need to be clarified. This work investigates data marketplaces that provide structured data in tables, more precisely in relations, according to the relational model of data introduced by Codd [Cod70]. Furthermore, in this thesis, the basic query language relational algebra will be used to express queries to a data marketplace. The following elaborations are based on [Vos08] but can similarly be found in [Cod90; AHV95; LL99].

2.2.1. The Relational Model

Data in a tabular format with given column names or attributes can be described as a relation. A relation r has n unique attributes $A_i, 1 \leq i \leq n$ and each attribute A has a domain $\text{dom}(A)$ which allows differentiation between at least two values (true/false). Additionally, it is supposed that a unique null value $\perp$ exists that is part of each domain: $\perp \in \text{dom}(A_i), 1 \leq i \leq n$. For the purpose of this work, a null value means that no value for this domain is contained in the relation, regardless of whether it is unknown or does not exist. While in some cases it might be useful to differentiate different null values, it is not necessary in the context of this work. The set of all $A_i, 1 \leq i \leq n$ is denoted as: $X = \{A_1, \ldots, A_n\}$. The domain of X is defined as the union of all domains $\text{dom}(A_i), 1 \leq i \leq n$:

$$\text{dom}(X) = \bigcup_{A \in X} \text{dom}(A)$$

Next, a tuple (i. e., a row in the table) μ is defined as a combination of one value for all (f. a.) attributes $A_i \in X$ stemming from its according domain $\text{dom}(A_i)$. Formally, this is expressed as:

$$\mu : X \rightarrow \text{dom}(X) \text{ s.t. } \mu[A] \in \text{dom}(A) \text{ f. a. } A \in X$$

The set of all possible tuples over X is denoted as $Tup(X)$. As a consequence, r can also be seen as a subset of $Tup(X)$:

$$r \subseteq Tup(X)$$

The set of all possible relations r over X is denoted as $Rel(X)$. In reality, not all possible relations (i. e., $Rel(X)$) are meaningful. Consider, for instance, a relation with attributes *user-name* and *email-address*. In this case it would not make sense to allow different users (e-mail-addresses) to register the same user-name. Therefore, relations can be restricted by a number of *intra-relational* constraints σ that have to hold in order for a relation over R to be valid. The sum of all constrains over r regarding the attributes X is denoted as $\Sigma_X(r)$. An intra-relational constraint could, for instance, require certain attributes (e. g., username) to be unique.

The relational schema $R = (X, \Sigma_X)$ describes all relations r that satisfy Σ_X. While *intra-relational* constraints exist, they are not needed for the purpose of this thesis, leading to the simpler notation of a relational schema being described by its set of attributes and constrains that are not further specified:

$$R = (X, \cdot)$$

A set of k relational schemas $\boldsymbol{R}$ is defined as:

$$\boldsymbol{R} = \{R_1, \ldots, R_k\} \text{ with } R_j = (X_j, \cdot) \text{ for } 1 \leq j \leq k$$

The according relation instances r_j can be considered to form a database d, which is defined as:

$$d = \{r_1, \ldots, r_k\}, r_j \in Rel(X_j) \text{ for } 1 \leq j \leq k$$

Similar to relations, databases follow a database schema. This contains all relational schemas ($\boldsymbol{R}$) as well as the sum of *inter-relational* constraints ($\Sigma_{\boldsymbol{R}}$) that have to hold in order for the database to be valid. An example of *inter-relational* constrains is foreign key constrains, which impose restrictions on attributes that occur as key (identifying attribute) in one relation and as a reference in another relation.

Together, this is denoted as: $D = (\boldsymbol{R}, \Sigma_{\boldsymbol{R}})$. Again, it is acknowledged that there will usually be constraints; however, this work refrains from explicating them. Hence, the following short notation will be used: $D = (\boldsymbol{R}, \cdot)$

2.2.2. Relational Algebra

Having defined relations, it is further necessary to define operations that are supported on these relations by means of relational algebra, a set-based query language. To this end, let $r \in Rel(X)$ be a relation according to the schema $R = (X, \cdot)$ and $Y \subseteq X$ with the according set of tuples $Tup(Y)$. A tuple $\mu \in r$ from which only a subset of attributes $Y \subseteq X$ is regarded, is denoted as $\mu[Y]$, which is per definition an element of $Tup(Y)$. Then, a projection of r (choosing only certain attributes) can be defined as:

$$\pi_Y(r) := \{\mu[Y] | \mu \in r\}$$

Furthermore, two types of selection (choosing only certain tuples) can be defined. Firstly, selecting tuples that satisfy an externally given condition. Let $r \in Rel(X)$ be a relation according to the schema $R = (X, \cdot)$ and $A \in X$ be an attribute. Furthermore, let $a \in \mathrm{dom}(A)$ be a value for that attribute A and let $\Theta \in \{<, \leq, >, \geq, =, \neq\}$ be a comparison operator. Then, selecting tuples by an external condition can be defined as:

$$\sigma_{A\Theta a}(r) := \{\mu \in r | \mu[A] \Theta a\}$$

Secondly, selecting tuples that satisfy an internal condition. Let $r \in Rel(X)$ be a relation according to the schema $R = (X, \cdot)$ and $A, B \in X$ be attributes. Furthermore, let $\mathrm{dom}(A) = \mathrm{dom}(B)$ and $\Theta \in \{<, \leq, >, \geq, =, \neq\}$ be a comparison operator. For both selection types, it should be mentioned that if $\Theta \in \{<, \leq, >, \geq\}$, then $\mathrm{dom}(A)$ and/or $\mathrm{dom}(B)$ have to be of at least ordinal scale. Consequently, selecting tuples by an internal condition can be defined as:

$$\sigma_{A\Theta B}(r) := \{\mu \in r | \ \mu[A] \Theta \mu[B]\}$$

Sometimes, when working with multiple relations, it is necessary to combine them. There are a number of ways to achieve this. Here, natural join and full outer join will be focused on.

Generally, joins combine two relations $r_1 \in Rel(X_1), r_2 \in Rel(X_1)$ with according schemas $R_1 = (X_1, \cdot), R_2 = (X_2, \cdot)$ based on at least one common attribute, i. e., $X_1 \cap X_2 \neq \emptyset$. As auxiliary means, let I be the intersection of X_1 and X_2, i. e., $I = X_1 \cap X_2$ and V be the union of both, i. e., $V = X_1 \cup X_2$.

In a natural join, tuples $\mu[I] \in r_1$ that do not have a matching tupel $\mu[I] \in r_2$ or vice versa are omitted, i. e., if $\mu[I] \notin r_1 \vee \mu[I] \notin r_2$ tuples are not regarded.

Based on this, the natural join is defined as:

$$r_1 \bowtie r_2 := \{\mu \in Tup(V) \mid \mu[X_1] \in r_1 \wedge \mu[X_2] \in r_2\}$$

Full outer joins, in contrast, are not restricted to tuples that have a match in the other relation. If no match can be found, tupels are extended with null values. Thus, the outer join can be defined as follows:

$$r_1 ⟗ r_2 :=$$

$$\left\{ \mu \in Tup(V) \;\middle|\; \begin{array}{l} \mu[X_1] \in r_1 \wedge \mu[X_2] \in r_2 \\ \vee\, \mu[X_1] \in r_1 \wedge \mu[I] \notin \pi_I(r_2) \wedge \mu[A] = \perp \text{ f. a. } A \in X_2 \backslash X_1 \\ \vee\, \mu[X_2] \in r_2 \wedge \mu[I] \notin \pi_I(r_1) \wedge \mu[A] = \perp \text{ f. a. } A \in X_1 \backslash X_2 \end{array} \right\}$$

2.2.3. Relational Algebra Examples

To illustrate the elaborations on the relational model and relational algebra, the two relations r_1, containing weather forecast data for airports, and r_2, containing address data of said airports, are considered. A tabular representation of these relations can be found in Table 2.1 and Table 2.2, respectively.

Table 2.1.: Relation r_1: Weather Data.

Station	AirPressure	Humidity	Temperature	Precipitation	Date
FRA	1021	52	17	0	2017-05-08
FRA	1020	43	19	0	2017-05-09
FRA	1005	40	15	41	2017-05-10
LHR	1025	69	16	17	2017-05-08
LHR	1008	82	14	85	2017-05-09
LHR	1003	70	12	70	2017-05-10

Relation r_1 follows the schema $R_1 = (X_{r_1}, \cdot)$, with $X_{r_1} = \{$*Station, AirPressure, Humidity, Temperature, Precipitation, Date*$\}$. In this example, the according domains are: $\mathbb{N}$ for AirPreasure, Humidity, Temperature and Precipitation; valid dates for Date; and valid airport codes for Station. Similarly, relation r_2 follows the schema $R_2 = (X_{r_2}, \cdot)$, with $X_{r_2} = \{$*Station, Street, HouseNo, PostCode, City, Country*$\}$. In this example, the according domains are: valid airport codes for Station; $\mathbb{N}$ for HouseNo; and respective valid address strings for Postcode, City, and Country.

Table 2.2.: Relation r_2: Address Data.

Station	Street	HouseNo	Postcode	City	Country
AMS	Evert v/d Beekstraat	202	1118 CP	Schiphol	Netherlands
FRA	Hugo Eckener Ring	⊥	60549	Frankfurt	Germany
LHR	Western Perimeter Road	⊥	TW6 2GA	London	United Kingdom

Example 1: Supposing, one is only interested in the country an airport is located in, an according projection query can be formulated as:

$$r_3 = \pi_{\{Station,\ Country\}}(r_2)$$

The resulting relation r_3 is presented in Table 2.3:

Table 2.3.: Relation r_3: Restricting r_2 to Station and Country.

Station	Country
AMS	Netherlands
FRA	Germany
LHR	United Kingdom

Example 2: Supposing, one is only interested in weather data from FRA, an according selection query can be formulated as:

$$r_4 = \sigma_{Station=\text{‘FRA’}}(r_1)$$

The resulting relation r_4 is presented in Table 2.4:

Table 2.4.: Relation r_4: Weather Data from r_1 for FRA only.

Station	AirPreasure	Humidity	Temperature	Precipitation	Date
FRA	1021	52	17	0	2017-05-08
FRA	1020	43	19	0	2017-05-09
FRA	1005	40	15	41	2017-05-10

Example 3: Supposing, one is interested in the temperature at FRA for different dates and one wants to know the exact address of the station[4], an according query can be formulated as:

$$r_5 = \pi_{\{\mathit{Temperature,\ Date,\ Station,\ Street,\ HouseNo,\ PostCode,\ City}\}} \left(\sigma_{\mathit{Station}=\texttt{FRA}}(r_1) \bowtie r_2\right)$$

The resulting relation r_5 is presented in Table 2.5:

Table 2.5.: Relation r_5: Specific Weather and Address Data for FRA Using Natural Join.

Temperature	Date	Station	Street	HouseNo	Postcode	City
17	2017-05-08	FRA	Hugo Eckener Ring	⊥	60549	Frankfurt
19	2017-05-09	FRA	Hugo Eckener Ring	⊥	60549	Frankfurt
15	2017-05-10	FRA	Hugo Eckener Ring	⊥	60549	Frankfurt

This example implicit made the assumption that only records with a match in both relations should be returned; consequently, a natural join operator has been used. For the next example, this assumption is relaxed, which means a full outer join will be used.

Example 4: Supposing, one is interested in the temperature at AMS, FRA, and LHR for the date 2017-05-08 and one wants to know the exact address of the stations[5], regardless of whether weather data is available for them, an according query can be formulated as:

$$r_6 = \pi_{\{\mathit{Temperature,\ Date,\ Station,\ Street,\ HouseNo,\ PostCode,\ City}\}} \left(\sigma_{\mathit{Date}=\text{‘2017-05-08’}}(r_1) \mathbin{⟗} r_2\right)$$

The resulting relation r_6 is presented in Table 2.6:

Table 2.6.: Relation r_6: Specific Weather and Address Data Using Full Outer Join.

Temperature	Date	Station	Street	HouseNo	Postcode	City
17	2017-05-08	FRA	Hugo Eckener Ring	⊥	60549	Frankfurt
16	2017-05-08	LHR	Western Perimeter Road	⊥	TW6 2GA	London
⊥	⊥	AMS	Evert v/d Beekstraat	202	1118 CP	Schiphol

Table 2.6 shows how the tuple containing the address data for AMS is extended by null values for Temperature and Data because no data is available.

[4] Knowing that FRA is in Germany, the attribute Country has been omitted for overview reasons.

[5] Again, the attribute Country is omitted to keep the resulting relation to a reasonable size.

Part I.

High-Quality Web Information Provisioning

3. Information and the Web

For millennia, humans have been striving for knowledge and for ways to make it persistent. Evidence suggests that the Sumerians invented the written word as a means of accounting during the fourth millennium BC. Being able to preserve knowledge was advantageous to ancient societies as it fuelled commercial and political advancement and the development of culture. Along with this went archiving of written documents (clay plates) in collections that can be considered early forms of libraries [Pot00] and also a good place for learning [Mac00]. One of the most significant ancient collections of knowledge was the famous library of Alexandria, founded around the turn of the 2nd century BC [BR11; Mur09; Mac00; Bar00]. It is noteworthy because this was among the first institutions that aimed at collecting all knowledge available to mankind [Bar00] and was the largest of its time with – depending on the source – up to 400,000 [Mur09] or 700,000 items [Bar00].

More recently, information has been discovered as production resource as described in [Nor11; RK96; Mac62] with Machlup [Mac62] being among the first to describe knowledge as a production factor. While in previous times labour and capital were considered the most important factors of production, North [Nor11] argues that we are at the dawn of the information age where information becomes a scarce resource. This, he argues, is due to three factors:

1. The change from labour or capital intense activities to knowledge intense products and services.
2. Globalisation, in particular the resulting acceleration of international learning processes.
3. The improved information and telecommunication infrastructure which increases information transparency and thus brings perfect markets a little closer.

Similarly Weiber [Wei00] postulates that at the beginning of the third millennium AD both society and economy are in transition, eventually resulting in an information economy. Earlier mentions of the information society or the information economy date back to the 1960s and 1970s. For instance, Machlup [Mac62] introduced the term *Knowledge Industry* for "firms that exclusively engage in selling information or advice." However, he also points out that there are different degrees of specialisation, i. e., some companies or industries provide information-related services without doing so exclusively. In 1969, Drucker [Dru69] declared that the world was undergoing great change with regards to both technology and the economy. Furthermore, he explicitly states that the American economy has shifted from manual to knowledge work and discusses implications of the transition towards this new knowledge society. Porat [Por77] then presented a measurement of the information economy, i. e., the information activity within the U. S. economy as well as the implications of the transition from an industry-based to a knowledge-based economy. Later, Castells [Cas10] discusses the network society and the information economy which he attributes to the technological changes the Internet and the Web have brought along.

From the literature, the long history and increasing importance of information is evident. In light of this, *information* will be defined in Section 3.1. Then, the history of the Internet and the Web as well as Web search will be recapped in Section 3.2, as networking is crucial to the emergence of the information economy [Cas10]. Furthermore, this section will highlight how search on the Internet has been – and currently is – approached to make the best use of the enormous amount of information available online. This includes a categorisation of search technology. Following this introductory section, Section 3.3 discusses three important research fields in the context of online search engines, namely 1) accessing the so-called Deep Web, 2) information extraction from Web sources, 3) ranking of search results, and 4) social search. Penultimately, data curation is introduced as means to organise data, followed by a discussion of the profession of information broker in Section 3.4. Finally, Section 3.5 presents the method of this part of this work and points out the limitations of actual search technology before the focus of this part of the work will be concluded. Some parts of this chapter have previously been published in [SV12].

3.1. Defining Information

It is important, for the remainder of this work, to have a precise understanding of what the term information actually means. This, however, is not an easy task as there is no common definition of *information* across domains and even scientists within the same domain have failed to reach a consensus, as pointed out in [Fer03; LLS10]. Nevertheless, most commonly the terms *character, data, information,* and *knowledge* are defined together in a hierarchical manner [LLS10; Nor11; RK96; DP00].

The lowest element in this hierarchy, usually, is a *character*. Hansen and Neumann [HN05] state that if more than a very simple datum (a single character) is to be persisted or transmitted, usually a number of characters – a *character string* – is needed. The authors consider a transmitted character string to be a message. Other authors use the term data for what Hansen and Neumann [HN05] describe as a character string. For instance, Rehäuser and Krcmar [RK96] follow DIN 44300 and define data as "represented by characters which are presented for processing"[1]. Similarly, in [LLS10; DP00] it is stated that (raw) data represent events but in a not human-readable format. North [Nor11] defines data as being characters following a syntax. On its own, however, data does not have an inherent importance or relevancy [DP00].

It is not until a meaning or context is added to the data – it is brought in a form meaningful to humans – that data is transformed into *information* [LLS10; Nor11; RK96]. Similar to Hansen and Neumann [HN05], who stated that a message (data) becomes *information* if it has meaning to the receiver, Davenport and Prusak [DP00] define information as being a message that informs someone or as "data that makes a difference". Consequently, it is at the discretion of the receiver of a message whether the message contains information or not [DP00], a view shared by Machlup [Mac62] who states that informing is the act of conveying *knowledge*. Knowledge is the last building block in many definition hierarchies. Continuing his somewhat circular definition based on verb forms, Machlup states that *knowing* is the state in which knowledge (information) is already at hand. Other authors emphasise that knowledge requires information but also relies on connecting information to prior experiences, expectations and contexts [LLS10; Nor11; RK96; DP00].

Other authors such as Ferber [Fer03] use the terms information and knowledge vice versa, i. e., knowledge as data with a meaning and information as knowledge applied to specific situation. However, he also points out that there are other definitions and one should always be careful in what sense the terms are used.

To clarify, consider the number (character string) 42; it is data, consisting of the digits 4 and 2 of the character set {0..9}. On its own, it does not mean anything. If supplemented by the words degrees centigrade, 42 becomes infor-

[1] German original wording: "Daten [werden] durch Zeichen repräsentiert, die zur Verarbeitung dargestellt werden."

mation – it can now be the temperature outside or a person's body temperature. Once conveyed to the receiver this information is converted into the knowledge that a person has a dangerously high fever, or that the next day is going to be warmer than the central European standard. As this example shows, prior knowledge of the human body and the European climate has to be present in order for information to become knowledge[2]. At this point, it has to be clarified that knowledge in itself can be differentiated by various types, such as implicit and explicit knowledge [LLS10]. Given that this thesis, will focus on information and data, there is no added value in discussing these knowledge types here; interested readers are referred to [LLS10; RK96; Mac62] for an overview of the topic.

In the technical domain, the information concept of SHANNON [Sha48] is widely used which is concerned with the efficient transmission of messages. Discussing information transmission over telegraph and telephone, however, he clearly states that he does not consider meaning, but only how messages are conveyed. This understanding of information, while highly relevant in its domain, is utterly different from the information definition presented above and is only mentioned here for completeness' sake. In this work, information is always assumed to have meaning as this aspect of the definition is of highest importance for this work.

BAEZA-YATES AND RIBEIRO-NETO [BR11] do not even define the term information in their fundamental work *Modern Information Retrieval: The Concepts and Technology Behind Search.* To be fair, they differentiate Information Retrieval (IR) from data retrieval and, thus, indirectly define both. In their sense data retrieval refers to retrieving data that matches a query exactly, while IR allows for deviations such as synonyms and aims at satisfying information needs.

A rather simplistic definition from RUGGE AND GLOSSBRENNER [RG97] states that information is something that someone wants to know and is willing to pay for. While this is not a scientific but a practitioner's definition, it incorporates the connection to knowledge and highlights its property of having a value. NORTH [Nor11] extends the knowledge hierarchy, introduced above, in a corporate contexts by *act, competency,* and *competitiveness.* According to him, knowledge only has value for a company if it is transformed into an ability which leads to actions being taken. Competency, the ability to take the right actions at the right time, is based among other things on experience in acting. Having unique competency in an enterprise can give it the edge over the competition

[2] Conversely, with different prior knowledge, 42 may as well be the answer to life, the universe, and everything [Ada95].

and, thus, make it more competitive. Figure 3.1, based on an original figure by North [Nor11], depicts this relationship, showing which part of the hierarchy is dealt with in what discipline. While computer science mainly focuses on the elements up to and including information, management studies usually start at the level of information. Thus, information is seen as the point of intersection between the two disciplines and this is exactly where this work is placed. It aims at building on techniques from computer science to ultimately provide high-quality information to consumers and businesses.

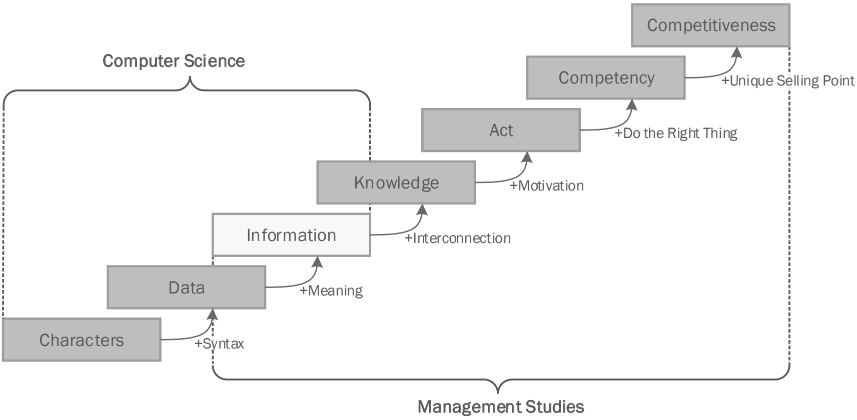


Figure 3.1.: The Hierarchy from Characters to Competitiveness According to North [Nor11], Extended with Attribution to Different Domains.

3.2. Web and Web Search

Having defined the terms *data, information,* and *knowledge*, as well as having outlined their relationship, this section elaborates on a ubiquitous data source – the Internet, which is most commonly accessed through the WWW-Service. To this end, this section first provides an overview of the history of the Internet and the Web. Then, a history of Web search will be presented. Finally, it highlights how search on the Internet has been and still is approached to make the best use of the enormous amount of information available online. This includes a categorisation of search technology.

3.2.1. A Short History of the Web

The Internet has a rather short (but lively) history of just over 50 years, which will be described here as in [KR12; W3C00]. Following the growth of the computer industry, in the 1960s, it was only logical to investigate how computers could be connected to each other in order to be used by users from remote locations. Packet switching, as a robust communication solution, was developed and in 1967 the Advanced Research Projects Agency (ARPA) proposed a framework for the ARPAnet. This went live in September 1969, connecting 4 U. S. universities (California LA, Stanford Research Institute, California Santa Barbara, Utah) by the end of that year. By 1972, ARPAnet had grown to 15 nodes and the first e-mail program was written. By the end of the decade, numerous other networks had emerged, among them the Hawaiian ALOHANet and the French Cyclades. During the course of this development, the Defense Advanced Research Projects Agency (DARPA)[3][4] funded research for the connection of networks, for which the term *Internetting* was coined. Consequently, by the end of the 1970s, the technical foundations of the Internet were laid and about 200 hosts were connected to ARPAnet. About 10 years later, the then formed Internet had around 1,000 hosts connected. During that time, various other networks were added to the Internet, most of which were from academic institutions, the services operated on these networks mainly being file transfer and e-mail.

The Internet as an information source only became widely recognised after the advent of the World Wide Web – or Web for short. This service, running on the infrastructure of the Internet, and often confused with the Internet itself, is based on a project proposal by Berners-Lee [Ber89] in March 1989. He suggested a “universal linked information system” based on the idea of *hypertext*, which allowed “a place to be found for any information or reference which one felt was important, and a way of finding it afterwards”[Ber89].

The idea of linking or joining pieces of information together goes back to Bush [Bus45], as described in his seminal work ‘As we may think’. The idea of joining documents together was picked up on by Nelson [Nel65] in his vision of the *Evolutionary List File*, in which content is stored only once, but is included in many lists by means of links. In this work, Nelson [Nel65] introduced the term *hypertext* “to mean a body of written or pictorial material interconnected in such a complex way that it could not conveniently be presented or represen-

[3] http://www.darpa.mil/, accessed: 2015-05-31.

[4] ARPA and DARPA are practically the same organisation which changed its name several times back and forth over the last decades: http://www.darpa.mil/About/History/ARPA-DARPA__The_Name_Chronicles.aspx, accessed: 2015-05-31.

ted on paper" [Nel65]. This idea inspired Berners-Lee to envision a document structure in which pieces of text can be linked to other parts of a document, or even other documents. This resulted in the development of a Web server, the Hypertext Markup Language (HTML) as a means of creating hypertext documents, the Hypertext Transfer Protocol (HTTP), and a Web browser to view documents [KR12]. A hypertext document is commonly referred to as a Web page; all Web pages reachable under a single domain are referred to as a website.

Despite his inspiration for the Web, Nelson [Nel99a] views this implementation as belying his vision. In particular he criticised ever-broken links, unidirectional links, and no proper traces to the original source, which he aims to overcome with his project Xanadu[5]. Nevertheless, the WWW is extremely popular today and, as can be observed every day, it did indeed enable better information interconnection and access, revolutionising the way mankind searches for information.

The power of the Web was also recognised by companies that started to commercialise the network in the early 1990s. During this time, the infrastructure provisioning of the Web was also successively taken over by private companies [KR12]. This development was facilitated by the fact that the WWW received a Graphical User Interface (GUI) around that time, by the appearance of the first browsers [KR12; W3C00].

This part of this thesis will focus on the Web and Web search. In order to do so, the basic ideas behind searching and the evolution of search engines will be discussed in the context of the evolution of the Web, showing how the basic paradigm of search has been extended over the years. However, before this can be done, some details will be provided on how the Web can be modelled, which will later clarify how algorithmic search engines work.

As the term *Internetting* suggests, the Internet is a network of networks [KR12; BFS03]. Given that a computer network can be seen as a graph, in which computers, network switches, etc. are nodes and the connecting wires are edges, the Internet as a whole can also be seen as a graph. While this is helpful at a technical level, for the purpose of understanding searching on the Web, it is not sufficient, as one server may host many documents. Consequently, when speaking about the Web, it is more sensible to view the Web as a graph, masking out the technical infrastructure and focusing on the logical structure of Web pages. In this view a Web page is represented as a node and a hyperlink as a directed edge (this is a result of the fact that a hyperlink can only point in one direction).

[5] http://xanadu.com/, accessed: 2015-05-31.

What is special about the Web graph is its dynamic. Web pages (nodes) and links (edges), are created, changed, and deleted continuously [BFS03].

A well-recognised study by Broder et al. [BKM+00] showed, using a Web crawl of more than 200 million pages and more than 1.5 billion links from May 1999, that the Web graph has the shape of a bow tie [BFS03; Lev10; LM06; VH07]. The study found that around 90% of all Web pages are connected by links (edges). If edges are treated as directional, this conglomerate of Web pages consists of 4 components of roughly the same size. Firstly, the Strongly Connected Component (SCC) which builds the core of the Web. Every Web page within the SCC is linked in such a way that any other Web page of the SCC can be reached by following edges. Secondly, the IN-group which can reach the SCC by following links but cannot be reached from within. Thirdly, the OUT-group which can be reached from the SCC but cannot reach it. Fourthly, there are tendrils which either exit IN or enter OUT but are not connected to the SCC. If a tendril of IN is linked to a tendril of OUT, this is called a tube. Finally, there are disconnected components. These relationships are illustrated in Figure 3.2.

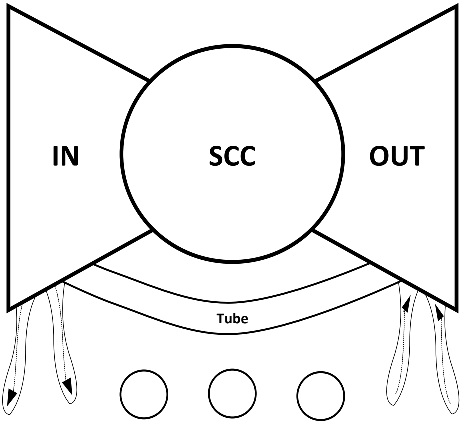


Figure 3.2.: The Bow-Tie-Structure of the Web.

However, it should be noted that the original study was conducted 15 years prior to the writing of this thesis and has never been repeated on such a large scale. As a consequence, it is not certain that the Web still maintains this particular structure. That said, it is likely that there are still more and less (maybe even not) connected parts as well as pages with an overhang of incoming or outgoing links, respectively. This, combined with the ever changing nature of the Web, poses severe challenges for automated Web crawlers.

3.2.2. A Short History of (Web) Search

Having roughly outlined the history of the Internet and the Web as well as provided some technical background, this section will now focus on the history of Internet search, again focusing primarily on Web search.

In the pre-Web era up to the early 1990s *Archie, Gopher,* and *Wide Area Information Servers (WAIS)* were tools used to retrieve Information [Gil93; Com95]. Archie servers maintained an index of files available on other machines. To achieve this, they regularly contacted other file servers and requested a list of all files available which was then merged into the index [Com95]. This approach was one of the first used to automatically search the Internet. Gopher, in contrast, relied on humans [RG97; Gil93]. When users connected to a Gopher server they were given a menu (created by humans). Each menu item would refer either to a document or another Gopher menu from the same or another server [Gil93; Com95]. Using Gopher one needed to know where to look for information, therefore a service known as Very Easy Rodent-Oriented Net-wide Index to Computer Archives (VERONICA) was used. VERONICA applied the Archie principle to Gopher by indexing Gopher menu items. While Archie and Gopher require accessing a remote server, WAIS is based on the client-server-principle [Gil93] explained for instance in [KR12]. Furthermore, WAIS automatically indexes document contents rather than document descriptions and operates based on word frequency.

From 1991 onwards, the WWW grew rapidly [KR12]. As a means of retrieving information, lists of hierarchically sorted hyper-links, known as Web directories, were established, with the intention of serving as entry points to the Web. These indices were mainly concerned with different research areas and can be seen as the application of cataloguing techniques, used in libraries, to the WWW. In fact, The WWW Virtual Library[6] was the name of the first Web catalogue [Lib14]. As the WWW underwent commercialisation, private companies also started to offer professionally maintained Web directories, one of the most prominent examples being Yahoo![7], which was founded in 1994 as "Jerry and David's Guide to the World Wide Web" [yahed]. However, they announced the closure for

[6] http://vlib.org/, accessed: 2015-05-31.

[7] http://www.yahoo.com/, accessed: 2015-05-31.

December 2014 [Rosed]. As a present day example of a Web directory the Open Directory Project (dmoz)[8] can be mentioned.

A key characteristic of such Web directories is that they are frequently maintained manually, implying that humans are involved in the analysis of website content in order to categorise it [GBR09; BHW11]. For instance, if the website of the Magic Circle – a British organisation dedicated to promoting the art of magic – http://www.themagiccircle.co.uk/ were to be categorised, it could be sorted under "arts"→ "performing arts"→ "magic"→ "societies", in fact that is exactly where it can be found in the dmoz. Users searching for information through a Web directory may either browse through the aforementioned hierarchy or utilise a tool that searches through titles and descriptions of websites as supplied by the maintainers.

It is reasonable to suppose that manual maintenance ensures quality and leads to an intuitive structure. However, directories do have two drawbacks, 1) maintenance effort and 2) scalability, which makes it hard for directories to cope with the ever-growing WWW [GBR09; BHW11]. This can, for instance, be seen by the fact that the aforementioned dmoz lists more than 4 million websites[9] while the whole indexed Web contains more than 4.5 billion Web pages according to [Wor15]. This means there are 1000 times as many Web pages automatically indexed than websites that have been manually collocated. Even though this compares websites (which makes sense for a catalogue) to Web pages (which is sensible for automated indexes), it shows impressively the relation between automated indexing and manual cataloguing. For this reason, technical means to retrieve data have been used from the early days of the WWW. Despite its growth, the Web's decentralised and heterogeneous structure – due to a lack of standardisation in how information is presented – made it difficult for humans to find all of the information available online. As a solution, Web crawlers were introduced in the early 1990s. These crawlers – sometimes referred to as spiders – are autonomous programs that follow hyperlinks to automatically retrieve new and updated websites. Around the same time, search engines based on crawlers were also built [MSHP09].

The architecture of search engines can be divided into three parts. Firstly, crawlers traverse the Web starting from a set of seed Uniform Resource Locators (URLs) subsequently following all encountered hyperlinks. Then, the retrieved websites are temporarily stored and analysed to extract key words and various other meta data. These key words and data are linked to their original source

[8] http://www.dmoz.org, accessed: 2015-05-31.

[9] As of 2015-03-28

and subsequently stored in the search index. This process runs – independent of user queries – infinitely in the background. The last part is the user interface or runtime system. It translates user-queries into machine-understandable terms, retrieves relevant entries from the index, ranks the results by relevance, and presents them to the user [VH07; LM06; Lev10]. A simplified schematic diagram showing the structure of a search engine is presented in Figure 3.3.

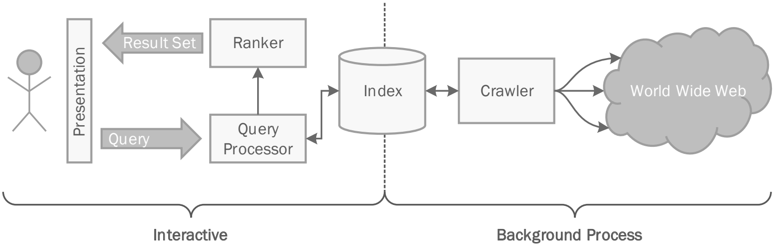


Figure 3.3.: Simplified Schema of a Web Search Engine.

Accordingly, research has been conducted in all three areas, which will be covered in Sections 3.3.1, 3.3.2, and 3.3.3. Firstly, Section 3.3.1 discusses which approaches are pursued to improve crawling or more generally data gathering, focusing on how sources that are hard to access can be tapped, also referred to as Deep Web Mining. Following on from this, Section 3.3.2 elaborates on approaches to data analysis and data extraction. In this regard it is – among other things – investigated how non-textual sources can be accessed and evaluated. Thirdly, in regard to ranking and presenting results to users, this section examines what actions have been taken to determine what a user's intention is and, consequently, in what order results should be presented.

3.2.3. Approaches to Search

Having briefly described how Web search evolved over the past two and a half decades, this section will classify the different approaches to searching. Firstly, it can be distinguished between manually maintained directories and modern search engines based on automatically assembled indexes. The latter type can be further divided into three categories: *general purpose, special purpose,* and *archive* search engines [Lew09b]. The purpose of the first two is to serve immediate information needs, while the purpose of the latter is to permanently conserve Web pages in order to create an archive that keeps Web documents available, even after they have been removed from the Web by their owners or creators. A search engine that falls into this last category is the Wayback Archive[10]. As archive search engines do not primarily aim to satisfy immediate information needs, rather they function at preserving information, they shall no longer be considered in this work.

General purpose search engines are search engines providing average results – i. e., not necessarily in-depth and little personalised information but covering a wide array of topics – to (average) users. To this end they do a wide crawl of the Web trying to collect as much data as possible, which is also referred to as the horizontal approach to searching. However, when following this approach, there is always a risk of accidentally excluding specific, and important, niche information in the course of reducing the result set to a manageable size. This is one of the reasons that different general purpose search engines return different results as they build their own indices, another being that they have different approaches to ranking.

Specialised search engines in contrast focus more precisely on a specific subject area or type of information. Examples of this include: searches for up to date information (e. g., news), searches for specific document types (e. g., image files), or searches in certain domains (e. g., scientific papers) [Lew09b], where one domain may consist of several sub-domains. For instance, the domain travel may consist of the domains flights, hotels, car rental, etc. [Cer10].

Different search engine providers – on the whole using different types of search engines or at least different implementations – usually offer different search options and return different results. In order to combine the strengths of various search engines, *meta search engines* have been developed. In contrast to conventional search engines they do not maintain an index of their own but redirect queries to a number of search engines. The retrieved results are either presented to the user without any processing, or more commonly, duplicates are eliminated and a new order is computed to offer a persistent ranking [Cer10]. Meta search engines themselves can fall into any of the three aforementioned categories, depending on their application scenario.

While it is an advantage of meta search engines that they can cover a broad spectrum, search functionality might be limited as only search operators supported by all underlying search engines can be used [GBR09]. In a study from 2009, Griesbaum et al. [GBR09] mention several examples of meta search engines, of which only two out of six were still available online at the time of writing

[10] https://archive.org/, accessed: 2015-05-31.

this thesis. Interestingly, the two sites still available are the people search engines PIPL[11] and YASNI[12]. This could lead to the assumption that search engines for people are of high demand. However, the fact that 123PEOPLE[13] went out of business in late April 2014 [fut14] is a contradicting indicator.

The big search engine providers (such as GOOGLE[14] or BING[15]) first and foremost offer a general purpose search services, which they enhance by providing special purpose search engines alongside. Initially, both general purpose and special purpose searches were offered through the same website but through different interfaces. However, more recently, special search results started to be merged with general Web search results, making the search in a sense *universal* [Gooedb; Jased]. In terms of functionality, *universal search* can be compared to meta search engines; a query is answered by retrieving results from various specialised vertical search engines (i. e., search engines focusing on a special domain, or file type), before the results are integrated and displayed to the user [Qui09; Mared; Dav07]. While these sub-search engines belong to the same operator in the case of universal search engines, this is not necessarily the case for meta search engines.

A different type of service, which provides information and should therefore be mentioned, although not technically a search engine, is knowledge-based search systems. These systems deliver answers to queries by computing them based on built-in data, rather than collecting the information from the Web [Wei10]. An example of such a system is WOLFRAM|ALPHA[16], which according to their website, maintains more than 10 trillion pieces of data [Woled] that have been gathered by at least 150 human editors prior to its launch in 2009 [Gil09]. Along with the collection of data, there is an on-going maintenance process with the aim of "providing a single source that can be relied on by everyone for definitive answers to factual queries" [Woled]. The processing of queries posed in natural language is not yet perfect, which may annoy users [Gil09]. Another disadvantage, that should be mentioned, is that while WOLFRAM|ALPHA is superior to conventional search engines for queries to well-structured data, it is not able to serve the whole scope of complex searches, such as up-to-date Web data [Cer10]. GOOGLE, too, is creating a knowledge base of facts to power its search. In contrast to WOLFRAM|ALPHA, GOOGLE does no longer rely solely on

[11] https://pipl.com/, accessed: 2015-05-31.

[12] http://www.yasni.de/, accessed: 2015-05-31.

[13] http://www.123people.de, accessed: 2015-05-31.

[14] http://www.google.com, accessed: 2015-05-31.

[15] http://www.bing.com/, accessed: 2015-05-31.

[16] http://www.wolframalpha.com/, accessed: 2015-05-31.

crowd-sourced humans to achieve this task but builds its so-called knowledge vault entirely based on algorithms [Hod14].

3.3. Challenges in Web Search

This section discusses four important research fields in the context of online search engines. Firstly, some details on how to access the so-called Deep Web will be provided. Secondly, it will be discussed how information can be extracted from Web sources in order to be meaningfully stored. Thirdly, it will be elaborated on how search results can be sorted. To this end, ranking will be important. Finally, the broad topic of social search including extending search technology by means of crowd-sourcing will be introduced.

3.3.1. Mining and Integrating the Deep Web

Despite all of the crawling, indexing, and retrieval efforts, many Web sources exist that currently cannot be accessed in automated ways, and their number is growing [Wried]. This phenomenon is not new, it is at least known since 1994, when, according to a not primarily scientific marketing white paper [Ber01], Dr Jill Ellsworth coined the term *Invisible Web* for non-indexable websites. In analogy to an ocean, Bergman [Ber01] introduced the term *Deep Web* for this phenomenon because, like the deep ocean, he considers these parts of the Web to be visible despite not being accessible. Following his analogy, he refers to the accessible Web as *Surface Web*. Similarly, Stock [Sto03] divided information on the Web into the two categories: information available on the Web (Surface Web) and available through the Web (Deep Web). Following the argument of Bergman [Ber01], here the terms Deep Web and Surface Web will be used for simplicity. There are various reasons why the Deep Web may be inaccessible [GBR09]:

- Access restrictions, such as password protection
- Websites exclude themselves from being indexed through the use of a Robots Exclusion Protocol[17] which uses text-files to state which resources a robot may access – while this is not a technical prevention, it is good manners to obey them

[17] http://www.robotstxt.org/, accessed: 2015-05-31.

- The content is in a disconnected part of the Web, i. e., a portion of the Web which has out-links and may be interconnected but which has no in-links and, therefore, no incoming connection from the rest of the Web by hyperlinks and can thus not be found (see Figure 3.2)
- Technical restrictions, such as database-driven designs, i. e., websites are a result of user-specific queries to a database.

Most commonly, the last point is implied, when the term Deep Web is used [CHL+04; MKK+08; Lew09b; Raj09; RB09; XCZ+10]. With regard to size, Bergman [Ber01] estimated that the size of the Deep Web is roughly 400 to 550 times larger than the indexable Web, with more than 200,000 Deep websites. Chang et al. [CHL+04] estimated in 2004 the number of Deep Web databases to be around 450,000 (with a total of 1.258 million query interfaces) by extrapolating from a sample of 1,000,000 IP addresses to the whole IP space (excluding reserved areas). Researchers from Google stated that there were about 10 million high-quality Deep Web forms in 2008 [MKK+08], or about one billion pages of structured data, in 2011 [CHM11]. Even though these figures may neither be validly compared with one another nor be accurate today, they support the assumption that the Deep Web does exist and is growing.

Madhavan et al. [MKK+08] propose three principles to enhance Deep Web access. First, judging the quality of sources; second, crawl only subsets to ensure the load on crawlers does not become too high; third, develop heuristics to discover similarities between data sources as it is not likely that domain specific methods scale to the Web. In terms of judging the quality of Deep Web source Xian et al. [XCZ+10] developed a utility maximisation model that can be used to assess the value of Deep Web sources and thus help to decide which Deep Web sources to choose if a number of sources are available.

In 2011, Google's goal of making the Deep Web accessible to search engine users was emphasised by Cafarella et al. [CHM11] and underpinned by the proposal of two approaches to Deep Web data collection. Firstly, they mentioned the vertical search engines, as described earlier; however, they consider this approach impractical as a subject area may be hard to define and a significant proportion of human interaction is needed to integrate the various sources. As a solution, they propose to *surface* Deep Web content by posing queries to inaccessible databases and indexing the resulting pages. According to Cafarella et al. [CHM11], using this method, Google was able to index the content of several million Deep Web databases in a completely automated manner. Consequently, they consider their approach superior to any manual approach.

Despite this success, Cafarella et al. [CHM11] confessed that there were still major challenges to meet. These challenges include developing or using semantic services to improve posing queries to Deep Web databases, as well as gathering data from other Deep Web sources, such as the social Web (i. e., social network sites such as Facebook). Regarding the semantic problems, ontological approaches have been proposed by [J L10] and [LLD11] as a solution.

Marques and Figueiredo [MF10] found that it was difficult to surface Deep Web content that is hidden behind POST forms as they do not provide unique URLs to the underlying content. As a solution, they introduce *stigmergic hyperlinks* to address Deep Web search. While looking like common HTML hyperlinks, stigmergic hyperlinks are server side objects that have a life attribute that increase whenever users click the link and decreases overtime if it is not clicked. This was inspired by ants leaving pheromones, i. e., users leave a virtual pheromone trace by clicking. More precisely, the authors suggest using these hyperlinks to reference Deep Web content. To this end, stigmergic hyperlinks are further enhanced by search functionality. Eventually, this will lead to an index of Deep Web Content with additional meta data about relevance for users. Furthermore, stigmergic hyperlinks allow for the index to be downloaded and, thus, be used by third-parties.

Rajaraman [Raj09] differentiates two types of Deep Web access. In addition to the previously described "query deep Web sources and indexing the results" approach to Deep Web access – referred to as *Deep Web crawl* – he mentions *Federate Search*, where an Application Programming Interface (API) is used to access sources at query time. Having discussed both types, his suggestion – implemented in the Kosmix Explore Engine[18] – is using a hybrid approach combining the comprehensiveness of Web search with the specificity of federate search.

Leaving quality concerns aside – this review shows that retrieval and indexing of documents is no longer a severe problem and that automatically exploiting the Deep Web is currently being addressed by various researchers [BR10].

However, there are more issues to be addressed in order to fully satisfy current information needs. For instance, in 2009, Dopichaj [Dop09] stated that it was technically impossible to meaningfully answer queries such as "return all pages that contain product evaluations of fridges by European users." Ceri [Cer10] came to a similar conclusion one year later. Whilst Dopichaj [Dop09] favours the Semantic Web as a solution, Ceri [Cer10] proposes the so-called search computing paradigm that aims at integrating (i. e., combining) Deep Web sources in

[18] Formerly available at: http://www.kosmix.com/.

order to be able to answer such queries. The need for this was also identified by Chang et al. [CHL+04]. Some years earlier, Masermann and Vossen [MV00] proposed a query language that allows the querying of a number of databases on and off the Web without schema-knowledge. This, however, explicitly targeted databases, rather than general information available online.

The framework envisioned by Ceri [Cer10] would work as follows: Initially, a search query would be processed by a query optimiser that chooses suitable underlying search services for different parts of the query. Then, results of these search services would be joined and displayed to users who would be given the chance to modify their queries dynamically, referred to as *liquid query processing* [BBCF10; BBC+11]. The entire framework builds on the creation of two new user groups or communities. On the one hand *data providers* who offer data services and, on the other hand, *developers* who build search services based on these data services. In this scenario, data services are built on top of data sources which are collections of data regarding similar domains. They may consist of scraped Web pages or be Web services themselves. This exemplifies the fact that search services would integrate data services in a transitive manner.

Baumgartner et al. [BCGH10] describe how a data service (also referred to as data mart) can be built using Deep Web mining on a given Web source. Their approach – called Lixto – is to create wrappers for any Deep Web data source (and others) by manually training computers (using a graphical user interface) to fill out forms and to retrieve data. These wrappers then form the basis for Web services to be included in data marts based on a service-oriented architectures (SOA)[19]. Building on this, Campi et al. [CCG+10] describe how data marts can be built and registered with the framework. An overall architecture for search computing is suggested in [BBC+10].

3.3.2. Extracting Meaning

As stated before, search engines crawl the Web to retrieve Web pages and analyse their contents in order to index them, as indicated in Figure 3.3. The indexer, the part that creates and updates the index by analysing Web content, relies on techniques from IR, in order to summarise and categorise retrieved documents. In the context of the Web, the term Web Mining is often used for this action. As Baeza-Yates and Ribeiro-Neto [BR11] argue, this is even more challenging because data is distributed, often volatile, and heterogeneous. Furthermore, data can be both structured and unstructured, the quality of data is questionable, as

[19] An introduction to SOA can be found in [NL05]

it does not undergo an editorial process, and finally, its volume is tremendous and still growing.

In order for the indexer to work properly, content mining, i. e., text or multimedia mining has to be applied. Given that in Western culture text is the main vehicle for information, this section will focus on textual information. Compared to data mining (finding unknown patterns in data), text mining has the clear advantage that the information is not hidden. Beyond that, text on the Web is often enriched with markup language that can help to structure the contents [WFH11].

Nevertheless, text in itself is not structured and as such poses challenges. Fields of interest in this area include:

- *Clustering* [LRU14], i. e., taking many websites as an input and return topic-centred clusters;
- *Link Analysis* [LRU14; Liu07], which focuses on how information on the Web is related. This includes developing hubs, which are sites linking to many pages on different topics, and authorities, which are sites that focus on a specific topic;
- *Text and Website Pre-Processing* [Liu07] by means of stemming (i. e., normalisation of words to their stems), removal of stop words, identifying important parts of a website such as links and other HTML elements, finding content blocks within a website as well as finding duplicates of websites or pages for efficient indexing; and
- *Knowledge Mining* from the Web, examples of which include *YAGO2* [HSBW13], a knowledge base built automatically from Wikipedia[20] and *WebChild* [TMSW14], a knowledge base that also compromises common sense knowledge, such as apples are round, which is currently lacked by computers, further examples of knowledge bases are GeoNames[21] and WordNet[22]; a good general introduction to information extraction is given by Balke [Bal12b].

For the sake of completeness multimedia media mining shall also be briefly touched upon. There are two utterly different approaches to indexing multimedia content. On the one hand, related text (textual meta data regarding or text

[20] http://www.wikipedia.org/, accessed: 2015-05-31.

[21] http://www.geonames.org/, accessed: 2015-05-31.

[22] http://wordnet.princeton.edu/, accessed: 2015-05-31.

surrounding the multimedia content) is analysed. On the other hand, multimedia content is analysed directly, which is computational more challenging [Lew09b]. However, analysing the content itself offers the possibility of enhancing results found through the first method. Furthermore, it allows for the retrieval of multimedia content in the absence of annotations [LSDJ06].

3.3.3. Understanding User Queries and Ranking Results

Once the index has been built, it can be queried. However, user queries commonly have to be pre-processed in order to determine a user's intention because human language is not machine-understandable yet. Human language, more commonly referred to as natural language, entails several challenges, two important ones being synonyms (multiple words for one object, e. g., house as a building to live in versus house as an aristocratic family line such as the house of Windsor) and polysemy (one word for different objects, e. g., bank as in river bank versus bank as a financial institution) [LM06].

One of the ways to address this challenge is semantic search [Dop09], which exploits the Semantic Web. The idea of the Semantic Web is to enhance texts available through the Web by semantic information in order to make it machine-understandable. It was initially proposed in 2001 by Berners-Lee et al. [BHL01]. The personal view of Berners-Lee [Ber09] is that the key value proposition of the Semantic Web is linking – this was also the case with the Web initially – in a way that both humans and machines can understand relationships between data and discover new data by exploration. On a technical level, Berners-Lee et al. suggested the Resource Description Framework (RDF) [W3Ced], a standard intended to facilitate data interchange.

However, Dopichaj [Dop09] found that the Semantic Web and its enabling technologies were not really being used until 2009, the reason, most likely being due to the maintenance efforts it requires. To overcome this and to increase Semantic Web usage, the search engine providers Bing, Google, and Yahoo! announced, in June 2011, that they would join their efforts to foster the Semantic Web by focusing on a single standard. Nevertheless, according to [Fox11] this standard would only be used for displaying additional information on result pages, rather than being used for the actual search. In 2012, Google announced and introduced Semantic Web technologies in their actual search [Efr12]. As of today, it remains unclear whether the Semantic Web will eventually be better supported by websites, which is the necessary first step in order to provide more possibilities for search based on natural language.

Once user intentions are clear, the index can be queried accordingly and the results can be presented. The previously mentioned early search engines were only able to compute whether a search string was contained in a document or not. Despite finding documents that users might not have found themselves and extracting documents that have a certain relevance, users were left with potentially thousands of results to go through manually in order to find the sites most relevant to them. For this reason, it was necessary to order results according to a "guessed" relevance to users [Dop09]. To this end, Brin and Page [BP98] – the founders of Google – suggested *PageRank* as an ordering criterion in 1998[23].

PageRank is commonly viewed as a breakthrough in search technology because it does not rely solely on information presented on a website but incorporates linking factors [LM06; MSHP09; GBR09]. The technique used in PageRank is based on the underlying idea that the Web can be seen as a graph (recall that pages are nodes and hyperlinks are edges). Furthermore, edges of this graph are interpreted as citations or recommendation, e. g., if website A links to website B, A attributes some credit to B [BP98; Kle99; LM06; GBR09]. This implies that B is probably a good source in context of whatever is linked. For the actual calculation of the PageRank, a method known from citation analysis or recommendation behaviour is used, which was traditionally used to judge the importance of a person or source. In this way relevant sources – also known as *authorities* – can be distinguished from non-relevant sources [BP98; VH07; Dop09]. Formally, PageRank was originally expressed as:

$$PR(A) = (1-d) + d \sum_{t_j \in T} \frac{PR(t_j)}{C(t_j)}$$

As can be seen, calculating the PageRank for a website A is an iterative process. It is based on the PageRank of all in-linking documents (citations) T divided by the number of out-links for these documents $C(t)$. The constant d, also referred to as dumping factor, is a probabilistic simulation for the chance that a user stops browsing and starts at a new page. This is also known as the random surfer effect [BP98].

Besides PageRank, a number of other ranking algorithms exist that also build on linking factors, of which Hypertext Induced Topic Search (HITS) based on [Kle99] is most well-known [FLM+06; LM06; GBR09]. In contrast to PageRank, which only includes in-links pointing to a Web page, HITS also considers out-

[23] N. B. later Google introduced many other factors.

links pointing from a Web page. As a consequence, a node in HITS receives two values: a *hub* and an *authority* value. Both values are circularly linked, in that the hub value is calculated by summing the authority values of the outlinked Web pages and the authority value is calculated by summing the hub values of the in-linked Web pages. Even though HITS was developed at roughly the same time as PageRank, it was initially not intended to be commercialised and was only used from 2001 by a commercial search engine [LM06]. The major difference between PageRank and HITS is the document basis. Due to the complex and circular nature of HITS it potentially takes very long to execute. Therefore, it is commonly only applied to a subset of potentially relevant documents [Dop09]. While this offers the possibility of a more precise ranking, it does so at the expense of potentially leaving important documents out of scope. A third approach, by Lempel and Moran [LM00], called Stochastic Approach for Link-Structure Analysis (SALSA) includes ideas of both HITS and PageRank [FLM+06].

Google's PageRank algorithm was later improved and additional criteria were taken into account allowing better result rankings to be calculated [VH07]. One example of this can be found in [KL02]. According to Google, their search incorporates more than 200 factors in order to be able to judge relevance [Gooeda; Mitedc; Sma09; Sul10]. Unfortunately for the scientific community, these factors are not publicly available as they are regarded as trade secrets. Nevertheless, in 2009 Smarty [Sma09] published a community-generated list[24] of nearly 130 factors which are likely to influence Google's search result ranking. The factors are grouped into the following categories: Domain (e. g., domain age); Server-side (e. g., geographical location of the server); Architecture (e. g., HTML structure); Content (e. g., content uniqueness); Internal Cross Linking (e. g., number of internal links to the given page); Website factors (e. g., overall site update frequency); Page-specific factors (e. g., content duplication with other pages of the same site); Keyword usage and keyword prominence (e. g., keywords in the title of a page); Outbound links (e. g., quality of pages the site links to); Backlink Profile (e. g., quality of Web pages linking in); Each Separated Backlink (e. g., anchor text of a link); Visitor Profile and Behaviour (e. g., number of visits); Penalties Filters and Manipulation (e. g., Keyword over usage); Brand/Author Reputation (e. g., Use of the domain in Google AdWords[25]).

[24] The list can be found at: https://spreadsheets.google.com/pub?key=tMc56KQJFjYOBMcEq263r4g, accessed: 2014-05-09.

[25] https://www.google.com/adwords/, accessed: 2015-05-31.

Griesbaum et al. [GBR09] describe four categories of factors on which search result ranking is based: *on-page, on-site, linking* and *user behavioural* factors. Most of the categories presented by Smarty [Sma09] can be mapped to one of the broader categories defined by Griesbaum et al. [GBR09], as is presented in Table 3.1. However, the category *penalties* and *Brand/Author Reputation* are not covered by Griesbaum et al. Given that the broader categories cover the most important factors and that in particular *Brand/Author Reputation* is very specific to Google, this work focuses on the broader four categories, presented in the order in which they were implemented by various search engines.

Table 3.1.: Mapping of Ranking Factor Categories by Griesbaum et al. and Smarty.

Categories according to Griesbaum et al. [GBR09]	**Categories according to Smarty [Sma09]**
On-Page Factors	Architecture, Content, Page-specific factors, Keywords usage and keyword prominence
On-Site Factors	Domain, Server-side
Linking	Internal Cross Linking, website factors, Outbound links, Backlink Profile, Each Separated Backlink
User Behavioural Factors	Visitor Profile and Behaviour

On-page factors – for example proximity, function, and format of words within a text [GBR09] – were first used to solve the ranking challenge. With regards to function and formatting, HTML-markup was used to determine the importance of words. To calculate similarity and proximity, vector space models that translate documents into vectors, on which mathematical calculations are based, are commonly used [Dop09]. In the middle of the 1990s, Lycos[26] gained considerable market share after including the word proximity, i. e., the proximity of (multiple) search terms within a document, in their ranking algorithm [MSHP09]. The drawback of this mechanism is, however, that it is vulnerable to spam. In the context of search engines spam refers to Web pages of presumably low quality that aim at receiving a high ranking in popular search engines by exploiting knowledge of how search engines work. As a practical example, it is very simple to generate Web pages repeating the same words many times in important functions such as headings [Dop09].

[26] http://www.lycos.com/, accessed: 2015-05-31.

On-site factors are technical factors regarding the entire domain on which documents are hosted. Both [Sma09; GBR09] mention domain registration details as an example of this category. However, Griesbaum et al. [GBR09] state that factors in this category are only guesswork. Other factors identified by Smarty [Sma09] include, among others, non-linked domain mentions and geographical location.

Linking factors influence the search based on the mechanisms explained, when discussing PageRank and HITS. Besides these two measures that are mainly based on counting in-links and out-links, the quality of the linked websites is also determined in terms of its content Smarty [Sma09].

One of the problems with link analysis-based ranking algorithms, in general, is their selectivity towards established sites because it is easier for established sites to gain new in-links which again increases their ranking value. This in turn increases their visibility which may well lead to more in-links and so on and so forth [GBR09].

While it has to be said that these linking mechanisms are harder to exploit than on-page-factors it is still possible to create spam when linking algorithms are used. As a practical example, link farms [Dop09] shall be mentioned. A link farm is a group of websites with the sole purpose of linking to each other in order to increase their ranking [GBR09; LM06].

Recently, *human behavioural* factors were integrated into search engines. They are based on the assumption that if two people pose an identical query their information need is not necessarily the same [Dop09]. In this regard, it was shown at Microsoft Research that implicit measures of human interest (for instance the number of results that were visited or the time spent on a particular result) can reliably be related to ratings of user satisfaction [FKM+05]. Based on this observation, it was later also shown that considering implicit measures in rankings can indeed improve the ranking [ABD06]. Riemer and Brüggemann [RB09] came to the conclusion that incorporating human behavioural factors, more precisely the personal characteristics of users, can increase the perceived relevance of search results. Additionally, their work provides an overview of personalisation techniques and shows which search engines utilise them.

Within the sphere of human behavioural factors is the aspect of social factors. For instance, Google offers the possibility to recommend search results to Google+[27] friends and exploits the social graph, i. e., users' relations to their

[27] https://plus.google.com/, accessed: 2015-05-31.

friends and what these liked, in order to provide better search results. However, this also bears tremendous challenges relating to privacy.

3.3.4. Crowds and Social Search

During the last decade, the Web experienced a trend of socialisation which led to the transition from the "*read-only*" to the "*read/write*" Web. This has also led to the term *Web 2.0* in 2006. Prior to that, the Web was mainly used to consume information. The practical consequence of this transition was that users could now easily contribute their own content to the Web [VH07]. This trend is ever growing as can be seen by way of example of YouTube[28] where users upload about 30 hour of videos every single minute [Koo13]. Furthermore, the advancement of the Web has led to a phenomenon called *crowdsourcing*, i. e., out sourcing of tasks to the *crowd*, which has been defined as "new pool of cheap labor: everyday people using their spare cycles to create content, solve problems, even do corporate R & D." [How06] or "the collective of users who participate in the problem-solving processes." [Bra08].

In the context of social search engines, crowds enhance search results or data collections through social interaction and sharing. Burghardt et al. [BHW11] classifies the following approaches: *social indexing, social question answering, collaborative filtering,* and *collaborative search.*

Social indexing refers to collaboratively maintained Web directories; examples include the aforementioned dmoz but also tagging, as for example, on bookmarking services like Delicious[29]. The clear advantage of such systems is that searchers and taggers are essentially the same group of people. Consequently, there should be a common understanding of terms. However, this common understanding might be a learning process if new individuals or individuals from other groups are included. In particular for non-text-based content, tagging can be used as a means of indexing in order to make certain document types searchable (tag-based search). While this so-called free tagging works well on some platforms, such as hash-tags on Twitter[30], for general-purpose search this approach has not yet been applied. Furthermore, quality cannot be assured as taggers are free in their tag choice. This could only be overcome through the use of professional taggers.

Question and Answer Systems offer users the possibility to pose questions which can then be answered by a community or paid experts, sometimes both

[28] http://www.youtube.com/, accessed: 2015-05-31.

[29] http://delicio.us/, accessed: 2015-05-31.

[30] http://twitter.com, accessed: 2015-05-31.

[ABD06; BHW11]. One issue with this is that if answers are provided by professional experts, these services are usually not free to use.

Collaborative filtering refers to recommender systems that filter relevant documents based on the similarity of users, i. e., if the service believes that user A is similar to user B, it will provide user B with search results that user A liked, either explicitly through a liking mechanism or implicitly through not refining the search or staying long on a result.

Finally, *collaborative search* can be seen as a prime example of social search. It serves as platform that enables independent users to jointly work on their search tasks. An example of such a search engine was Eurekster[31]. However, at the time of writing this thesis it remains unclear what happened to their personalised social search (for more details see [Sul04]), as they now advertise innovative banner ad services. One alternative called Rollyo[32] suffered the same fate [Bra12], however another, still operational, is the online search engine provider blekko[33]. Here, users can create and make publicly available so-called slash-tags to determine which websites to be searched.

Despite the advantages social search has to offer, such as working jointly on a task or receiving search results based on friends' experiences, there are people who would rather rely on objective standards than on their social graphs [Miteda].

3.4. Added Value Through Human Involvement

Having discussed technical means to Information Retrieval (IR) in the preceding sections, the following section focuses on adjacent topics. It has been established that ultimately only humans can judge the quality and relevance of information. Consequently, curation of data and the idea of information brokering will be introduced in the following subsections.

3.4.1. Data Curation

From museums and exhibitions stems the idea of *curation*. The term curation is a back-formation from the word curator and refers to the selection, organisation, and care-taking of items in a collection or exhibition [Oxfeda].

[31] http://www.eurekster.com/, accessed: 2015-05-31.

[32] Formerly available at: http://rollyo.com.

[33] http://blekko.com/, accessed: 2015-05-31.

Rather recently, the topic of digital curation started to be discussed in library and information science, where it still vastly resides today. Unlike IR, which focuses on quickly finding information, digital curation focuses on preserving (huge) data sets gained through scientific experiments (e. g., physical, biological, or astronomical data) for later usage [Cho08; PAM11; DT07; Hei11].

Main tasks in this digital curation are cleaning data and, probably more importantly, updating data to new technical standards in order to make it accessible for future use [GST+02; LMLG04; RPS+05]. This is to counteract the fear of a "digital dark age", a point in the future where information from today cannot be read any more because file formats are unknown, something that actually happened at the National Aeronautics and Space Administration (NASA)[34] who had great difficulties to read data from decades ago, e. g., from lunar missions [Bla90].

In library science, electronic repositories storing data and information are referred to as institutional repositories [Smi08; Cho08]. These repositories can – to some degree – be compared to indexes of search engines or search catalogues as they build the core of the respective information systems.

Highly relevant in this context is the research topic of data provenance in databases, an overview of which is given in [Tan07]. Data provenance, also referred to as data lineage, describes means for every datum to be able to trace its origin and reconstruct modifications. This is important for two reasons: firstly, knowing the origin of a source can increase trust (for instance shown in [SPM11] by the example of security knowledge). Secondly, it allows for crediting the right people when using their data, similar to referencing works by other authors in the scientific community, which could otherwise lead to copyright infringements.

The research of Buneman et al. [BCCV06] in this context has been institutionalised in the British Digital Curation Centre (DCC)[35]. The DCC suggest a curation life cycle model [DDCed] consisting of the following phases:

1. Data is *created or received* in order to enter the curation process according to some rules.

2. Data is evaluated and selected for curation, which is referred to as *appraise & select*.

3. The data is *ingested* into the data storage system used in the curation process.

[34] http://www.nasa.gov/, accessed: 2015-05-31.

[35] http://www.dcc.ac.uk, accessed: 2015-05-31.

4. The *preservation action* prepares data for long-term storage to ensure integrity and its readability.

5. The data is *stored* adhering to relevant standards.

6. Once stored, the data has to be kept available for *access, use, and reuse.*

7. Finally, *transforming* it allows for new data to be created based on the stored data, e. g., by translating it to a different file format.

This sequence is accompanied by a number of auxiliary tasks which are not particularly relevant to this work. These tasks are mainly administrative in nature and comprise *preservation planning* and *community watch and participation* to name but a few. The complete set can be found online [DDCed].

3.4.2. Information Brokers

Information brokers are professionals who specialise in providing clients with information gathered from various sources. While the term is widely accepted, Rugge and Glossbrenner [RG97] point out that information brokers should really be called IR specialists or information consultants because they do not really broker (buy and sell) information but retrieve information that a client specified a priori. Thus, it is a service of finding, analysing and presenting the information that is paid for, rather than the information itself. In this context, IR is understood as finding the information a client wants.

This field of operation has been described for an English audience in [RG97] and for a German audience in [Bac00]. Both works consider themselves manuals for people who want to become self-employed information brokers. As a prerequisite, they state people wishing to become information specialists should already be domain experts in a specific domain. Given their nature, these works focus mainly on the practical tools of data retrieval, for example online and offline databases, libraries, and telephone interviews. However, given that over a decade has passed since their publication, many facets of retrieval they describe, particularly online aspects, can be considered outdated.

Self-management, project management, and time management, as well as marketing oneself, are discussed with the focus that information brokers sell themselves or their competencies rather than actual information. Further aspects include pricing and legal aspects. Pricing will be extensively discussed in Chapter 5 and Chapter 6, whereas legal aspects will be touched upon in the context of use case discussions in Section 4.3.

Furthermore, both highlight common problems in finding information. Even though the authors do not state this explicitly, the problems they describe are very similar to those that have been discussed in the previous sections on search engines. Common problems are:

- Understanding what the client wants
- Obtaining data from various sources explicitly including interviews and telephone enquiries besides the aforementioned offline and online source
- Rating and cleansing of the retrieved data (e. g., determining accuracy and handling duplicates, maybe even determine that there is no information on a topic)
- Presenting the information to the client

Thus, information brokers are comparable to search engines but are humans and, therefore, should supposedly yield better quality results.

In 2000, Bachmann [Bac00] reported, referencing the old website of Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis e.V. (DGI)[36] [Infeda], a number of 1,000 to 5,000 employed information intermediaries as well as up to 200 self-employed information specialists. Importantly, both sources stated that most of the 200 information specialists also do other business such as consulting work. However, the new website of DGI [Infedb] does not state these figures any longer. Additionally, it does not even describe the position *information broker* any longer, but just links to Wikipedia. Also, the special interest group information brokering of DGI (German: AG Infobroker) has no description of its own on the new website [Infedc], which just links to the old site [Infedd], according to which the last meeting was held in October 2010.

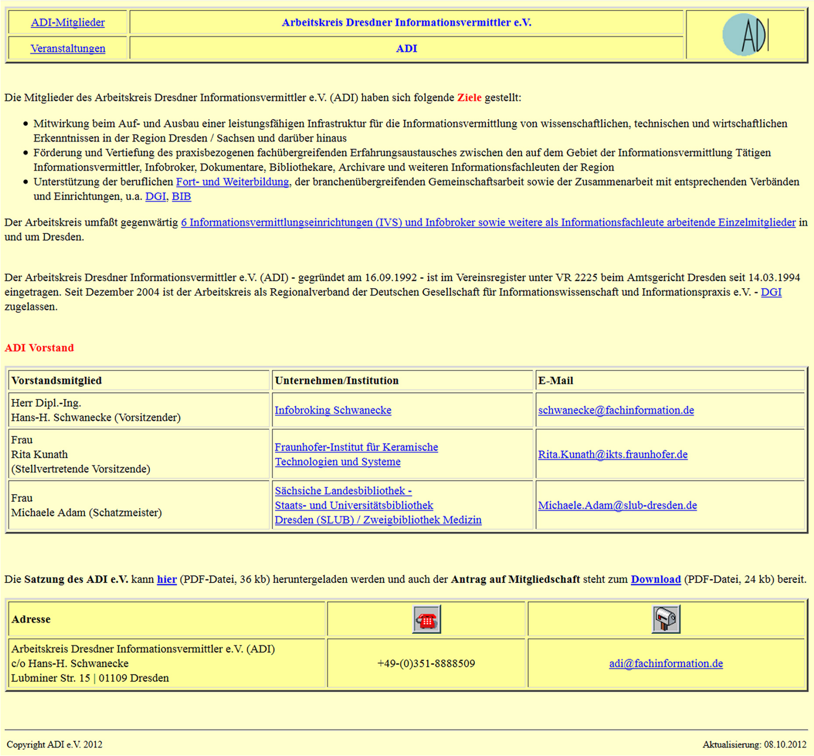
The AG Infobroker recommends working in regional groups. However, the number of regional groups has dropped by 23.08 percent, as evident when comparing the new [Infede] and old [Infedf] website as shown in Table 3.2[37]. Furthermore, the professionalism of some of these groups can be questioned by looking at their websites – one of which is presented in Figure 3.4 as an example.

This serves as evidence that entirely manual information provisioning produces high-quality output but also has had its time. Another hint in this direction is that fact that both works [Bac00; RG97] have been written at the turn of

[36] http://www.dgi-info.de/, accessed: 2015-05-31.

[37] For readability reasons, not all groups have been spelled out in Table 3.2.

the millennium, which was only at the start of Google's success. Even Bachmann [Bac00] states that online sources are of increasing importance. Furthermore, already in 1995, the idea of digital information brokers was discussed [FEFP95]. All of this suggests that while the knowledge maintained in information brokers' heads is very valuable expert knowledge, an entirely manual process is not going to survive on its own.



ADI-Mitglieder	Arbeitskreis Dresdner Informationsvermittler e.V.	ADI
Veranstaltungen	ADI	

Die Mitglieder des Arbeitskreis Dresdner Informationsvermittler e.V. (ADI) haben sich folgende **Ziele** gestellt:

- Mitwirkung beim Auf- und Ausbau einer leistungsfähigen Infrastruktur für die Informationsvermittlung von wissenschaftlichen, technischen und wirtschaftlichen Erkenntnissen in der Region Dresden / Sachsen und darüber hinaus
- Förderung und Vertiefung des praxisbezogenen fachübergreifenden Erfahrungsaustausches zwischen den auf dem Gebiet der Informationsvermittlung Tätigen Informationsvermittler, Infobroker, Dokumentare, Bibliothekare, Archivare und weiteren Informationsfachleuten der Region
- Unterstützung der beruflichen Fort- und Weiterbildung, der branchenübergreifenden Gemeinschaftsarbeit sowie der Zusammenarbeit mit entsprechenden Verbänden und Einrichtungen, u.a. DGI, BIB

Der Arbeitskreis umfaßt gegenwärtig 6 Informationsvermittlungseinrichtungen (IVS) und Infobroker sowie weitere als Informationsfachleute arbeitende Einzelmitglieder in und um Dresden.

Der Arbeitskreis Dresdner Informationsvermittler e.V. (ADI) - gegründet am 16.09.1992 - ist im Vereinsregister unter VR 2225 beim Amtsgericht Dresden seit 14.03.1994 eingetragen. Seit Dezember 2004 ist der Arbeitskreis als Regionalverband der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis e.V. - DGI zugelassen.

ADI Vorstand

Vorstandsmitglied	Unternehmen/Institution	E-Mail
Herr Dipl.-Ing. Hans-H. Schwanecke (Vorsitzender)	Infobroking Schwanecke	schwanecke@fachinformation.de
Frau Rita Kunath (Stellvertretende Vorsitzende)	Fraunhofer-Institut für Keramische Technologien und Systeme	Rita.Kunath@ikts.fraunhofer.de
Frau Michaele Adam (Schatzmeister)	Sächsiche Landesbibliothek - Staats- und Universitätsbibliothek Dresden (SLUB) / Zweigbibliothek Medizin	Michaele.Adam@slub-dresden.de

Die **Satzung des ADI e.V.** kann **hier** (PDF-Datei, 36 kb) heruntergeladen werden und auch der **Antrag auf Mitgliedschaft** steht zum **Download** (PDF-Datei, 24 kb) bereit.

Adresse	☎	📠
Arbeitskreis Dresdner Informationsvermittler e.V. (ADI) c/o Hans-H. Schwanecke Lubminer Str. 15 | 01109 Dresden	+49-(0)351-8888509	adi@fachinformation.de

Copyright ADI e.V. 2012

Aktualisierung: 08.10.2012

Figure 3.4.: Website of Arbeitskreis Dresdner Informationsvermittler (ADI) as an Example Website of a DGI Regional Group, accessed: 2015-05-31.

Table 3.2.: Overview of DGI Regional Groups.

Regional Group	Old List [Infedf]	New List [Infede]
ADI (Dresden)	✔	✔
AIT (Thüringen)	✔	✔
AKI Hamburg	✔	✖
AKI Magdeburg	✔	✔
AKI Rheinland	✔	✔
AKI Rheinland-Pfalz/Eifel	✔	✔
AKI RheinMain	✔	✖
AKI Rhein-Neckar-Dreieck	✔	✖
AKRIBIE (North-Rhine Westphalia)	✔	✔
BAK (Berlin)	✔	✔
BRAGI (Brandenburg)	✔	✔
Infotreff Ruhrgebiet	✔	✔
MAID (Munich)	✔	✔

3.5. Method and Focus of this Part

In this section, the focus of this part of this work will be outlined by identifying shortcomings of current search technology and highlighting the aim of this work. Then, the method applied for this part of the thesis at hand will be presented.

3.5.1. Limitations of Search Technology

The previous sections have given an overview of important historical developments and recent approaches to information organisation and retrieval with a major focus on Web search. It can be stated that search engines were developed for one main reason: to satisfy ever-growing, increasingly complex information needs.

While, generally speaking, retrieval and indexing of Web documents is no longer the problem it was during the early days of the WWW, it is still far from being solved in its entirety [BR10]. For instance, there remain limitations to current search technologies which will be outlined in this section to lay the foundation for describing an improved search system in the following section.

One of the biggest concerns about search engines is the fear of manipulation. This is owing to the fact that search engine providers do not publish their algorithms and reasoning behind rankings, also referred to as the black box nature of search engines. This was for instance pointed out by Google founders Brin and Page [BP98] who are now themselves being accused of manipulating search results [WM15].

Knowing that algorithms set one search engine provider apart from another, it is understandable that they treat their methods as trade secrets [Sul10]. However, the operators of search engines recognised this lack of trust and have started to fight the issue, as is evident in this Google blog post [Hufed].

Another major point of criticism is the fact that personalisation and socialisation, albeit offering chances for better results [Dop09], are commonly observed with scepticism. On the one hand, using such techniques results in a different Web for different users – at least a different presentation – [Mitedb], on the other hand, these methods raise privacy concerns because they build extensive user profiles [FKM+05; GBR09; Dop09]. Wiechert [Wie09], a leading German privacy officer, concludes that European privacy laws are violated by personalised search. As a result, he demands that companies improve their privacy standards, that internationally agreed-upon common privacy standards are enacted, and that existing privacy standards are updated to be applicable the Internet era.

A remaining technical issue is to improve Deep Web access (either through manual [BCGH10] or automated [MKK+08] training). Furthermore, systems capable of answering domain queries on multiple heterogeneous data sources such as "name all European universities that offer an information systems degree in cities larger than 40,000 inhabitants and within 50km to an airport" have to be developed.

In addition, Web spam, as described earlier, is a problem that is not solved and probably never will be, as spamming is an on-going competition between spammers and search engines, often referred to as the spam war [Dop09; BR10; Liu07]. This manifests itself in the observation that whenever search engines can detect one kind of spam another evolves. For this reason quality of algorithmic search engines cannot be trusted all of the time.

As a fact, only humans can tell whether information has the quality they want it to have. In this regard, the search computing paradigm is a sensible step to fight spam and consequently to increase the result quality. However, the initiators themselves state that it is all but clear whether there is demand for search computing, as most users are (most of the time) happy with the simple search offered by major search engines [BR11]. While Ceri [Cer10] recognised that generally special purpose search engines outperform general purpose search engines, the search computing paradigm has a relatively broad user base with heterogeneous interests as a target audience.

Finally, the search computing paradigm lacks clear incentives for the involved parties (data providers, developer, and domain experts). Furthermore, the loosely connected organisational structure that is imposed on the different participants is intuitively harder to manage than having a single organisational body.

3.5.2. Focus of this Part

It could be seen how important data and information are in the overall business processes of the present day. Quoting the somewhat self-ironic, but nevertheless true, computer science rule "garbage in, garbage out", it can be seen that only if high-quality information is provided to a business, it can acquire knowledge which eventually will lead to competitiveness and competitive advantages. However, this problem is not limited to corporations. Individuals need high-quality information, too, in order make decisions in their lives which could be rather trivial, such as where to go on holiday, or more serious, such as what pension fund to choose or which house to buy.

Furthermore, it has been pointed out that for many users the average search results, which are provided by general purpose search engines, are sufficient. However, there are certain niche domains – such as treatments for rare medical conditions – where this is not the case. Consequently, new approaches to searching have to focus on, and provide solutions for, users that are currently not able to satisfy their potentially complex information needs. Indeed, it can be said that the value of information largely stems from its customisation for a given purpose or a special interest group [GRC05]. From this, the main aim of this part of this work is deduced:

> *Develop a software artefact that answers a user's domain-specific informational queries levering curation to satisfy their high-quality information needs.*

The technical means to achieve this are, however, secondary. It is true that Google gained dominance after its introduction of innovative ranking mechanisms including PageRank. However, it can be argued that its utility rather than the actual underlying technology made users choose Google over its competitors. Evidence for this is given by a number of user quotes regarding Google in [LM06], with one referring to it as the Swiss Army knife.

In this regard, also the concept of transparency comes into play. While Google started with the vision that search processes should be made more transparent [BP98], they have since moved away from this point of view. As of today, they conceal their exact method and are very quiet about what factors they use [Sul10]. Furthermore, because of the latest European rulings, they are even legally forced to modify their results if a person's privacy is violated [TA14]. While this is favourable in the interest of privacy, it could also be seen as the beginning of censorship [Sch14]; at least the corresponding infrastructure is now provided.

In order to overcome the related lack of trust, the proposed solution builds on openness regarding user data usage. This means the search system will be designed in a way that users understand how it works and in what way their personal data is used. This openness can be underlined by the publications on the system mentioned above.

Regarding quality, it can be stated that throughout the existence of the Web, high-quality content was always connected to humans who select and organise the data underlying the IR process, which is now known as digital curation. It is self-evident that only humans can decide what is really relevant to them. As an example Rugge and Glossbrenner [RG97] stated that Gopher menus were superior to automated search engines owing to their curation. Similarly, Bachmann [Bac00] stated that finding high-quality information in the sheer mass of information available has to be done by information specialists.

This suggests – and thus the system will be developed that way – that high-quality sources should be tapped and curated to create a new single point of access to the entire knowledge of a domain in order to offer the required information at the right level of detail. In contrast to traditional search engines, the data will be stored in the curated state (including links to its original source) rather than just storing links to the original source which prevents the issue of having outdated links in an index. The main difference to the search computing paradigm is that the repository is centrally maintained rather than having the sources registering themselves with the repository.

In a sense the aim of this work is to re-include humans into the search process in order to satisfy demanding information needs in niche domains based on curation. This curation shall go beyond what has been done in the context of directories. The targeted form of information curation in the sense of selecting only high-quality information and extracting the essence is done by the aforementioned information brokers.

Therefore, the aim of the next chapter is to develop a generic process that – focusing on information quality – yields better search results for niche domains than automated search engines by exploiting human curation as an add-on to technology rather than on its own. This aim is not too far-fetched as it is in accordance with the observation reported on in Section 3.5.1 that principally special-purpose search engines can outperform general-purpose search approaches.

There have been approaches to achieve the same. One of these approaches was proposed by Sanderson et al. [SHL06] who suggest – focusing on a single domain – a curated harvesting approach. In their approach, predefined queries are sent to pre-registered services on a nightly basis. Then, relevant information is harvested and temporarily stored. Afterwards, the retrieved information is audited by data curators who decide upon its relevance and only relevant data is kept. Moreover, care is taken to prevent harvesting the same data more than once.

This approach is similar to data consolidation in the life sciences, as described, e. g., by Kulikova et al. [KAA+04]. For instance, different institutions contribute to a nucleotide sequence database. In order to ensure consistency across various instances of this database, data is synchronised on a nightly basis [KAA+04].

A similar harvest and curate approach was suggested by Lee et al. [LMHS09] with respect to their retrieval tool ContextMiner[38]. However, using this tool, users can configure key-word based searches on a number of sources, e. g., Twitter, but neither can they make these so-called campaigns available to others, nor can they include private data such as personal files.

The online search engine provider blekko also claim to provide innovative, categorised search results based on curated data. However, they do not provide any information of how this is achieved but simply state to achieve this based on proprietary technology and their own index. Similarly, Scoop.it![39] claim to be a collaborative search and curation platform. However, their focus is on social media publishing and curating content is to be used in social media advertising rather than for answering search queries. Moreover, they also do not provide any insights regarding their technology.

Regarding the involvement of humans in a curation process, *Data Tamer* [SBI+13] is similar to the approach introduced in this work as it also relies on humans in case of doubt. However, it is focused towards schema integration and entity consolidation and consequently operates on a database level, while

[38] http://contextminer.org, accessed: 2015-05-31.

[39] http://www.scoop.it/, accessed: 2015-05-31.

the approach suggested here is more high-level and more concerned with information rather than data. Similarly, [FOTB14] presents an idea for harvesting and curation. However, it focuses on linked data.

In contrast to all of them – in fact there is hardly any more in-depth research on combining data curation and IR – the approach described in the next chapter incorporates a number of sources including privately held data. Furthermore, a generally available repository is built, the search on which being enhanced by asking users to provide special interests in order to apply techniques from recommender systems when presenting the results. Additionally, information is presented in an interlinked manner so that searching and browsing are re-integrated on a single platform similar to the way a *memex* would present the information. Supporting the idea that this is a valuable approach to pursue is the fact that DARPA has recently made a call for advanced niche domain search systems [DAR14].

Most importantly, all search approaches discussed so far – including general search engines – are only accessible when being online and do not allow for results to be persisted offline. However, even in this day and age, there are times when an Internet connection cannot be guaranteed at all times, be it because of bad coverage in rural areas, because of air-travel where to date Internet is only sparsely available, or simply because of prohibitively expensive roaming fees. Thus, it is necessary to make information available offline and it is an equally important goal of the proposed information provisioning platform to do so even when no Internet connection is available. Hence, the proposed solution will from now on be referred to as *Web in your Pocket (WiPo)*. Interestingly, the thought of having the Web readily in a shirt pocket was already discussed by Brin et al. [BMPW98] when describing a subset of the Web that they used among other things to test PageRank. However, whether their suggestion that all human-generated text on the Web will fit on a device in a shirt pocket within a few decades is questionable by simply looking at the amount of content that is created each minute on the Web today presented in [Koo13]. For instance, Twitter processes 100.000 tweets in that time.

Other areas of research that will be touched upon, because they have been largely neglected so far, are the organisational form and business models of innovative search systems. Although it has been acknowledge that the Deep Web offers significant value [Cer10] little has been published regarding these topics; exceptions to this rule are [BBC+10; BCDP11].

3.5.3. Applied Method

According to the research methods literature, e. g., [Cro98; Pun05], a method has to follow the question that is sought to answer. It is the aim of this work to develop an alternative search process and to demonstrate its applicability to selected use case. To this end, the *Design Science* approach described by Peffers et al. [PTRC07] will be followed because it focuses on creating an artefact rather than on theorising. Based on the observation that even though Design Science was introduced to the *information systems community* in the 1990s (e. g., in [MS95]), no common framework for Design Science research existed and it was used and rectified on a case to case basis, Peffers et al. [PTRC07] developed a framework which has been highly recognised in the literature with more than 1,200 citations[40].

Based on an extensive literature review of prior work from which they chose seven representative papers, Peffers et al. [PTRC07] developed a 6-step-approach to Design Science and demonstrated its use in four case studies. The six phases of the approach are:

Identify Problem & Motivation is the first step and aims to identify and define the research problem and motivate why it is important to solve the identified issue.

Define Objectives of a Solution specifies how the problem described in step one can be feasibly solved. The objectives can be qualitative or quantitative in nature.

Design & Development can be seen as the core component of the overall Design Science process, where the actual "artefact" (i. e., an entity that represents a possible solution) is created; for instance this could be a model, a method, a construct, or a software.

Demonstration is the logical next step after the development and serves to demonstrate how the problem is solved using the developed artefact.

Evaluation attempts, then, to measure how well the artefact solves the problem. On a conceptual level, this step can make use of appropriate metrics which provide empirical evidence or logical proof.

Communication aims at making the knowledge gained in the process available to other researchers. This includes, among other things, the problem, the solution, and its effectiveness.

[40] According to: http://scholar.google.com/citations?user=zwOdzCgAAAAJ, accessed: 2015-05-31.

The overall process is depicted in Figure 3.5. At this point, it should be emphasised that the process includes a feedback loop, i. e., the results of later steps can be used to fuel a subsequent iteration of the process. Furthermore, the fact that the method is intended to support the creation of an artefact makes it the ideal method for this research project which is situated – as mentioned before – at the junction of management studies and computer science (see Figure 3.1).

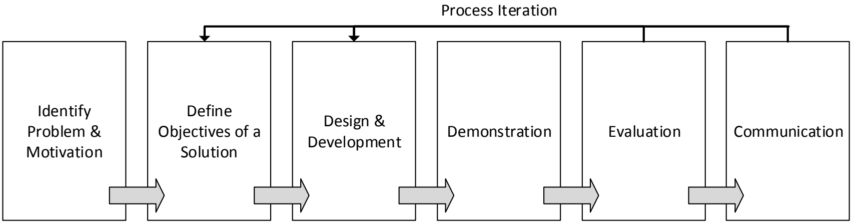


Figure 3.5.: The Design Science Research Process, Adapted from [PTRC07].

This work focuses on the first three steps of the process. In the preceding extensive literature review, the problems of insufficient information provisioning in niche domains, as well as search result offline availability, have been identified. These issues serve as *motivation* that was outlined in the previous section to build a new search model. The ideal solution, i. e., the *objectives*, is then highlighted at the beginning of Chapter 4. Then, the development of the artefact is described in Section 4.1 as a model and in Section 4.2 as actual software implementation. *Demonstration* and *Evaluation* have been conducted as part of the case comparison in Section 4.3 by the example case of New Zealand Land Search and Rescue (LandSAR)[41], the New Zealand search and rescue organisation. *Communication* has started with numerous works on the model [DSV12; DSVR13; DSVR13; DRSV], its implementation [SGH+14b; SGH+14a], and a demonstration [SGH+15]. These efforts peak for the time being in this thesis.

[41] https://www.landsar.org.nz/, accessed: 2015-05-31.

4. High-Quality Information Provisioning using WiPo

Until recently, encyclopaediae, maintained by a group of trustworthy people, were the single point of reference used by the majority of the population. They were maintained by a group of people who were believed to be trustworthy. However, encyclopaediae were expensive and became outdated at a rapid rate; particularly with the ever faster generation of knowledge over the last decades, resulting in a shorter half-life of knowledge.

As previously argued, the WWW has improved the availability of information tremendously. However, it has also decreased the trustworthiness and traceability of information. This has led to a situation in which a plenitude of available data has not yet been turned into information. In a corporate context, Gary Cokins coined the phrase "organizations are drowning in data but starving for information." [Tho13]. Similarly, Han and Kamber [HK01] pointed out in 2001 that data mining has gained attention in the information industry due to the need for turning vast amounts of data into useful information.

Having discussed the history and current state of search technology and outlined important shortcomings and possible solutions, this chapter is dedicated to exploring a new approach of extracting information from the sheer mass of data available online and providing users with exactly the information they need. While this can be seen as the motto of all search technology [BBB+11], here it will be pursued in light of a shift from document retrieval towards satisfying information needs. This approach follows an observation by Baeza-Yates and Raghavan [BR10], who stated that "People do not really want to search, they want to get tasks done". This phenomenon is also recognised in library and information science, where a focus shift from collections to users and their needs can also be observed [SC11].

In order to satisfy modern information needs, this work proposes to use an approach going back to the roots of information provisioning, by adopting the idea of the *Memex* created by Bush [Bus45] in 1945. A *Memex* is a machine in which humans can keep all information relevant to them in a central repository interconnected by trails or links. While search engines today provide at least a single point of entrance to the information available online, they do not generally provide the results in an interlinked manner but only point to indexed documents. This is supported by the observation that, for instance, GOOGLE's universal search still has limitations and mainly works for persons of public interest [HSBW13; HMW14].

Roughly based on the *Memex* idea, a new tool for information provisioning based on the concept of curation will be introduced in this thesis. To some degree, this is very similar to the idea of encyclopaediae, which have been manually assembled by a group of experts, or the idea of catalogues in the early days the Internet. However, in contrast to these concepts, this new tool is aimed at combining manual curation with existing Web retrieval techniques and recommender systems.

Combining the strength of these different approaches into one tool, an information platform can be realised that serves high-quality information needs, particularly if highly integrated and interlinked information is needed. That said, it should be emphasised that it is not the aim of this work to replace existing general purpose search engines as they are sufficient in a number of cases. Nevertheless, as has been pointed out in the literature, there are many queries for which the standard search engines of today do not suffice [Cer10; Dop09]. This work specifically focuses on scenarios in which, to date, topic-centric data collections do not exist or lack interconnectivity. Examples of such scenarios include search and rescue and tourism [DRSV14]; these cases as well as others will be elaborated upon in Section 4.3. This thesis proposes the Web in your Pocket (WiPo), a topic-centric curated information service configured to a user's needs, as a solution to address these issues.

The remainder of this chapter is structured as follows. Firstly, the WiPo approach will be outlined, based on and extending [DSV12], in Section 4.1. The next section will then show how the concept has been implemented by presenting the architecture of a prototypic implementation, based on and extending [SGH+14a; SGH+14b]. Next, Section 4.3 will discuss scenarios in which WiPo can be applied, extending [DRSV14] by enhancing the comparison model, adding a fourth case, and an in-depth evaluation of one particular case, namely LANDSAR. Finally, this chapter concludes with an outline of potential extensions to WiPo and how those parts of the concept that have not yet been implemented can improve information provisioning in the future.

4.1. The WiPo Approach

In essence, WiPo is a topic-centric information service, tailored precisely to a user's needs and budget, that is based on the curation of information sourced from the Internet, with the option of persisting information in an offline mode. With regard to the categorisations presented in the previous chapter, WiPo can be seen as special purpose search tool.

In contrast, in established search engines such as Google or Bing, search queries are run against a pre-assembled index [VH07; Lev10; LM06; Lew09a], which has been automatically gathered without knowing users' intentions. In these search engines pruning [MSF+05; SJPB08] reduces the result set to a manageable size focusing on satisfying average and not special information needs. Furthermore, most of the time, allowed user input is nothing more than keywords with a simple syntax or the restriction to specific domains or top-level domains. As discussed in the previous chapter, approaches do exist to capture a user's intuition; however, to date, there is no widely known and openly accessible approach incorporating personal, offline data with external, Web-sourced data.

In practical terms, WiPo users provide keywords, supply a list of relevant links and have the option to upload files through an easy-to-use GUI. Similar to traditional search engines a user query is run against an existing document base, the so-called *Curated Database*. In addition, WiPo uses all inputs provided by users to extend the curated database. To this end, given Web sources are crawled. The crawling results, together with documents supplied by the user, are then examined by experts in order to only store high-quality (i. e., correct and relevant) content in the curated database. This step is supported by data mining. As this approach targets niche domains an initial Web search using established search providers may be a starting point. However, in many cases such as healthcare, expert knowledge is important when integrating the results.

The final results are presented in an integrated manner that makes it easy to see relations between results. In that way, information presentation and interlinking of information by WiPo is somewhat similar to *Xanadu* [Nel99b; Nel08], a linkage based electronic documents system, in that both approaches are based on the observation that the Web in its current form does not suffice as an information repository in many cases. Unlike *Xanadu*, WiPo is built atop of the Web rather than as an alternative offline solution for text management. Furthermore, WiPo will not be limited to desktop computers and will be extended to all kinds of devices including mobile devices, enabling users to carry their information repository around, hence the name: Web in your Pocket (WiPo).

As pointed out in the previous chapter, particularly the integration of different sources is key to improving information provisioning. WiPo is unique in that it does not rely solely on public sources but allows for the inclusion of private sources, very much in the spirit of the *Memex* [Bus45], which was supposed to contain exclusively private information[1]. In that sense, WiPo can also be seen as a portable *Memex* for the Internet age. Indeed, the fact that DARPA has recently launched a programme, by the name of Memex, to fund the development of a search paradigm that allows for tailoring of indices and search results to subject areas and individual needs [DAR14], supports the rational that the WiPo approach is addressing a relevant problem.

High-quality information is achieved by exploiting curation, which, in the context of this thesis, shall mean the long-term selection, cleansing, enrichment, preservation, and assembly of data and information and their respective sources. Curation is well-known in museums where it is the task of curators to assemble and maintain collections and exhibitions. The concept, as defined above, has already been successfully applied to scientific data, for example the DCC.

Curation can be achieved by means of manual labour, which may be provided either by a number of selected individuals or an anonymous crowd. The first can be compared to an expert team creating an encyclopaedia, the latter to the processes on Wikipedia. Of course, curation is supported by algorithms in order to decrease the work load of curators.

Besides curation, the biggest difference between the WiPo approach and established search engines is a paradigm shift from *pull* (i. e., users need to request information) to *push* (i. e., users configure a service to their needs and are provided with updates whenever new information becomes available).

Another noteworthy difference lies in the kind of sources both approaches access. While for both the structure of the source does not matter, the way of accessing them does make a difference. Established search engines are limited to publicly available sources and have only just started to automatically gather data from semi-public sources such as online databases with subscriptions or online user groups with limited access, like forums. In contrast, WiPo is able to easily access semi-public sources as human experts will hand-pick the most relevant sources for a topic, and determine the best way to integrate them. In addition to that, WiPo also has the aim of accessing sources that have not yet been made available to the public, as well as allowing users to provide their own files. Therefore, unlike established search engines, WiPo will be able to access entirely private sources.

[1] This is little surprising given that it was described long before the advent of the Web.

Furthermore, WiPo allows for sophisticated personalisation going beyond what has been described for algorithmic search engines in Chapter 3. It enables users to configure their own explicit search profile which helps in finding relevant documents and improving ranking. Moreover, the WiPo service can even be configured so that users are informed whenever new information is available for them. Furthermore, the quality of the curation can eventually be made dependant on the price a user is willing to pay.

It should be pointed out that WiPo is not intended as a substitution for traditional search engines, its role is to provide more comprehensive information in areas were Web searches can be tedious and time-consuming. For such scenarios, WiPo offers a unique and innovative approach to information provisioning, compared to traditional Web search engines, an argument which is summarised in Table 4.1.

Table 4.1.: Comparison of WiPo and Web Search, Adapted from [DRSV14].

	Search	**WiPo**
General Approach	Crawling and automated indexing	Crawling and extensive curation
Data Sources	Public, semi-public	Public, semi-public, private
Type of Sources	Online only	Online and offline
Data Availability	Online only	Online and offline
Update Paradigm	Pull Only	Pull and push
Data Sophistication	Limited – one size fits all	Customizable, flexible
Data Quality Determinant	Algorithms	Algorithms, crowds, and human experts

In order to demonstrate the WiPo process at the end of this section, first a general introduction to information needs will be given and it will be discussed what kind of information needs can be addressed using WiPo. Based on this, the WiPo approach is presented in detail. While in the course of this work an extensive discussion of how WiPo can help in different scenarios is presented in Section 4.3, for the initial presentation tourists wishing to prepare for their holidays shall be considered as a running example because it can be grasped intuitively.

4.1.1. Quality Information Needs

The concept of *information needs* dates back to the 1960s when Taylor [Tay62], and others, investigated information systems. Within the literature it remains unclear as to where the human need for information initially stemmed from, with the general assumption being that humans just need information [Neh65] (as cited in [FE74]). However, information needs or data needs are commonly related to decision making [Mil12b].

An early overview of the general subject area of information and information needs was given by Faibisoff and Ely [FE74], in 1974. In their work they show that information needs can be seen as a general concept for both information demands and information desire. The first being a concrete articulation while the later may also simply be a feeling or even unknown to the user [FE74]. As this differentiation is not necessary with regard to this thesis, the more general term of information need will be used, supposing that users know that they have an information need and can articulate it in order to pose a query to a search engine or an information system in general.

Nehnevajsa [Neh65] (as cited in [FE74]) argues that the usefulness of information depends on four attributes. Information is useful if it:

1. is of the right *kind*,
2. has the right *quantity*,
3. has the right *accuracy*, and
4. has the right *timeliness*.

Decades later, using the terms data and information quality interchangeably, Wang and Strong [WS96] presented a well-recognised, comprehensive data quality framework based on two data consumer surveys. They arrived at four high-level data quality categories:

1. Intrinsic data quality, i. e., criteria that stem from the data itself, including *accuracy* and also *believability*
2. Contextual data quality, i. e., criteria that depend on the context, including *timeliness*, *relevancy* (which can be seen as equivalent to *the right kind*), and *amount* (which can be linked to *quantity*)
3. Representational data quality, i. e., criteria that show how well the data is represented, including *ease of understanding*
4. Accessibility data quality, i. e., criteria that depend on how well the data may be accessed and how securely.

It is the aim of WiPo to provide high-quality information by means of curation. The concept of data quality will be discussed in more detail in Chapter 6 in the context of data marketplaces where data quality will be used in order to price information goods. For now, this higher level of abstraction is sufficient. The last two points, namely *representational data quality* and *accessibility data quality*, can be viewed as dependent solely on the implementation but are rather subordinate in classifying information needs as it should be common practice to make search results available in an easily accessible and understandable manner. Furthermore, it is supposed that through the human involvement the *intrinsic data quality* is ensured by WiPo. Also, finding the right *kind* of, i. e., the *relevant*, information and ensuring its *accuracy* – here to mean correctness – can be seen as main curation tasks in order to achieve high-quality information. Thus, it is implied that information provided through WiPo satisfies these data quality criteria by means of human double-checking.

As a result, only some *contextual* quality criteria are left to differentiate customers' information needs. Namely, *quantity* and *timeliness* are variables to differentiate information needs. *Timeliness* can be interpreted as how recent a piece of information is or how often it is updated, referred to as *update frequency*. The information *quantity* can refer to the amount of information that is returned. Here, the number of sources that have been tapped and integrated regarding a certain topic will be used to address information quantity which is reasonable as more sources also increase the amount of information obtained. In the initial publication of WiPo [DSV12] the term *data broadness* was introduced to describe this concept, and this nomenclature shall be used from now on for consistency.

This results in a two-by-two-matrix[2] along the dimensions of *data broadness* and *update frequency* presented in Figure 4.1. The following provides examples for the four resulting scenarios (based on [DSV12]):

1. **Low update frequency and low data broadness**: For instance, air travellers who have an unexpected delay and wish to explore the city they are stranded in. They only need the relevant information once, i. e., no updates and one reliable source is enough as it seems to be an over-proportional effort to compare and consider plenty of sources.

2. **High update frequency and low data broadness**: For example, a small-time investor who wishes to keep abreast with recent stock market devel-

[2] Of course the dimensions are rather continuous than discrete but using two extreme values simplifies the description without losing any meaning.

opments and needs the most up-to-date price information, but again one reliable source is enough.

3. **Low update frequency and high data broadness**: This is the case for tourists planning a trip to a destination they have not been to before, wishing to gain information from a variety of sources such as travel agencies, tour providers, blog posts, video material, and map providers, to name but a few. Despite the diversity, the update frequency does not usually need to be high as a holiday is usually a one-off endeavour.

4. **High updated frequency and high data broadness**: For instance, LandSAR organisations which have comprehensive pre-planning for all sorts of rescue activities (e. g., children, tourists, dementia patients). Such pre-plans involve information on different areas such as accessibility of national parks, characteristics of the flora and fauna in the area as well as information on the types of people that commonly loose orientation, for instance information regarding dementia. Of course, this information has to be kept as recent as possible to be able to deal with unexpected or sudden operations.

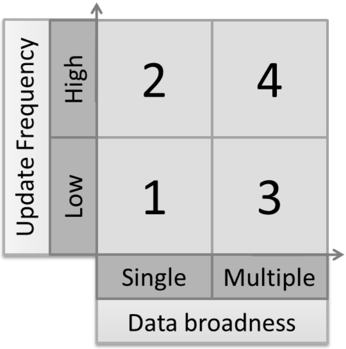


Figure 4.1.: Dimension of Information Needs, Adapted from [DSV12].

For cases 1 and 2, established search technology is sufficient most of the time; these users would query a search engine of their choice to find the one source that satisfies their needs. For cases 3 and 4, the situation is different. In these cases a number of sources have to be found, consulted, and combined to form a comprehensive picture, which can be – depending on the extent of the information needs – a very labourious task. This is further increased if the resultant body of knowledge has to be kept up-to-date. Consequently, cases 3 and 4 are cases where WiPo could be most useful. But even for a stranded air traveller WiPo can provide some benefits if no Internet connection is available. This, of course, requires the person to use a tourism WiPo on their mobile device with up-to-date information synchronised in advance. Besides the tourism and LANDSAR case, a healthcare example and a business case will serve as further examples in Section 4.3.

4.1.2. The WiPo Process

The WiPo process is generic in nature and not use case dependant. However, the actual implementation will always have to consider domain specifics and employ domain experts when being developed. Here, the overall process will be described from a searcher's point of view rather than from a curator's view because, although the process of curation is an important part of the overall process, the ultimate goal is to satisfy users' needs.

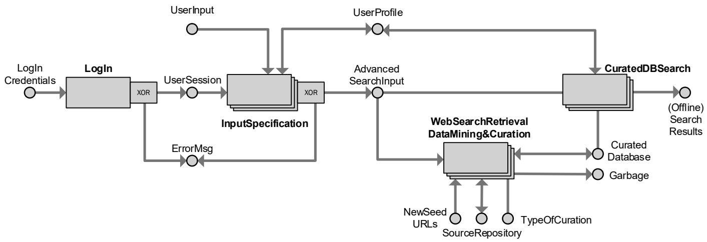


Figure 4.2.: Overall WiPo Process.

The overall WiPo process is demonstrated in Figure 4.2. First, users provide *LogInCredentials* – in form of a user name and password combination – which can be either wrong, resulting in an *ErrorMsg*, or correct, resulting in a *UserSession*. The *UserSession* is processed by the abstraction *InputSpecification* which also takes *UserInput* into account. If the *UserInput* is erroneous, an *ErrorMsg* is produced. Otherwise, *AdvancedSearchInput* is produced and the *UserProfile* is updated if necessary. *AdvancedSearchInput* is then read by the abstraction *CuratedDBSearch* which queries the curated DB and produces *(Offline)SearchResults*. It is then consumed by the abstraction *WebSearchRetrievalDataMining&Curation*, the core component of the WiPo-Process. In fact, it does not matter which of the two (*CuratedDBSearch* or *WebSearchRetrievalDataMining&Curation*) consumes the *AdvancedSearchInput*, it only has to be accessible by both to be eventually consumed. The *WebSearchRetrievalDataMining&Curation* component is dependent on the *TypeOfCuration* which determines whether curation is done automatically, by a human individual, or a crowd. It further takes *NewSeedURLs* as additional input and updates the *SourceRepository* as well as the *CuratedDatabase*. As not all retrieved content classifies as high-quality and ends up in the *CuratedDatabase*, it also produces some *Garbage*.

The WiPo process can be mapped to the Curation Lifecycle Model by the British DCC [DDCed] presented in Section 3.4.1. However, it was not modelled on it, as it is the primary aim of WiPo to curate information on a topic of interest – not archive data. The mapping of WiPo to the Curation Lifecycle Model can be found in Table 4.2. However, some of the tasks are only represented in refinements.

Table 4.2.: Comparison between the WiPo Process and the Curation Lifecycle Model.

WiPo	Curation Lifecycle Model [DDCed]
Web Crawling	Create or Receive
Curation	Appraise & Select
ContentExtraction and Curation	Ingest
Curation	Preservation Action
CuratedDatabase	Store
CuratedDBSearch	Access, Use, Reuse
Curation	Transform

Having outlined, the overall process in the preceding paragraphs, now, all abstractions will be discussed in more detail.

InputSpecification. Once users are logged in, they have to provide *UserInput* through a suitable GUI. Unlike traditional search engines, which commonly only support keywords and the ability to only search a given domain, WiPo supports four types of *ValidSearchInput*. Besides *Keyword* (in a given query syntax), these are a list of *URLs*, personal *Files* or *UserProfileData*. In the case of *UserProfileData*, the *UserProfile* is updated. This input is first checked for validity by testing if *keywords* are present and form a valid query or that supplied *URLs* are valid. If the *UserInput* is not valid, an *ErrorMsg* is produced. In a valid case, users either provided *ValidSearchInput*, i. e., actually wanted to perform a search, or *ValidUserProfileData* to later enhance their search results by telling the system to particularly look for content related to their personal interests. The *ValidUserProfileData* is matched against the *UserSession* and the according *UserProfile* is updated. *ValidSearchInput*, which contains at least *Keywords* but possibly also *URLs* and *Files*, is processed by the abstraction *PrepareSearchInput* which produces a *SearchInput*-document. This document is taken and combined with relevant information from the *UserProfile* to form *AdvancedSearchInput*. The input specification process is illustrated in Figure 4.3.

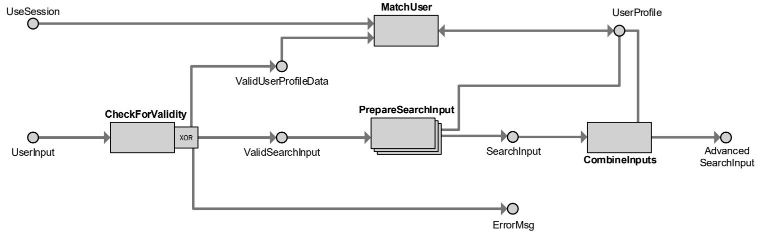


Figure 4.3.: WiPo: Input Specification.

PrepareSearchInput, as depicted in Figure 4.4, does the following: First, the *ValidSearchInput* document is split and the three types of input are processed. *URLs* are crawled and the results are fed into the *StandardiseDocuments*-transition. Furthermore, they are consumed by the *CombineInputs*-transition which creates the combined *SearchInput* of all input types. *Files* undergo the same *StandardiseDocuments*-process where they are transformed into a standard format, e. g., XML or key-value-pairs in JSON. These *StandardisedDocuments* are then processed, together with the *UserProfile* in order to determine *SearchTopics* – i. e., topics of interest to the user –, and kept to become part of the *SearchInput*.

Finally, *Keywords*, i. e., the users' search query, will also be used to determine *SearchTopics* and also to set *Pre-Filters* which limit search results by time, locality or similar. Eventually, *Keywords* and *Pre-Filters* are included in the *SearchInput*, too.

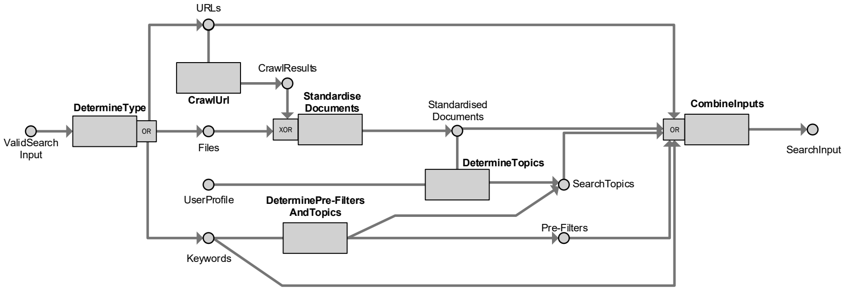


Figure 4.4.: WiPo: Search Input Preparation.

> **Example**
>
> A tourist is searching for an adventure hiking trip through the New Zealand outback. They supply the search string "one week hiking trip NZ North Island bush", provide a list of New Zealand tourism URLs they consider helpful, and a holiday diary entry from a friend who recently did a similar trip. From the files (diary), the keywords and the crawled websites, the topics *NZ bush, hiking,* and *canoeing* can be extracted. From the keywords pre-filters are derived that set the duration to approximately one week and the location to New Zealand's North Island. All of this information is then forwarded to the next processing steps.

WebSearchRetrievalDataMining&Curation. This is the abstraction containing WiPo's core components – as depicted in Figure 4.5 –, namely *SourceSelection,* which provides *EnhancedSelectedSources* based on *AdvancedSearchInput* as well as updates the *SourceRepository*; *WebRetrieval,* which takes *EnhancedSelectedSources* to deliver *CrawlResults*; *DataMining,* which transforms *CrawlResults* into *DataMiningResults*; and *Curation*, which creates curated documents and updates the *CuratedDatabase* based on the *DataMiningResults* and the *TypeOfCuration*. Furthermore, *Curation* takes *NewSeedURLs* as input and updates the *SourceRepository*. Since not all content created is eventually turned into documents for the *CuratedDatabase*, some *Garbage* is also produced. To keep the *CuratedDatabase* up-to-date, the auxiliary transition, *Maintenance*, regularly re-feeds URLs from the *SourceRepository* into the whole process as *EnhancedSelectedSources* in order to keep the content of the *CuratedDatabase* as recent as possible. To this end, it regularly checks the timestamps of sources, using their last crawl date, and adds them as *EnhancedSelectedSources* if they are above a certain threshold. The check for timestamps can be implemented as a filter known from XML nets, as discussed in Section 2.1.

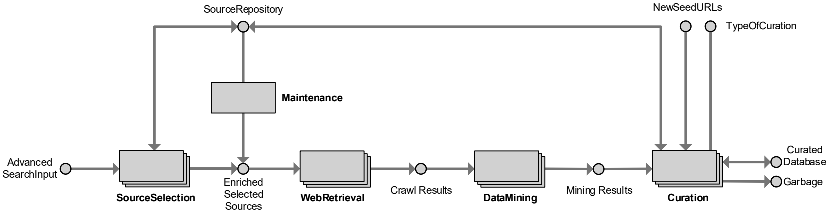


Figure 4.5.: WiPo: Web Search, Retrieval, Data Mining & Curation.

The first step of *WebSearchRetrievalDataMining&Curation* is *SourceSelection*, which is used to determine the most suitable Web sources from the *SourceRepository* in order to satisfy the information needs expressed by the user. As can be seen in Figure 4.6, this is achieved using *SearchTopics*, the *UserProfile*, *Keywords*, and *Pre-Filters*, all of which are previously compiled into *AdvancedSearchInput*. These pieces of information are then combined into one *AdditionalInfo*-document. While *SelectSources* only reads the tokens, *CombineToInfo* actually consumes them. Thus, it has to be programmatically ensured that only information that has been read can be consumed.

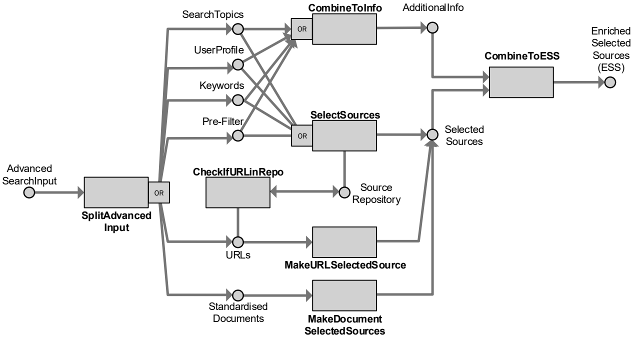


Figure 4.6.: WiPo: Source Selection.

StandardisedDocuments are directly transformed into *SelectSources*, while URLs are checked first to see whether they already are contained in the *SourceRe-*

pository and then added if required. Only then, they are transformed into *SelectSources*. At this point, it is worth noting that URLs can be contained in two ways in *SelectSources*, once as URL and once as standardised document (i. e., the document behind the URL). Even though this might be confusing at first, it is sensible as, in this way, the document available at the URL can be used for an in-depth analysis during *DataMining* (see Figure 4.8) while the URL itself can be used as a seed for focused crawling going beyond the actual document (see Figure 4.7). Finally, *SelectSources* are combined with *AdditionalInfo* to form *EnhancedSelectedSources(ESS)*.

> **Example**
>
> At the end of this step, the *EnhancedSelectedSources* contain URLs from the system – for example tourism websites as well as general hiking and canoeing sites – as well as the user supplied URLs and standardised user files.

Next is *WebRetrieval*. Here, *URLs* and *StandardisedDocuments* – if present – are separated from *AdditionalInfo*, which is always present, thus, an inclusive OR operation. Available URLs undergo a *FocusedWebCrawling*, which also follows out-links of the given URL. In this way, it creates a much broader basis than the individual crawl carried out in *PrepareSearchInput* (see Figure 4.4). Furthermore, *FocusedWebCrawling* takes the *AdditionalInfo* into account and is therefore much more focused on the actual query and the user preferences. Moreover, as a side-effect, *FocusedWebCrawling* creates meta data for the *SourceRepository* such as crawl timestamps. Implicitly, *FocusedWebCrawling* returns results in the same format as all other *StandardisedDocuments*. If some of those have been passed to *WebRetrieval*, they are joined with the crawl results.

From all these documents, the content is extracted and combined with the original *StandardisedDocuments* and the *AdditionalInfo* to create the *CrawlResult*-tokens. *StandardisedDocuments* are kept as there is a possibility that further data mining steps can make use of them. *Content*, in this context, refers to the actual content of the documents as well as meaningful annotations such as tags and a reference to the original source, but not to other meta data contained in the document, such as menu structures of websites. All this is illustrated in Figure 4.7, where it can also be seen that the two main transactions of this process can be refined; this, however, is use-case-dependent and has therefore been omitted in this instance.

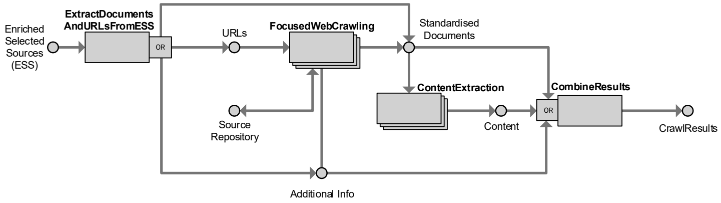


Figure 4.7.: WiPo: Web Retrieval.

As soon as *CrawlResult*-documents are available, *DataMining*, as depicted in Figure 4.8, takes over. Obviously, it is not feasible to show all possible data mining techniques in this process. As a consequence, *Clustering, LinkAnalysis,* and *EntityExtraction* have been chosen as examples, but other methods are also possible alternatives. In this case, they would just be added parallel to the depicted methods. Additionally, it would be possible to conduct some steps in sequence. As with the other steps, the input document (in this case *CrawlResults*) is split in its relevant parts. Then, *StandardisedDocuments* go directly into *LinkAnalysis* resulting in *CurationMetaInformation*, i. e., information that is valuable for curators but is not a candidate in itself. For instance, link analysis results tell curators how documents and entities relate to each other but this is not information that can be viewed as a document on its own. Similarly, *AdditionalInfo* is transformed into *CurationMetaInformation*. Before this is done, *AdditionalInfo* and *Content* are processed by *Clustering, EntityExtraction,* and *CreateCandidateFromContent*. While the latter simply takes the content and transforms it into a *Candidate*, *EntityExtraction* analyses the content and tries to find meaningful entities and descriptions which become *Candidates. Clustering* reads the same inputs but produces *CurationMetaInformation* rather than actual *Candidates.*

Content and *AdditionalInfo* are only read by transitions but never consumed to ensure the information is available to all *DataMining* steps that need it. To prevent documents piling up in these places, a garbage collection has to run in the background, which has been omitted in the illustration for simplicity reasons. Finally, *Candidates* and *CurationMetaInformation* are then combined to form *MiningResults* which are then passed on.

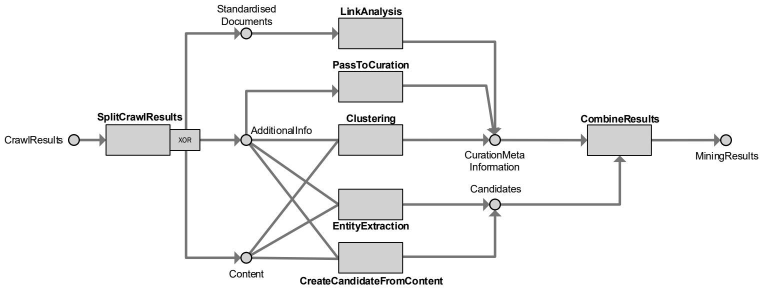


Figure 4.8.: WiPo: Data Mining.

> **Example**
>
> After crawling the selected sources and retrieving the documents, content about hiking and adventure tour providers can be extracted and *Candidates* derived. *EntityExtraction* extracts individual providers and their services. Furthermore, a *LinkAnalysis* provides insights regarding the importance of individual providers.

Last in line is *Curation* (illustrated in Figure 4.9), which is highly dependent on the *TypeOfCuration*. As pointed out previously, *Curation* can be done either manually, by individual experts or a crowd, or algorithmically. However, this differentiation mostly affects the actual implementation rather than the overall processes. This is made evident in the illustration by indicating that all transitions have a refinement that takes the *Type of Curation* into account. However, since these refinements are mainly concerned with the question of which user interface to present and when to apply which algorithm, this does not help the understanding of the overall process flow and is therefore omitted. For now, the *TypeOfCuration* will be seen as externally given and depending on the use case.

One curation task independent of the actual search process is the *AddToRepository* transition, the task of which is to add *NewSeedURLs* to the repository that can then be automatically crawled through the *Maintenance* mechanism, as shown in Figure 4.5.

When *MiningResults* arrive from *DataMining*, the first action to be taken is to check their quality, which can either lead to a *GoodCandidate* or to *Garbage*. Similarly, *GoodCandidates* and sources form the *SourceRepository* are re-checked from time to time. This process step ensures that the same or very similar content is not entered as candidate more than once. To keep the illustration in Figure 4.9 simple, it is supposed that both *CurationMetaInformation* and *Candidates* are contained in the *GoodCandidates* tokens. As an alternative, they could be passed on after a quality check into their own places, which in turn would be connected to all curation transitions that now only one place is connected to.

GoodCandidates serve as input for the actual curation tasks, leading to curated documents stored in the *CuratedDatabase*. These tasks are *CreateCuratedDocument* and *CombineManyCandidatesToCuratedDocuments*, as well as *AddContentToExistingDocument*. The first two are put-connections, whereas the latter is an update-connection to the *CuratedDatabase*.

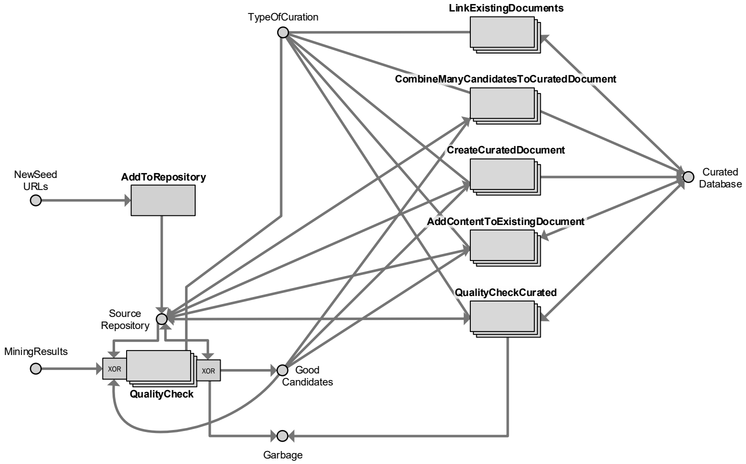


Figure 4.9.: WiPo: Curation.

It should be pointed out that candidates are only read by these transitions, which means they can be used multiple times. To ensure they do not become outdated in the course of time they are regularly fed back into *QualityCheck*. This can be considered a timestamp-based filter on the connection from *GoodCandidates* to *QualityCheck*.

There are two more curation tasks: *LinkExistingDocuments*[3] and *QualityCheckCurated*, which do not involve *GoodCandidates*. These are mainly concerned with managing curated documents in the *CuratedDatabase*. *LinkExistingDocuments*, the step in which documents that are related are also linked,

[3] Linking, as an example, takes *CurationMetaInformation* into account which is not depicted for the sake of a good overview.

updates the curated database, while *QualityCheckCurated* is the only step to remove objects from the database. Conversely, *Linking* is the only transition that does not write back meta data about curated objects, such as used parts of candidates and the like, to the *SourceRepository*.

> **Example**
>
> Some of the obtained results are discarded because they do not meet the level of expectation. All other results, become *Good Candidates*. Subsequently, an expert creates new documents for each tour provider discovered, as well as a single document containing links to the respective individual documents. Also, documents about canoeing and hiking in general are created during curation.

CuratedDBSearch Last in the sequence is the actual *CuratedDBSearch* presented in Figure 4.10. This step takes the *AdvancedSearchInput* and converts it into a query. For instance, *Keywords* could be the basis for the query, enhanced by the calculated *Pre-Filters* or some *UserProfile* information, such as personal preferences. The composed query is then run against an *Index* built by an *AutomatedIndexer* based on all documents within the *CuratedDatabase*. The result of this step is *DocumentIDs*. Next, the relevant documents are pulled from the *CuratedDatabase* and form the *PossibleResults*. These undergo a *Ranking* based on the *UserProfile*, leading to *SearchResults*. The *Ranking*, detailed in Figure 4.23 in Section 4.2.4, is strongly implementation dependent. Users can then provide feedback which implicitly updates their preferences. Furthermore, users can choose to synchronise selected results – or even all of them – with their mobile device in order to access them when being offline, leading to *(Offline)SearchResults*.

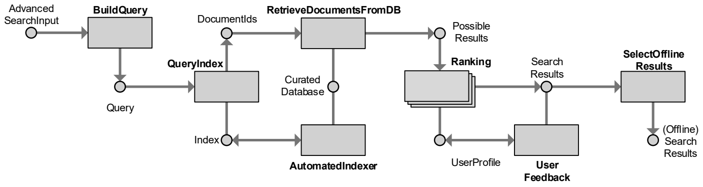


Figure 4.10.: WiPo: Search.

Example

Given the user query “one week hiking trip NZ North Island bush”, the index would return *DocumentIDs*, and eventually documents, on hiking tours as well as additional documents on the New Zealand bush such as national parks on the North Island, retrieved from the *Curated Database*. These include, for instance, the just created document containing all tour providers and also individual documents, containing particularly interesting tour providers separately. Since the user has previously made explicit to the system that they are interested in the Maori culture, tours containing Maori culture will be ranked higher in the *Ranking* step, despite “Maori” not being contained in the query. The results are ranked and presented to the user, who can then rate the results in order to provide feedback to the system by explicating their linking of the results. Just before boarding the plane to New Zealand the user opts to download all potentially interesting documents to their mobile device to continue reading on their way.

4.2. The WiPo Architecture

This section is dedicated to describing the current implementation of WiPo and parts of this section have been previously published in [SGH+14b; SGH+14a; SGH+15]. In order to keep the implementation project to a manageable size, the first prototypical implementation has some restrictions compared to the processes described in the previous section. While the overall WiPo process has been implemented, some transitions and refinements have only been realised rudimentarily, sometimes omitting functionality. This is evident from the illustration in Figure 4.11, where parts that are not fully functional yet are coloured in light grey. On the highest level, the only indicator for this is the fact that *TypeOfCuration* is not yet existent, therefore, manual curation by an expert is presumed throughout the entire process of curation.

The remainder of this section will focus on the actual implementation and will refer to the process view, only when appropriate. Furthermore, this section focuses on the architecture and interrelation of the used components rather than on implementation details, such as source code. Finally, one should bear in mind that the implementation of a working prototype for further experiments was done in a way that aimed at reusing as many existing tools as sensibly possible.

The overall WiPo architecture – depicted in Figure 4.12[4] – follows the client server principle. To this end, a Representational State Transfer (REST) Web interface (4) has been developed that can be accessed by an arbitrary application (2). For the first prototype a decision was taken to implement a browser-based client (1) which is fed by a server-side Web GUI dispatching module (3) because of the wide-spread use of browsers today.

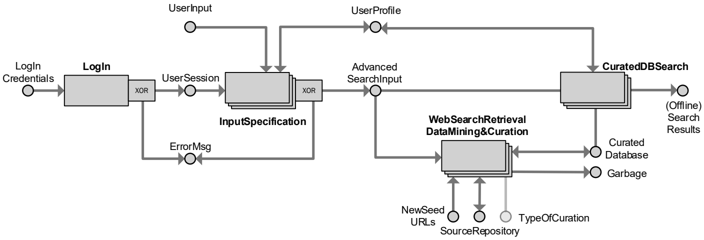


Figure 4.11.: The WiPo Process as Implemented.

The server infrastructure, which has been built in a modular fashion to provide the functionality as outlined in Section 4.1, can be accessed through the Web interface (4) which is supported by user management (5). The core component – as shown in Figure 4.12 – is the curated database (11). It is fed by curation (7), which in turn receives input from the candidate database (10). This database is fed by data collection (9), composed of scheduler, crawler, data extraction, and the source repository. Providing search functionality to users is done by the search module (6) which relies on an index (8) that is constantly updated from the curated database (11).

In order to describe this unique and complex system architecture, the client and its connection to the server will first be elaborated upon. Next, data collection is discussed, followed by an outline of the implementation of curation, before the section concludes with a detailed description of implemented search functionality.

4.2.1. Client-Side

The client side serves as a means for both searchers and curators to communicate with the system. It is accessed through the GUI, the main task of which is to simplify user interactions with the rather complex underlying infrastructure.

[4] Numbers in parentheses in the following description refer to Figure 4.12.

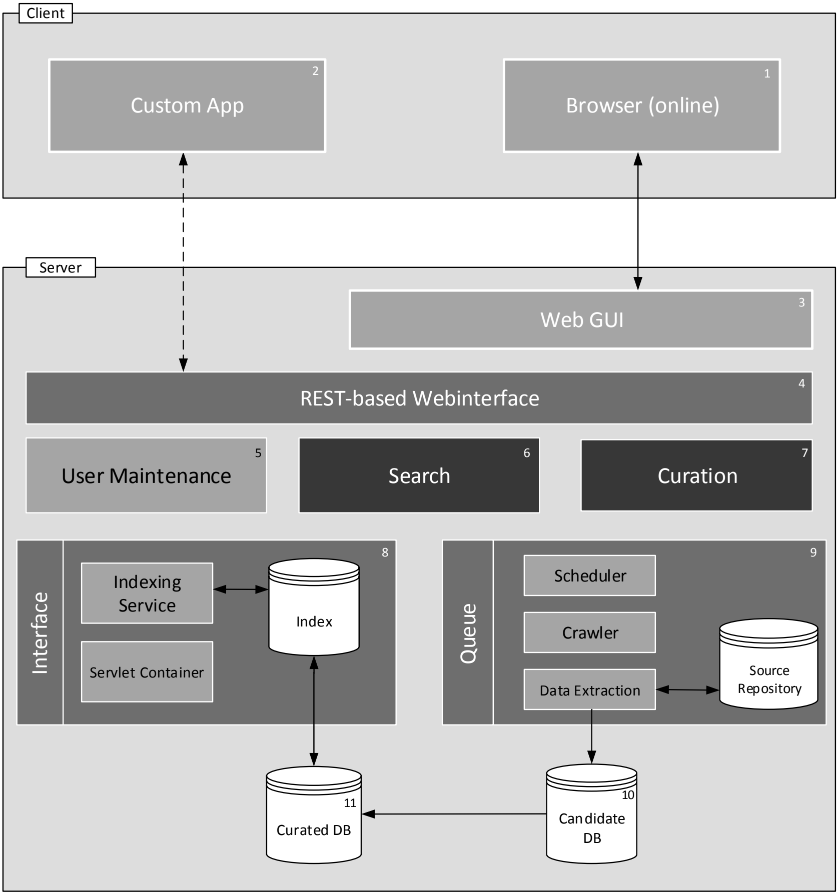


Figure 4.12.: WiPo Module Overview, Adapted from [SGH+14a] (numbers are explained in the text).

The prototypical implementation uses a browser-based Web GUI (1 & 3), rather than building the entire software client within the browser, for instance, using JavaScript. A server-side PHP Hypertext Preprocessor (PHP)[5]-proxy

[5] http://php.net/, accessed: 2015-05-31.

using Kohana[6] and the PHP implementation of libcurl[7] was developed to ensure compliance with the same-origin-policy enforced by JavaScript while allowing for the GUI and the WiPo core system to reside on different machines, which is necessary for the scalability of the system.

For browser implementation, the standard tools HTML and JavaScript were used, in particular the following libraries and frameworks: jQuery[8], jQuery UI[9], and Bootstrap[10]. Nevertheless, as pointed out earlier, an arbitrary application (2) can be developed as a user front-end, as long as it is able to communicate with REST Web services implemented in the Web Interface (4).

The REST Web interface (4) is the interlink between any client and the core system. As such, its main task is to forward requests from clients to the according internal module. It has been implemented as Java servlets, using Tomcat[11] as a servlet container and the Jersey[12] framework for implementing RESTful Web services in Java.

The User Management (5) is an adjacent module to the Web Interface. It takes care of registering new users and creating all necessary entries, such as profile data in an internally maintained user database, implemented using PostgreSQL[13]. Subsequently, user access and session management (using Tomcat), as well as user maintenance, are also done by this module.

4.2.2. Data Collection

Data Collection (9) is part of the *WebSearchRetrievalDataMining&Curation* step of the overall process. Similar to it, the *WebSearchRetrievalDataMining&Curation* step, the refinement of which is depicted in Figure 4.13, is implemented but limited to manual curation. However, this is only relevant for the curation step that will be discussed in Section 4.2.3. More precisely, Data Collection is equivalent to the steps *Web Search, Retrieval, Data Mining* of Figure 4.5 or Figure 4.13, respectively.

In order to be able to explain the implementation of data collection, it is important to first describe how the Input is generated. *InputSpecification* as a whole has been implemented but it does not yet consider the *UserProfile* as part of the

[6] http://kohanaframework.org/, accessed: 2015-05-31.

[7] http://curl.haxx.se/libcurl/, accessed: 2015-05-31.

[8] http://www.jquery.com, accessed: 2015-05-31.

[9] http://http://www.jqeryui.com, accessed: 2015-05-31.

[10] http://getbootstrap.com/, accessed: 2015-05-31.

[11] http://tomcat.apache.org/, accessed: 2015-05-31.

[12] http://jersey.java.net/, accessed: 2015-05-31.

[13] http://www.postgresql.org/, accessed: 2015-05-31.

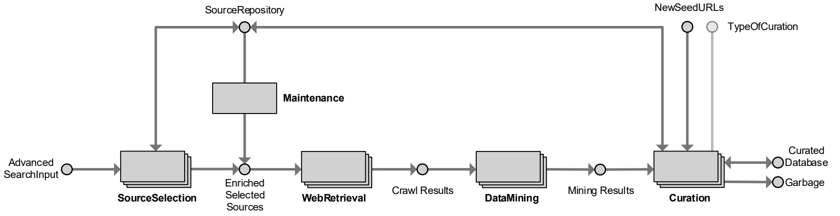


Figure 4.13.: WiPo: Web Search, Retrieval, Data Mining & Curation as Implemented.

AdvancedSearchInput as depicted in Figure 4.14. However, the profile can be updated for latter use in the ranking step (see Figure 4.23 in Section 4.2.4). There are also limitations to the supported input types; thus, *PrepareSearchInput* has been implemented in a scaffold manner, which causes some limitations shown in Figure 4.15. As of now, *Keywords* and *URLs* are supported but not *Files*. Furthermore, neither are URLs pre-crawled nor are keywords pre-processed. This also implies that no *StandardisedDocuments* are created and that *SearchTopics* are not yet being determined. As a consequence, there is no differentiation between *SearchInput* and *AdvancedSearchInput*.

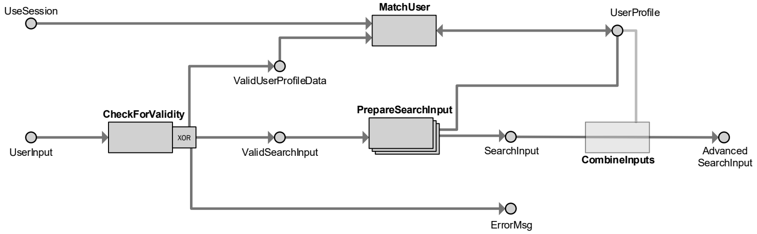


Figure 4.14.: WiPo Input Preparation as Implemented.

Consequently, only *Keywords* and *URLs* can be considered when selecting sources. Indeed, the *SourceSelection* is limited to adding URLs to the repository and choosing them as *SelectedSources*. Currently, URLs are added to the source repository, if not already contained, and are then piped to the *DataMining* process. No other selection of sources is in place at the moment. As a consequence, *EnhancedSelectedSources* are in fact not enhanced at the moment but the term is used for consistency (see Figure 4.16).

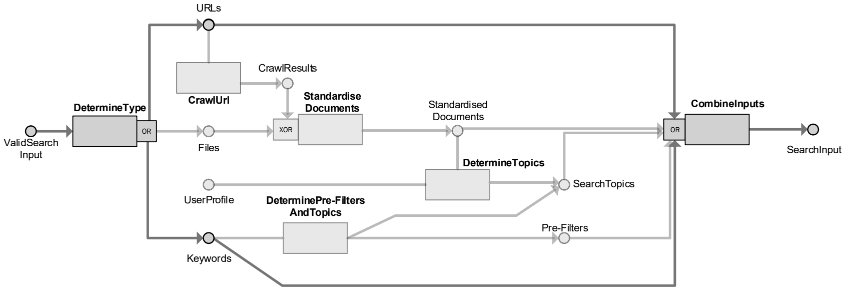


Figure 4.15.: WiPo: Search Input Preparation as Implemented.

To obtain documents from selected source URLs, the crawler Apache Nutch[14] is used for *FocusedWebCrawling* (see Figure 4.17), i. e., it downloads websites and stores them as dump files on the WiPo server. Given that selected sources are not only user-provided but can also be sources that should be updated, the whole crawling process is managed by a scheduler daemon that creates so-called *CrawlJobs*, which are then executed by the crawler. However, this additional meta information (last crawl time, etc.) is contained within the *SourceRepository*. Thus, the crawl is not focused by *AdditionalInfo*.

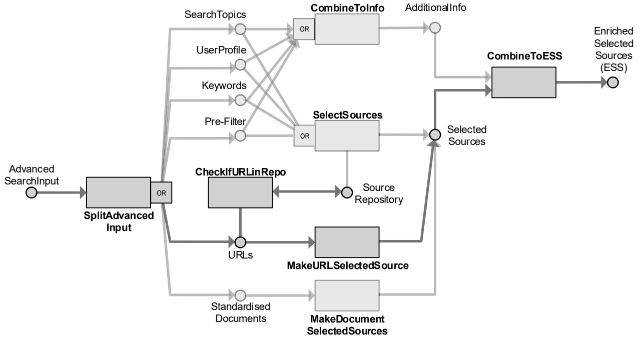


Figure 4.16.: WiPo Source Selection as Implemented.

[14] https://Nutch.apache.org/, accessed: 2015-05-31.

Given that the crawler is not focused in the sense that it decides autonomously on the relevance of crawled documents, it is depicted here as a standard transition rather than as an abstraction (cf. Figure 4.7).

The daemon differentiates between three types of *CrawlJobs*: *UserCrawlJobs, ReCrawlJobs,* and *DataCrawlJobs*. The latter two are automated using the Quartz Scheduler[15], which corresponds to the maintenance step from Figure 4.13. All *CrawlJobs* are managed in a priority queue and are presented here adapted from [SGH+14b]:

UserCrawlJobs are created when users add one or more URLs to their search which are unknown to the system (*CheckIfURLInRepo* in Figure 4.16). Once received, the scheduler creates a *UserCrawlJob* for these URLs with the highest priority. This implies that *UserCrawlJobs* are executed before any other *CrawlJob* to ensure that users receive their desired information as quickly as possible.

ReCrawlJobs are used to update outdated data – based on the last crawl timestamp longer than a pre-set threshold. All outdated URLs are added to the *ReCrawlJobs* which is then added to the scheduler's priority queue with medium priority.

DataCrawlJobs are used to enrich the WiPo database with new information. They have the lowest priority of all CrawlJobs.

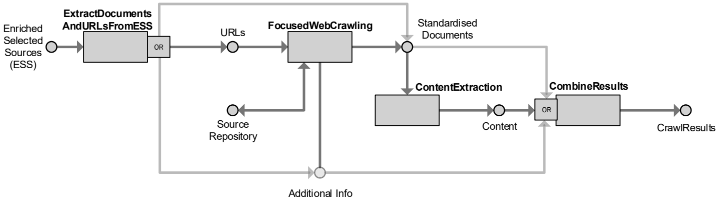


Figure 4.17.: WiPo: Web Retrieval as Implemented.

The first two *CrawlJobs* have a depth of 1, which means that only the given URLs are crawled. In contrast to this, *DataCrawlJobs* have a configurable depth greater than 1, which means that out links from the provided URLs are also

[15] http://quartz-scheduler.org/, accessed: 2015-05-31.

crawled. Given that there are no search topics, this crawling cannot be considered focused in any way other than being steered by the URLs fed to the crawler. Yet, as a step of the crawling process, meta data is extracted, which includes, for example, fetch time and last modified date. This meta information is stored in a PostgreSQL database known as the source repository from the Petri Net representation (see Figure 4.2). After the crawler successfully downloads a page its content is extracted. Within this step, the content is stripped from all unnecessary information including mark-up languages. The biggest challenge in this process is the extraction of content from HTML irrespective of whether it complies with World Wide Web Consortium (W3C) standards. A further challenge is that Web pages commonly contain menu structures, advertisements and similar non-content-related data. In the first prototype, Boilerpipe[16] was used to detect and remove non-content. Given its simplicity, it was decided to depict this transition as a standard transition, similar to *FocusedWebCrawling*.

The *CrawlType* is also used to prioritise a curator's work, similar to the way crawls of different types are executed with different priorities. To achieve this prioritisation, results from different *CrawlTypes* are presented separately to curators so that they can attend to user requests first and to other tasks later.

Furthermore, *DataMining* has not been fully developed yet and currently, *CurationMetaInformation* is not being generated. As depicted in Figure 4.18, from the process point of view, *DataMining* is restricted to create candidate documents from *CrawlResults* which is the only output of this process.

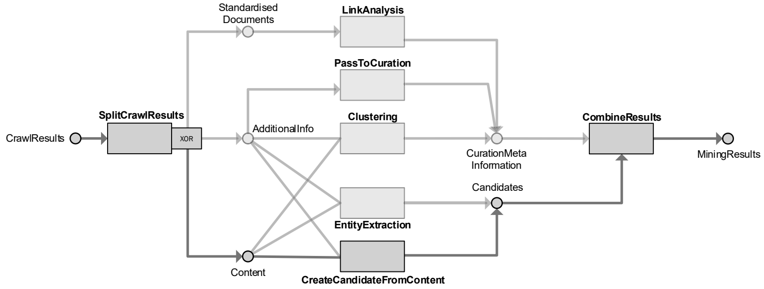


Figure 4.18.: WiPo: Data Mining as Implemented.

[16] http://code.google.com/p/boilerpipe/, accessed: 2015-05-31.

4.2.3. Curation

The *WebSearchRetrievalDataMining&Curation* process, depicted in Figure 4.13, shows that the curation module (7) is the link between *MiningResults* in the form of candidates (candidate database; 10) and the curated databases (11). It enables curators to build curated documents from candidates in a number of ways (see Figure 4.19) by interacting with an advanced GUI, illustrated in Figure 4.20. As a side product of curation, meta data is generated during these steps and stored in the source repository.

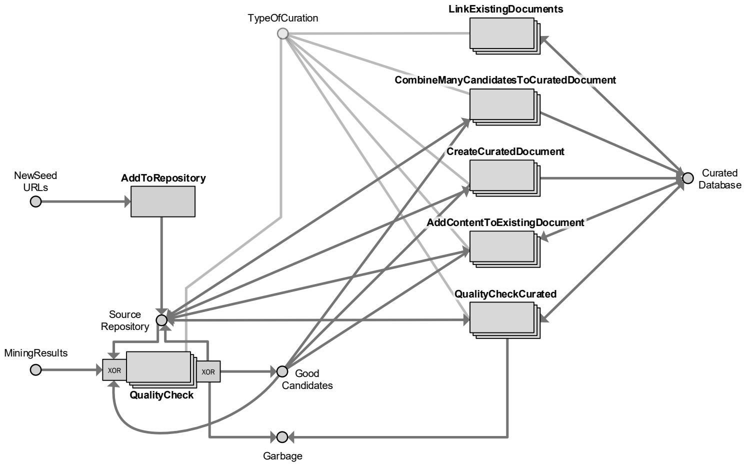


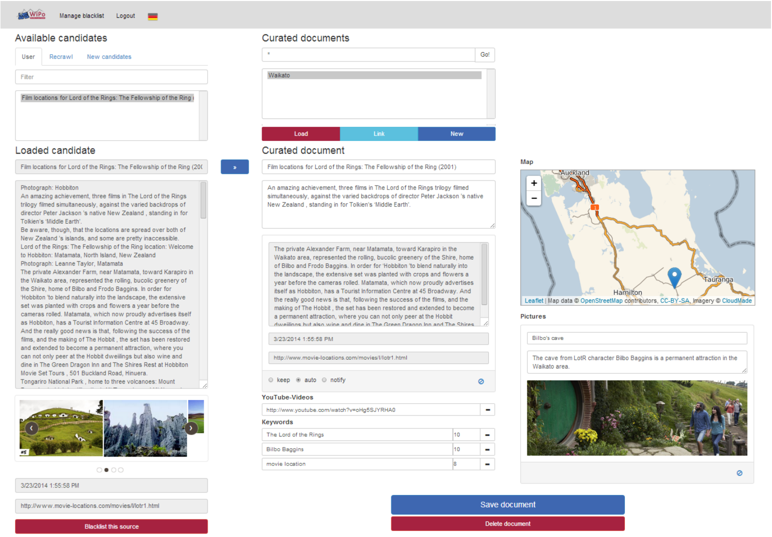
Figure 4.19.: WiPo: Curation as Implemented.

Curation is nearly completely implemented as presented in Figure 4.19, which fulfils the main aim of the prototype development. However, there are two minor limitations; *CurationMetaInformation* cannot be considered because it is not provided by previous steps and, at present, only human expert curation is implemented, despite the overall concept incorporating other forms of curation, such as the crowd or algorithmic curation. Thus, at the moment most of the tasks are not supported by algorithms, which is a number one priority for future work.

Candidate and curated database are both instances of MONGODB[17], a schema free document-oriented Not only SQL (NoSQL) database, with a JSON document structure used for both candidate and curated documents. All curation tasks –

[17] https://www.mongodb.org/, accessed: 2015-05-31.

such as creating curated documents, quality checks on candidate and curated documents, extending curated documents with new content, and linking existing documents – are implemented by this module. When creating new curated documents – whether they come from an individual or from multiple sources – curators have two options. Either they use the entire document or only relevant paragraphs, referred to as sections.



Available candidates

Filter

Film locations for Lord of the Rings: The Fellowship of the Ring (

Loaded candidate

Film locations for Lord of the Rings: The Fellowship of the Ring (20(

Photograph: Hobbiton
An amazing achievement, three films in The Lord of the Rings trilogy filmed simultaneously, against the varied backdrops of director Peter Jackson 's native New Zealand , standing in for Tolkien's 'Middle Earth'.
Be aware, though, that the locations are spread over both of New Zealand 's islands, and some are pretty inaccessible.
Lord of the Rings: The Fellowship of the Ring location: Welcome to Hobbiton: Matamata, North Island, New Zealand
Photograph: Leanne Taylor, Matamata
The private Alexander Farm, near Matamata, toward Karapiro in the Waikato area, represented the rolling, bucolic greenery of the Shire, home of Bilbo and Frodo Baggins. In order for 'Hobbiton 'to blend naturally into the landscape, the extensive set was planted with crops and flowers a year before the cameras rolled. Matamata, which now proudly advertises itself as Hobbiton, has a Tourist Information Centre at 45 Broadway. And the really good news is that, following the success of the films, and the making of The Hobbit , the set has been restored and extended to become a permanent attraction, where you can not only peer at the Hobbit dweiilings but also wine and dine in The Green Dragon Inn and The Shires Rest at Hobbiton Movie Set Tours , 501 Buckland Road, Hinuera.
Tongariro National Park , home to three volcanoes: Mount

3/23/2014 1:55:58 PM

http://www.movie-locations.com/movies/l/lotr1.html

Curated documents

*

Waikato

Curated document

Film locations for Lord of the Rings: The Fellowship of the Ring (2001)

An amazing achievement, three films in The Lord of the Rings trilogy filmed simultaneously, against the varied backdrops of director Peter Jackson 's native New Zealand , standing in for Tolkien's 'Middle Earth'.

The private Alexander Farm, near Matamata, toward Karapiro in the Waikato area, represented the rolling, bucolic greenery of the Shire, home of Bilbo and Frodo Baggins. In order for 'Hobbiton 'to blend naturally into the landscape, the extensive set was planted with crops and flowers a year before the cameras rolled. Matamata, which now proudly advertises itself as Hobbiton, has a Tourist Information Centre at 45 Broadway. And the really good news is that, following the success of the films, and the making of The Hobbit , the set has been restored and extended to become a permanent attraction, where you can not only peer at the Hobbit dweiilings but also wine and dine in The Green Dragon Inn and The Shires

3/23/2014 1:55:58 PM

http://www.movie-locations.com/movies/l/lotr1.html

○ keep ● auto ○ notify

YouTube-Videos

http://www.youtube.com/watch?v=oHg5SJYRHA0

Keywords

The Lord of the Rings	10
Bilbo Baggins	10
movie location	8

Map

Pictures

Bilbo's cave

The cave from LotR character Bilbo Baggins is a permanent attraction in the Waikato area.

Figure 4.20.: Screenshot of the WiPo Curation Interface.

In order to keep the time-consuming curation workload manageable, curators can choose how curated content is handled when the underlying source(s) change(s). The following options are available:

Keep The curated text remains in the curated database as it was created, even if the original source(s) change(s).

Notify Also called re-present; once the underlying sources have changed they are presented again to curators so that they can decide what to do when they see the changes.

Auto Changes to the underlying sources are propagated to the curated database without being double-checked by a curator.

With auto-updating it is important to identify individual text segments of curated documents. This is achieved by determining so-called tokens (text strings) before and after the segments in the original source. In short, the update process works as follows. When a URL is re-crawled, the document as a whole is checked for changes by comparing the hash of the first crawl (stored in the source repository) with the newly crawled data hash. If a change is detected, the response depends on whether segments have been registered for this document. If no segments are registered, the candidate document can be updated right away. However, if the document has changes and was used in segments, the tokens have to be identified in the re-crawled document. If tokens can be found, the update follows the mechanism specified by curators (*keep, notify, auto*), if not, the document changes into the notify state. If tokens cannot be uniquely identified, the most likely text segment is searched for. More information on tokens and how they are determined and found is provided in [SGH+14b].

Once a text segment has been found, it is processed according to the mechanism specified by the curator. If a notification is required, the changes are written back to the *CandidateDB* with the status notify. If no notification is required, it is checked whether an auto-update is desired or not. In the first case, changes are propagated to the curated database directly. In the latter, changes are discarded. The exact way of how changes are propagated is illustrated by the set of rules, depicted in Figure 4.21.

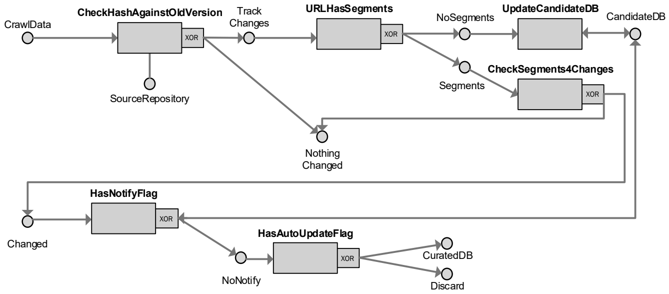


Figure 4.21.: WiPo: Update Cycle.

4.2.4. Search

The search process has been fully implemented as depicted in Figure 4.22. Before a search can be conducted, curated documents have to be indexed (8). To do so, the *AutomatedIndexer* implemented as MongoConnector[18], continuously running in the background, updates the employed Solr[19] *Index* whenever the mongoDB holding the curated database changes. The index contains only *DocumentIDs* but not actual documents.

The search module (6) is the central element for searchers to interact with the WiPo system. It is implemented in a way that it is responsible for the transitions *Check4Validity* and *PrepareSearchInput* from Figure 4.14 and Figure 4.15, respectively, which describe the *InputPreparation* step as well as for the actual search illustrated in Figure 4.22.

This means, the search module receives the user-supplied *Keywords* and separates them from the supplied URLs. While URLs are validated and forwarded to *WebSearchRetrievalDataMining&Curation*, *Queries* are built from the *Keywords* and sent to the Solr *Index* which responds with relevant *DocumentIDs*. Subsequently, *PossibleResults* are retrieved from the curated database.

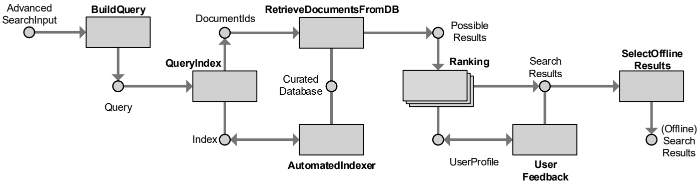


Figure 4.22.: WiPo: Search as Implemented.

After the retrieval of relevant documents, ranking takes place. This is done in a number of strictly linear steps, which allow for easy integration of new ranking steps and removal of those which are obsolete. Starting with all *PossibleResults*, documents that a user has previously marked irrelevant (hidden) are excluded, leading to *PreliminaryResults*. Then, documents and their keywords are ranked according to user preferences, i. e., the interest they stated in their *UserProfile*, producing *RankedResults1*. A second ranking step reorganises the result-set according to former ratings resulting in *RankedResults2*. Finally, the linked documents are loaded to complete the *SearchResult*. The process is depic-

[18] https://github.com/10gen-labs/mongo-connector, accessed: 2015-05-31.

[19] https://lucene.apache.org/Solr/, accessed: 2015-05-31.

ted in Figure 4.23. In the figure, the modularity can clearly be seen by the strict linearity of transitions. Also, it is evident how simple it is to add more modules.

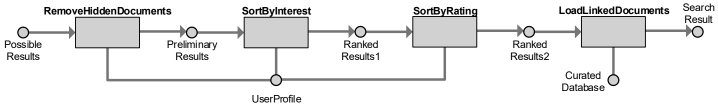


Figure 4.23.: WiPo: Ranking.

As a means to make results available offline, search results can be stored to the note-taking software Evernote[20], as evident in the overall search process shown in Figure 4.22. Currently, this is not technically part of the search module and is handled by the Web Interface (4), which acts as an intermediary between the Web GUI and the Evernote API. Similarly, *UserFeedback* is processed by the Web Interface (4).

4.3. Comparison of Selected WiPo Scenarios

Having discussed the approach and the implementation in some detail, this section is dedicated to exploring sample use cases to which WiPo can be applied. As stated in Section 4.1.1, WiPo can provide the biggest benefits if a number of sources are to be tapped and integrated, regardless of whether users are consumers, businesses, or even authorities or Non-Governmental Organisations (NGOs). Dillon et al. [DRSV14] demonstrated how WiPo can be applied to *tourism* (leisure consumers), *healthcare* (patients, i. e., health service consumers), and *New Zealand Land Search and Rescue (LandSAR)*[21], i. e., an NGO. In this section, these previous examples are extended by studying a business case, namely that of a company making use of WiPo in their environment analysis.

Together with the addition of one case, the comparison modus has been overhauled and has been placed on a more profound foundation. To this end, ideas from strategic management have been borrowed and applied to the use case analysis in order to allow for a more in-depth comparison. The method is developed and described in Section 4.3.1. Subsequently, the four cases will be discussed; commencing with the LandSAR case (Section 4.3.2) which, representatively for all cases, has been evaluated in-depth by means of an expert interview study (Section 4.3.3). Thereafter, the cases tourism, healthcare, and business will be

[20] http://evernote.com/, accessed: 2015-05-31.

[21] https://www.landsar.org.nz/, accessed: 2015-05-31.

discussed in Section 4.3.4, Section 4.3.5, and Section 4.3.6, respectively. This is followed by an overall comparison in Section 4.3.7.

4.3.1. Comparison Method

Strategic management is a holistic approach to medium and long-term business management which looks at future states and investigates what this means for the business today, rather than forecasting from the current state onwards [Bal12a]. The approach can be viewed as a three-step process [Bal12a], consisting of:

1. Information Analysis,
2. Strategic Concept,
3. Strategy Implementation.

Information Analysis gathers strategy relevant information and prepares it for the following strategic management steps. The *Strategic Concept* phase is about turning this information into different strategic steering options and choosing the most appropriate one. Finally, *Strategy Implementation* is concerned with bringing the strategy into operation and controlling if, and how, it takes effect.

It is obvious that the *Strategic Concept* and *Strategy Implementation* are not suitable for a use case comparison; however, in strategic management various tools exist to support *Information Analysis* which can also serve as framework for a use case comparison. This first step comprises the sub-steps *Environment Analysis, Enterprise Analysis* and *Strategic Analysis* [Bal12a]. Since the latter is concerned with combining information gained through the first two means and transforming them into strategy-relevant information, it does not provide significant added value to the context analysis aimed at here. This leaves the first two, *Environment Analysis* and *Enterprise Analysis*, as possible tools for a use case analysis which aims to provide a means to compare cases in order to identify similarities and differences. Readers interested in a more in-depth discussion of strategic management are referred to [Bal12a; Gou12; BH12].

While the *Environment Analysis* is generally seen to return *Opportunities* and *Threats*, the *Enterprise Analysis* is supposed to deliver *Strength* and *Weaknesses*, which can then be combined into a SWOT analysis, where SWOT is an abbreviation of *Strengths, Weaknesses, Opportunities, Threats* [Bal12a; Gou12]. On its own, a SWOT analysis is rather generic and will, therefore, not be applied here.

Nevertheless, there are means to structure internal and external factors, which will be discussed in the remainder of this section.

Gould [Gou12] identifies the following internal factors: *People, Finance, Company or Brand Image, Infrastructure and Scale, Specific Expertise, Total Proposition,* and *Customer Experience.* Whilst the first four factors are not particularly applicable to WiPo, the latter three are unique propositions of WiPo and shall hence be analysed in the comparison.

In contrast Baldegger [Bal12a] approaches internal factors from a high-level view. He looks at *resources and skills* which can be mapped to *people* and *infrastructure* of Gould [Gou12] and are, thus, not of interest. Furthermore, he suggests *strategic success position* which is not really applicable to the case comparison applied here, given that it is about use cases and not about businesses. Baldegger [Bal12a] also proposes *unique capabilities* and *core competences.* To detect these capabilities and competences he suggests using a check-list based approach or the value chain approach developed by Porter [Por85]. However, these approaches really look into business organisation which is non-existent here. Therefore, for now *Customer Experience, Specific Expertise,* and *Total Proposition* will be kept in mind, given that *core competences* and *unique capabilities* can be mapped to *Specific Expertise* and *Total Proposition.*

With regard to external factors, Gould [Gou12] considers the following environments: *Political, Economic, Technological, Social and Cultural,* and *Competitive.* Similarly, Barney and Hesterly [BH12] suggest investigating *Technological Change, Demographic Trends, Cultural Trends, Economic Climate, Legal and Political Conditions,* and *Specific International Events.*

Conversely, Baldegger [Bal12a] suggests analysing *Stakeholders, Customers and Output Markets, Competitors and Industry,* and the *General Environment* in order to conduct a *Scenario Analysis.* Contradictory to the use case comparison intended in this work, their suggested scenario analysis aims at identifying and comparing likely future developments of the company, which is not applicable here. Given that besides the operator of a WiPo instance, the customers and suppliers, no further relevant stakeholders exist, a stakeholder analysis is not really necessary. Nevertheless, it makes sense to look at those stakeholders identified.

While markets will mainly be determined by the use case in question, it is interesting to look at what products can be offered to customers and how they can be characterised in terms of willingness to pay. For analysing competitors and the industry, an analysis according to Porter's 5 Forces [Por97; Por85; Por08] can be used. This technique explores *actual competition* in a given market influenced by *potential new competitors, suppliers, buyers,* and *substitute goods.* It seems unnecessary to conduct a full analysis according to this schema but it should be kept in mind that it makes sense to look at the competition and potential substitutes, as well as customers.

As far as the general environment is concerned, Baldegger [Bal12a] suggests similar dimension as Gould [Gou12]; namely *Ecological, Technological, Economic,* and *Social* sectors.

Although the works cited here do not refer to the external analysis by name, what they are essentially doing is an analysis of the PEST dimensions (Political, Economic, Social, Technological) or a derivation thereof.

In 1967 Aguilar [Agu67] was most likely the first to identify these dimensions as relevant when scanning the business environment, although he did not use the acronym PEST. Furthermore, he mentioned the dimensions science, in connection with technology, and demographics, in conjunction with social aspects.

PEST and its derivations such as PESTEL (Political, Economic, Social, Technological, Environmental, Legal) are acronyms of the external dimensions they consider. PEST or PESTEL are described extensively in [CPT10; Gil11] but also in [BF12] focusing on Social, Technological, Economic, Ethical, Political (STEEP) dimensions. This method is also discussed on numerous management oriented websites, such as Strategic Management insight[22] [Jur13], the Chartered Institute of Personnel and Development (CIPD)[23] [CIP13], and RapidBI[24] [Rap07]. Today, various forms of this approach exist; here, a short list is given based on Strategic Management insight [Jur13], backed by other sources as cited:

PEST	= STEP	= ETPS = Political, Economic, Social/Socio-Cultural, Technological [CPT10; BF12; CIP13; Rap07]
STEEP	= PEST	+ Ethical [BF12]
STEEPLE	= PEST	+ Environmental + Legal + Ethical [CIP13; Rap07; CPT10]
STEEPLED	= STEEPLE	+ Demographic
SLEPT	= PEST	+ Legal
PESTEL	= PESTLE	= PEST + Environmental + Legal [CPT10; CIP13; Rap07]
PESTELI	= PESTEL	+ Industry analysis
PESTLIED	= PESTEL	+ International + Demographic [CPT10]
LONGPEST	= Local	+ National + Global + PEST

[22] http://www.strategicmanagementinsight.com, accessed: 2015-05-31.

[23] http://www.cipd.co.uk, accessed: 2015-05-31.

[24] https://rapidbi.com/, accessed: 2015-05-31.

Together, this results in the following potentially relevant dimensions: *Political, Economic, Social / Socio-Cultural, Technological, Environmental, Legal, Industry, Ethical, Demographic, International, Local, National,* and *Global.* Structuring the analysis in such a way helps to ensure that nothing of importance is omitted and that the analysis is not too narrow [CIP13]. Furthermore, because it is aimed at an analysis framework for use cases, the differentiation between external and internal factors will be omitted. However, input from both streams will be used to develop the final framework.

Not all of the aforementioned dimensions are relevant in the context of WiPo, therefore, a meaningful subset has to be chosen. Looking at the presented use cases as well as the general purpose of WiPo, *International, Local, National,* and *Global* factors are not relevant, as WiPo is targeted at niche domains and not a globally operating organisation. Thus, the scope of these dimensions is too wide.

Here, the *Technical* dimension focuses on the infrastructure and implementation of WiPo rather than on external technology changes as originally intended in [CPT10]. This can also be seen as part of the *Specific Expertise* mentioned earlier. Another part of this expertise is the curators and the organisation behind them, which are also relevant in the context of WiPo. If to be subsumed in one of the dimensions, this fits well into the *Economic* dimension – bearing in mind that the internal-external-divide has been waived. The *Economic* dimension usually considers the overall economy such as Gross Domestic Product (GDP) growth [CPT10], nevertheless, here it will be re-interpreted to also take into account the business perspective in terms of organisational structure and potential business models. However, no in-depth business analysis is to be expected. Furthermore, the *Industry* perspective including *Potential Substitutes* and *competition* is considered part of the *economic* dimension. Hence, this dimension will be referred to as *Economic / Business.*

The *Social* dimension looks at external people such as customers or potential customers [CPT10]; in this framework it will include relevant *Demographic* factors, as proposed in [CPT10]. As a consequence, this dimension is concerned with customers in particular, not society as a whole, for example buying habits and educational level as discussed in [Jur13], will not be part of this dimension. In addition, potential factors influencing curators will be discussed.

A PEST analysis does not differentiate between *Political* and *Legal* dimensions, only considering *Political* factors. This is also reasonable here as it can be argued that political issues are future laws. While it might be important to differentiate between the two in a corporate environment, for the case comparison, here, this is not relevant – bearing in mind that this work does not focus on legal aspects but rather on technical and organisational perspectives. The same is true for *Ethical* considerations, which are about complying with certain standards beyond legal needs, e. g., regarding recruitment standards [Jur13]. These are concerns which are not of particular interest in the case of WiPo. Nevertheless, some general statements regarding this dimension shall be made in the context of the *Political/Legal* dimension, therefore, *Political/Legal/Ethical* will be treated as a single entity.

Likewise, *Environmental* considerations are not yet in the main focus of standalone IT systems but will become of importance once a system is cloud sourced, a sector where the topic of green IT is largely discussed (for instance [BAB12; MLB+11; BAHT11]). However, for most WiPo cases it is not of particular relevance.

The analysis just introduced will be preceded by a general case description which will incorporate the dimensions *Customer Experience* and *Total Proposition*, as mentioned at the beginning of this section. From this a six-step discussion of the use cases has been derived, comprising the following dimensions: *Case Description*, *Technical, Social, Political/Legal/Ethical, Environmental,* and *Economic*. If a term had to be coined, this analysis could be named SPEET (Social, Political/Legal/Ethic, Economic/Business, Environmental, Technological) to clarify the difference to STEEP, which considers *Ethical* factors as a separate entity, but not *Environmental* factors.

This kind of analysis, whilst apparently not widely used, is supported by previous research. For instance, Peng and Nunes [PN07] have used PEST analysis for finding and focusing on a narrow study object in Information Systems research. Haselmann [Has12], has used similar dimensions – economic, legal, organisational, and technical – to investigate the phenomenon of cloud computing [Has12; HV10].

However, before the actual analysis can start, it has to be clarified what will be discussed for each domain. Dillon et al. [DRSV14] have described a number of attributes that specify use cases. These are *Target User, User Input, Additional Input, Source Selection / Service composition, Type of Information, Data Mining, Curation, Services Response Time, Service Refreshing Requirements, Visualisation and Output,* and *Business Model*, and all can be mapped to the SPEET-dimensions used here. Furthermore, additional attributes have been identified based on the above elaborations and added to the list. While this list may not be entirely comprehensive, in concordance with Dillon et al. [DRSV14], it can be stated that these dimensions and attributes are applicable to all cases described and can potentially be applied to any other WiPo scenario. In the following list, attributes adapted from [DRSV14] are marked with an asterisk (*).

Use Case Description While not being a real dimension for the analysis, the use case description does pave the way for it. In a sense, it elaborates on the use case to give the reader a thorough understanding of what the case is and how it provides value to the user.

Proposition & Experience This attribute combines the concepts of *Total Proposition* and *User Experience* and gives an overview of the provided service in each use case. Furthermore, it goes into detail about what a user can expect from the systems. In this respect, it will also elaborate on user benefits and how they are likely perceived.

Service Composition* This has originally been part of *Source Selection / Service composition* [DRSV14], which has been split into two for this work. *Service Composition* focuses on which sources are to tap in order to provide meaningful content.

Social Dimension The *Social* dimension focuses on people included in the WiPo use cases. On the one hand these are users of the system. To this end, users have been distinguished from *Customers* (termed *Target Customer* in [DRSV14]). Conversely, these are also curators, who have not been discussed so far.

User This term will be used to describe who the user will be and is not to be confused with customer - which is not necessarily the same group of people, although it can be. In the context of WiPo it is, for example, important to understand the user's needs and capabilities.

Curator This attribute is concerned with the qualities a curator should have – in the given use cases – and how curation is organised, i. e., individuals, small and large groups of experts or crowdsourcing are all considered.

Political/Legal/Ethic This dimension focuses on *Legal* and *Ethical* aspects of WiPo and potential changes due to politics and jurisdiction. However, it has to be kept in mind that a full legal analysis is beyond the scope of this work. Thus, this thesis will only touch upon some issues regarding the following two aspects:

Intellectual Property The question of who possesses the content gathered for WiPo, as well as who has the right to exploit this content, are topics of concern. This question is also discussed in regard to data and cloud computing [Hoe14] which highlights the importance of this attribute.

Privacy This is an important issue with regard to online search and cloud computing (see for instance [Hoe15]), as discussed in Section 3.5.1. Hence, it shall not be omitted here.

Economic/Business Dimension As elaborated on when defining the analysis framework, this dimension is primarily concerned with the market WiPo is operating in, as well as the business side of the use case in question. Consequently, it is not overly concerned with the economy in general.

Potential Customers* In [DRSV14] this was referred to as *Target User*. Here, it will be used to describe whether the service addresses individuals, groups or organisations. In contrast to the attribute *User* of the *Social* dimension, *Potential Customers* regards the customer who pays for the service, rather that its user, although it is recognised that the two parties could be one and the same.

Potential Providers This is the business side of *Service Composition* and describes the interaction with businesses providing input for WiPo.

Potential Competition This is concerned with similar services available on the Web which might serve as substitutes for WiPo.

Organisational Structure This is related to both curation and the *Business Model*. It explains how an organisation behind WiPo could be organised. This touches upon the technical infrastructure – single instances versus large server farms – as well as on the overall cost structure.

Business Model* Regarding this attribute, appropriate business models will be discussed, based on the above attributes.

Environmental This is concerned with the effect WiPo might have on the general environment in the specific use cases. However, it makes sense to distinguish between two ways in which the environment may be affected.

Direct Consequences for the Environment (ENV) These are consequences IT may have on nature, i. e., the issues that arise simply because WiPo is operational.

Indirect Consequences for the ENV This attribute deals with implications that are caused by people using WiPo rather than the consequences of WiPo running. For instance, an indirect consequence of WiPo, using the tourist analogy, is that natural areas may be destroyed by high tourist numbers resulting from increased awareness of the site.

Technological Dimension This dimension will extend on what has been discussed in the previous sections with regard to the WiPo implementation. To this end, it will be highlighted what technical specialities the given use case has. In particular the following attributes will be looked at:

Data Mining* While being a rather wide area it will be pointed out which techniques might be useful in the cases under discussion and what potential extensions WiPo could benefit from.

Service Response Time* Depending on the use case, users will have different expectations with regard to an acceptable timeframe for inputs to be fully processed. *Service Response Time* is concerned with the question of how this can be realised technically.

Data Freshness* This attribute investigates how current the data has to be and is dependent on the use case requirements and on the type of data that is mainly used (static versus up-to-date data). Whilst originally the term *Data Freshness* was not used in [DRSV14], this dimension can be seen as a link between *Service Refreshing Requirements* and *Type of Information* of [DRSV14].

Visualisation & Output* This attribute is concerned with the user interface and which special requirements with regard to visualisation the different use cases may have.

User Input* In different use cases, various types of inputs are common. This is addressed by this attribute, which also discusses how user inputs need to be processed by the system. Originally, a further attribute, *Additional Input*, was introduced by Dillon et al. [DRSV14]. However, this differentiation unnecessarily divides the discussion, since *Additional Input* elaborated on non-standard user input that is mined from the obvious user inputs. Therefore, in this work both have been incorporated into one *User Input* attribute.

Source Selection* The second part of the original *Source Selection / Service composition* [DRSV14] will discuss how sources to be used are determined from the available source.

Curation* This attribute addresses the different possibilities for curation, which can be achieved manually – by an individual or crowd – or algorithmically. Whilst manual curation is not technical per se, the process can be seen as a technique. Furthermore, it has to be supported by algorithms which are de facto technical. In this section appropriate curation methods will be elaborated on for each case.

The above dimensions and attributes are summarised in Table 4.3, which also states whether the definitions were first described in [DRSV14] or if they were developed for this work.

Table 4.3.: List of Investigated Attributes.

Dimension	Attributes as Discussed Here	Attributes as in [DRSV14]
Use Case Description	Proposition & Experience	———
	Service Composition	Source Selection Service Composition
Social	User	———
	Curator	———
Political/Legal/ Ethical	Intellectual Property	———
	Privacy	———
Economic/ Business	Potential Customers	Target User
	Potential Providers	———
	Potential Competition	———
	Organisational Structure	———
	Business Model	Business Model
Environmental	Direct Consequences for ENV	———
	Indirect Consequences for ENV	———
Technical	Data Mining	Data Mining
	Service Response Time	Service Response Time
	Data freshness	Service Refreshing Requirements Type of Information
	Visualization & Output	Visualization & Output
	User Input	User Input Additional Input
	Source Selection	Source Selection / Service Composition
	Curation	Curation

4.3.2. Scenario 1: New Zealand Search and Rescue

Use Case Description In recent years, the area of emergency and disaster management has undergone a substantial change and is professionalising [WO01]. Furthermore, the use of new technologies, such as social media services, is increasing [BESL13; EB13]; something that is evident from the proceedings of numerous regional and global Information Systems for Crisis Response and Management (ISCRAM)[25] conferences. In the context of WiPo, the use of mobile phones (e. g., [RDGG14]) and the use of online services, such as social media (e. g., [FHS+14; BESL13; EB13]), are particularly relevant.

In 2010, Bunker [Bun10] identified the need for collaborative information management in crisis situations in order to improve community warning and emergency incident response. Later, she and others investigated how social media such as Facebook and Twitter were used in the Christchurch earthquake [BESL13] and the Boston marathon bombings [EB13].

As an object of study, New Zealand Land Search and Rescue (LandSAR)[26] has been chosen, as this WiPo case is currently being explored in depth and a paper focusing solely on this scenario, including an extensive evaluation briefly discussed in the next section, is in preparation [DRSV]. In order to evaluate this use case, a series of interviews were carried out with a regional LandSAR group. Besides the evaluation results presented in the next section, these interviews gave an insight into the current protocol of LandSAR. For instance, it can be stated that technology adoption in the area of knowledge management is in its infancy and that plenty of expert knowledge resides within the heads of key members of the LandSAR group.

There are various situations in which LandSAR is called for help. According to their Website [Laned], it is LandSAR's aim to provide "land search and rescue services to lost, missing and injured all over New Zealand [...] in suburban, urban, wilderness and rural areas [...]". WiPo can provide the biggest benefit in rural areas and in the wilderness as in this terrain mobile Internet connections are generally weak and in New Zealand often non-existent. Parties who go missing include children, people with dementia, and tourists. Tourists are more likely than other vulnerable parties to go missing in rural terrain. Therefore, this case will consider the situation of a tourist reported missing after entering a national park.

In such a scenario, a typical LandSAR group would have made pre-plans or carry out ad hoc planning when an incident arises. Either way, a search manager

[25] http://www.iscramlive.org/, accessed: 2015-05-31.

[26] https://www.landsar.org.nz/, accessed: 2015-05-31.

would compile a plan of action from a number of sources, which may include theory texts about what to do in a particular case, historical search databases containing information gathered in similar previous cases, map data, including topological maps, and weather forecasts. This general information would be enhanced by case-based information regarding the lost party, for example: When were they last seen? What were they wearing? Do they have any special medical conditions?

One of the key challenges is getting all this information to the teams out in the field. Having it ready at operation headquarters is less complicated because – even though they are mobile and can be set up almost anywhere – a power supply will be provided by a generator and a satellite Internet connection will be established. In these headquarters, briefings are commonly conducted to convey all necessary information to the teams just before they set out. Traditionally, search team members would have to take notes and carry maps, whereas WiPo could help by keeping all the information on a mobile device. Furthermore, updated information does not have to be transmitted by radio which can be prone to errors but can instead be transmitted using the WiPo infrastructure pushing updated information from the central repository to the client. On the one hand this is a feature that needs to be built but is very similar to the notification mechanism discussed in the other cases and is as such not difficult to implement. On the other hand, the mobile Internet connectivity cannot be considered to be stable. As a consequence, the update service would have to run in the background and make use of even the weakest Internet connections it can get. An alternative could be using text messaging in the background which sometimes works even if no Internet connection can be established. However, this would require integration into WiPo.

In the background, WiPo can a priori serve as an information repository that is used to maintain static information such as theory texts and historic search information. Its data mining and Web search capabilities can be used to analyse past searches for patterns and to keep volatile information up-to-date. Examples include: the openings of certain tracks or trails, weather data, and tidal information. Additionally, data mining can be used to gain further information regarding certain locations. Most importantly, it can be used to extend potentially sparse information given about the lost party. For instance, additional information regarding illnesses can be added on the spot. Another distinct feature of this case is that non-curators (i. e., search team members) have to feed back new information through the app, which is then to be curated by the central emergency management.

While this WiPo case will be described concentrating on how it can be applied in a scenario where emergency management can prepare in advance, WiPo can also be useful if hardly any information is available a priori. For instance, Bunker suggested in a personal conversation to use WiPo in the context of a pandemic such as the recent outbreak of Ebola in Africa documented in [Wored]. This is a scenario in which information only becomes available as the crisis develops. Nevertheless, it is even more important to gather all information as quickly as possible.

Social Dimension *Users* in this WiPo case are mainly LandSAR team members going into the field and needing the most up-to-date information. This group is – as determined from interviews – rather heterogeneous with regard to their technical abilities. However, given the context it is feasible to offer special WiPo trainings to the teams. Nevertheless, WiPo has to be simple to use but also informative. Here, it is suggested that a simple text-based display, would be best, allowing for easy navigation through all case-relevant information and enabling the user to switch between different types of information such as images, maps, texts, and photographs. Additionally, the aforementioned function to send new information back to operation management is important and has to be easy to use.

Curators in this case will be specially trained LandSAR members who maintain the knowledge repositories and are most likely members of the operations management team. As they receive special training on the software, WiPo does not have to be utterly simplified, yet, it has to allow for a simple workflow and has to be tailored to the specific needs of LandSAR curators.

Political/Legal/Ethical Dimensions As with the other professional case, there should not be any major difficulties regarding *Intellectual Property* because content will be licensed. However, *Privacy* is relevant as far as personal information of the lost party is concerned. This has to be anonymised as far as possible after a case is closed. Nevertheless, this should be common practice today, in due consideration of any duty to preserve records.

Economic/Business Dimension The *Business Model* is rather difficult to approach because emergency organisations are addressed as customers. While they are unlikely to have large funds, they operate as incorporate and thus can probably afford to buy software as they buy other useful equipment. The question remains how to provide WiPo to them. Given the fact that the curation is highly domain-specific and the funding of emergency organisation limited, it makes sense to only provide the software because in contrast to funding, the resources – as far as volunteers are concerned – are not quite as scarce.

Regarding the *Organisational Structure*, it can be stated that WiPo would be mainly a software supplier rather than a full service provider. Therefore, the organisational structure would be the same as in the business case, only no curation department is necessary. As for the organisation within LandSAR either a countrywide WiPo would be run or some groups would maintain their own WiPo. Either way, this organisational unit would have to employ special curators to organise the relevant content.

From a business perspective the *Potential Customers* would be governments or emergency organisations such as LandSAR. However, these are customers who are commonly low on funds. Then again, the costs associated with running the service are not high so this should not be a major issue. However, the costs of developing the software and adapting it to the use case have to be covered. This problem could be overcome by using one organisation as a pilot and offering them a discount. Afterwards the system may be sold to other organisations at a higher price.

Since curation is out-sourced, the *Potential Providers* are not of particular importance but shall be mentioned here for completeness. These are mainly weather and map service providers as well as public websites containing information on tracks, for example whether they are open or closed and what their condition is.

The *Potential Competition* for such a WiPo is strong. For instance, there are the Canadian SAR Technology Inc.[27] and the US-based Mission Manger Inc.[28], which both provide incident management software that is cloud-enabled and provides advanced additional features such as team management. However, one weakness with both is the fact that they apparently do not allow for automated updates of Web-based resources.

Environmental Dimension The *Direct Consequences* that WiPo yields by running alone are not of particular relevance and can be covered in the general discussion on large-scale infrastructures and their energy consumption. The *Indirect Consequences* can also be regarded as of minor importance because LandSAR members should be trained on how to act in the wild. In fact using WiPo

[27] http://sartechnology.ca/, accessed: 2015-05-31.

[28] https://www.missionmanager.com/, accessed: 2015-05-31.

and giving information on how to behave in a given type of terrain may even have a positive impact in this scenario.

Technical Dimension The main task of *Data Mining*, as suggested above, is searching for specific problem spots in specific regions or enhancing existing knowledge with additional facts. Thus, *Data Mining* in this case is mainly about meaningful content extraction as well as entity recognition and possibly association rule mining. Integration of information is important, but the main focus will be on automatically discovering changes in the environment; the integration is of such a high relevance that it ought to be done by humans. Nevertheless, automated updates are an important feature. Hence, *curation* will mainly be a manual task done by a small group or an individual aiming at applying a set of context-specific rules to produce the required high-quality data.

Regarding *Service Response Time*, it has to be pointed out that this case is inherently time critical. Therefore, the database has to be updated regularly. For most information, a daily or even weekly basis may suffice. That said, as soon as case-specific information becomes available this gathering and mining process should be the quickest of all the cases discussed in this thesis and should be able to report reliable results within minutes, at most. As with other use cases, service response time should be instantaneous as far as collected data are concerned. However, the particularly quick mining process requires high-end hardware to be used. Along with this goes the observation that *Data Freshness* should be as high as possible for both static and dynamic information as a person's life might depend on the quality of information.

The *Visualization & Output* has to be well adjusted to mobile devices and has to feature an easy to use GUI in order to be helpful in the field. It will mostly depict texts including tables, images, and map data. Curation in this case involves editing large amounts of texts, which is not necessarily the case for other WiPo cases discussed here. Hence, the curation interface will have to support this form of text editing much better. The *User Input* in contrast will be much simpler as it is case-based. Thus, it will mainly be keywords. All other types of input are likely to be dealt with by curators. Nevertheless, given the fact that users can provide feedback this has to be accommodated.

The initial *Source Selection* will be predominantly manually inserted URLs. However, through focused crawling, fact knowledge is to be extended – for example, these sources can be determined by the system based on pattern matching.

4.3.3. Evaluation of WiPo for New Zealand Search and Rescue

In order to gain an initial insight in the usefulness of WiPo, a series of interviews was conducted with domain experts including a live demo of the use case. Four experts from a local LandSAR group participated in the study. All of the experts were part of the so-called incident management team which manages and coordinates search operations. The combined search and rescue experience of all interviewees exceeded 50 years; one of the interviewees was a founding member of the local group and held a senior role within the national LandSAR body focusing on IT.

While four is a rather small sample, it can be seen as sufficiently large for the purpose of an initial study given the interviewees' over-all experience and knowledge of LandSAR. Similar to previous work by the author [SPM11], in which an expert interview study was conducted to extract incentives to contribute to a security knowledge sharing platform, it was opted to limit this initial interview study to four experts to use the available time most economically and allow for an in-depth analysis of a few interviews as compared to barely analysing a greater number of interviews. Furthermore, the small sample covered the most important staff at the local support group and is therefore representative of a group which could serve as a pilot in an implementation.

This study was loosely based on the seven-step approach to interviewing by Kvale and Brinkmann [KB08], similar to [MSLV12]. They suggest a study should consist of the following seven steps: thematising the study, designing the study, interviewing, transcribing, analysing, verifying, and reporting/publishing. In this instance, transcription was omitted and the analyses were based on notes taken during the interviews. This is justified as the study is an initial gauge of whether WiPo can potentially work and in what direction it should be further developed. Thus, a full analysis of transcripts does not seem sensible at this early point. That said, it should be pointed out that investigating this particular use case for WiPo is a work in progress. In particular the steps verifying and publishing are currently being worked on, as a publication is in preparation [DRSV].

The interviews, which took between 45 and 90 minutes, were structured in three main parts. First, the interviewees were asked to describe the current state of technological affairs within their local group, as well as in LandSAR as a whole, and to give possible future directions. This was followed in part two by an introduction and a live demonstration of the WiPo system. Part three discussed how WiPo could be helpful to the area of crisis response management, what improvements WiPo would need, and what might hinder the usage of WiPo in the emergency management context. As a consequence of this structure, the interviews were conducted by two researchers, one focusing on the current technology usage and WiPo in use, the other – the author of this thesis – focussing on demonstrating the prototype and possible consequences for the implementation. Consequently, this work will focus on the evaluation part as the current state has been mentioned in the case description as far as appropriate. All interviews were digitally recorded and analysed with regard to their specific domains by the according researchers.

In general it can be said that all interviewees responded positively to the proposed solution and its applicability to Search and Rescue. In particular the option to have an auto-updating, centrally maintained, electronic information repository was deemed extremely useful. As indicated above, large amounts of information are maintained mentally by rescuers, and also physically in non-integrated online and offline locations such as flat files, photocopies of texts, and as books. Having all this information readily available and integrated improves the current situation and allows for automated updates to ensure the information is as accurate as possible.

In this context, it was mentioned by one interviewee that WiPo would be really powerful in regard to pre-planning for search and rescue operations. The curation process that is necessary was considered to be key to the whole process but was also identified as a bottleneck because it is important to get expert curators to conduct this crucial process.

Given that most of the LandSAR related information is rather volatile (tidal times for instance) the automated update function was evaluated as really useful. In particular the fact that WiPo provides three update modes (auto, notify, and keep) was seen as valuable and superior to the manual processes currently in place because up to now there is no regular check for the currentness of stored information, which may have severe consequences if discovered too late.

Finally, the ability to carry all the available information that has been curated on a mobile device into the field was described as a big advantage and even as being a critical attribute of WiPo. As mentioned before, the search base is usually connected but the teams lose connectivity as soon as they head into the field. In this context the possibility to push information to teams out in the field was perceived as very helpful, although the issue of technical feasibility was raised.

From the interviews, it can be concluded that WiPo is capable of providing a significant benefit to LandSAR. However, some modification to the current prototypic system, in particular to the presentation layer would be necessary.

Based on the findings above, WiPo can be beneficial in two main ways. First, it can be useful with regard to pre-planning. This means that all information available for a certain region or type of search is gathered before an actual search. This includes static information such as maps but also dynamic information such as weather forecasts and tidal information. The latter can be achieved owing to the automated update capabilities of WiPo which ensures that the information within the WiPo system is always up-to-date. Also, it could serve as a knowledge repository for information that has been gathered in debriefing sessions so that this information – which at the moment mainly resides in human memory and a non-indexed file structure – can be reused automatically. To achieve this, searchers need to be given the chance to provide feedback into the system.

The second advantage lies in getting information to the actual field teams. The WiPo search functionality can help searchers find information they specifically need. In particular in an urban search with Internet available this is not a problem but in areas where no connections are available rescuers can pre-load information on their way to the operating site while there still is a connection. Whilst it might not be possible to push information everywhere into the field, an auto-update when back at base would be a first step in this direction. Implementing a dedicated client, it would also be possible to realise it in a way that makes use of any connection it detects so that users do not actively need to synchronise their device. Along this line is also the possibility for search managers to push information to the teams in the field (for instance when updated information regarding tracks becomes available). In fact, with only minor modifications this would also be possible using the actual Evernote implementation. Nevertheless, the push functionality was highlighted as having a very good utility for LandSAR by a number of interviewees [DRSV].

At this stage it is not possible to integrate a user's own documents with a search. Adding this feature in a dedicated client would render note taking in briefings redundant, as searchers could take notes electronically and integrate them with information provided by search management and information could easily be shared with other team members.

In conclusion, WiPo for LandSAR was perceived as a useful tool by the interviewees [DRSV]. Nevertheless, a number of potential improvements to WiPo have been pointed out. In this respect the initial interview study fulfilled its purpose of serving as a primary evaluation in the design science approach according to Peffers et al. [PTRC07], which was utilised here.

4.3.4. Scenario 2: Tourism

Case Description The tourism case, first described in [DRSV14], built the basis for the implementation of WiPo described in Section 4.2. However, the actual implemented case for a conference demo ([SGH+15]) is more specific than the initial description of a generic tourism case in [DRSV14], as it has been applied to the area of filming location tourism. This fits with the idea that WiPo is of particular benefit in niche domains.

In any tourism case, holiday preparations are usually information-intensive, at least if tourists want to get the most out of their holiday by doing as much as possible in a short period of time instead of just relaxing under a palm tree. This means a number of sources have to be consulted and compared to find the most suitable activities in a chosen destination. Even when considering doing nothing but relaxing a destination has to be found, which is not trivial, and can often be a multi-step process. First, a country has to be chosen, then a region, and finally, accommodation as well as a potential beach to relax on. This can be seen as a decision which people feel that they never have enough information, a problem that has been analysed with regard to WiPo in [DSVR13].

In the case presented here, tourists who wish to combine their devotion to fantasy films with their love of unspoiled countryside and hiking are considered. As a country, New Zealand has been choose for two reasons; firstly, it is rich in nature as a consequence of its sparse population, and secondly, this natural beauty has resulted in the country being used as a location for numerous filming, including *Lord of the Rings* and *The Hobbit*.

It is presumed that these tourists will want information on film locations, on other points of interest such as shelters for hiking, and on flora and fauna, in order to have it readily available on their mobile device during their tour. However, mobile Internet coverage in the more rural areas of New Zealand, such as in the Southern Alps, is almost non-existent, making it an ideal case for WiPo.

Admittedly, there is a plethora of information regarding tourism in New Zealand available, in fact, even provided specifically for mobile devices by apps such as ITRAVELNZ[29] or TUHURA[30], but information provided through these apps is usually not well-integrated and cannot easily be persisted off-line.

In contrast, WiPo can be used in a unique way in this example. Firstly, users can provide a personal profile containing their interests (e. g., Lord of the Rings, local flora and fauna, etc.) and then query the curated database for Southern Alps and hiking. Furthermore, they may provide a URL to a blog of a friend who recently did a similar trip and personal files such as a pre-sketched itinerary with "must-see" places. All this allows for advanced personalisation.

[29] http://www.itravelnz.co.nz, accessed: 2015-05-31.

[30] http://www.tuhura.co.nz, accessed: 2015-05-31.

Supposing a user provides a city in New Zealand that they want to visit, together with the just mentioned information, first the database will be queried and immediate results will be returned. Meanwhile, a focused Web crawl will search for film locations close to the specified city. Also, information on accommodation, transportation, weather forecasts, emergency precautions, interesting scenery and local plants and animals, extracted from multiple websites and online databases will be returned. Regarding *Service Composition*, such sources could be: THE WORLDWIDE GUIDE TO MOVIE LOCATIONS[31], FIRSTLIGHTTRAVEL[32], NOMADICMATT[33], and YOUNG ADVENTURESS[34].

During data mining, these results can be classified according to the original keywords or clustered according to WiPo's findings. The curation step ensures the results are of good quality and interlinked so that they are meaningful. Once new results are available, users will be notified and the final result, consisting of a list of documents comprising all of the information relevant to the user, such as maps, a list of possible accommodation, information on plants to be found in the Southern Alps, and relevant roadside information, is presented. Users can then choose to make the whole database, or specific parts of it, available offline.

In summary, the *Proposition and Experience* can be described as follows: WiPo in the film tourism case presents information on film locations integrated with information that is helpful when planning a trip, all personalised to a user's needs. This is facilitated through an easy to use GUI which can be intuitively used. Thus, it serves as single point of truth that can even be persisted off-line and as such is of significant value to users.

Social Dimension *Users* of a tourism WiPo system can be considered as technology-savvy as they carry their mobile devices even when going on a nature-related holiday. Thus, it is probably safe to suppose that most of these customers tend to be rather young and used to dealing with technology that is simple to use and simply works by itself, like many apps on mobile devices do. Many users will be so-called digital natives, people who were born into an interconnected world [Pre01]; however, the exception may well prove the rule. Thus, the app will have to be sophisticated in order to provide significant value to these demanding customers. This is true for both the GUI design as well as for the provided information, which needs to be well integrated compared to alternative sources of information these users would consider.

[31] http://www.movie-locations.com/, accessed: 2015-05-31.

[32] http://www.firstlighttravel.com/lotr, accessed: 2015-05-31.

[33] http://www.nomadicmatt.com/travel-guides/new-zealand-travel-tips/, accessed: 2015-05-31.

[34] http://youngadventuress.com/2014/08/new-zealand-road-trip.html, accessed: 2015-05-31.

Regarding *curators*, it can be said that their expertise will have to be profound with regard to the country, area, or topic they curate. Thus, it is quite likely that they are locals who know about the filming locations, or film enthusiasts having explored the area themselves. Given the sheer amount of knowledge regarding a particular area, it is likely that this cannot be done by an individual or small group but rather by a large group or crowd of locals or former tourists. In this respect WiPo will have to provide good usability with regard to communication for curators, discussion, and version histories in order to be ready to be maintained by a large number of people. Furthermore, extensive support by algorithms, such as automated updates, has to be installed in order to overcome the workload. Nevertheless, due to the subjectivity of the task at hand, manual labour cannot be replaced by algorithms entirely.

Political/Legal/Ethical Dimensions This use case is probably the most complex of the cases presented here in terms of legal context as it heavily reuses content from third parties and thus conflicts with the owners of the original content are likely as their *Intellectual Property* rights are commonly protected. For instance, in New Zealand by the Copyright Act 1994 accompanied by the Copyright Regulations 1995 and by the Copyright (New Technologies) Amendment Act 2008 [Off14]. While in Germany intellectual property is protected by §87f to §87h UhrG [Urhed], the situation in other countries might be different but it is not surprising that content creators and owners want to be reimbursed by any potential profit generated by WiPo.

In contrast to search engines, it is WiPo's aim to build a database from existing texts and deliver them. In light of intellectual property rights, this is very unlikely to be legal for open public services. A starting point could be focusing on sources publishing their content under open licenses such as Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)[35], or GNU Free Documentation License (GNU FDL)[36] which Wikipedia is using [Wiked], or similar.

WiPo can be seen as an intermediary platform, similar to hotel booking sites. However, in contrast to these services, which might favour particular hotels based on the commission they receive, WiPo is intended as an inherently neutral tool. Therefore, providers of goods and services for tourists will probably be happy to make their information available through a centralised platform as it increases their visibility for free. If owners of other data, such as weather data,

[35] http://creativecommons.org/licenses/by-sa/3.0/legalcode, accessed: 2015-05-31.

[36] http://www.gnu.org/licenses/fdl-1.3.en.html, accessed: 2015-05-31.

are concerned, licensing is necessary, which will increase the organisational overheads.

Related to legal and ethical considerations is the idea of *Privacy*. In WiPo, it is relevant because extensive profiles are to be kept, in order to satisfy information needs to the best of WiPo's abilities. However, this is a matter of ethical debate, raising the question of whether user profiles should be exploited to the degree technically possible and to what extent users should be told about privacy. For WiPo, the decision has been made to give users full control over their profile by making it explicit and editable by them. In future versions even disabling profiling is an option; nevertheless, this will impact on search results.

Economic Business Dimension The *Potential Providers* have been extensively discussed in the previous dimension with regard to licensing their content. For this reason this attribute will not be discussed any further here.

Given the user demographics, the question from a business perspective is who the *Potential Customers* will be. With reference to previous section, there are two major options which are closely related to the *Business Model*. Option one is to charge users for the service, whereas option two is to offer the service for free and tap another source of revenue. In the latter case the actual customers may be content providers such as, tour providers, who would pay a commission fee. This would be similar to the business model of many hotel booking websites. Alternatively, the whole project could be advertisement-based in which case the actual customers would be advertisers. In particular for a crowd or community model where users and curators become the same group, an advertisement model would be well-suited. The advertisement-based model, and also the revenue-sharing model, have the advantage that the user profiles can be targeted by tailored advertisements and offers specifically aimed at the users' interests. However, this would conflict with previously discussed ethical considerations.

If a professional business is to be established with employed curators, a subscription model would be appropriate to cover the cost of running the system. To allow an initial penetration of the market, a freemium strategy is appropriate [SF08]; this would be done by providing basic services for free and charging for premium services such as offline functionality, for instance. Given the situation in regard to tourism information, it is more likely that a third party is covering most of the cost. In this context, freemium is still possible, a premium feature could, for instance, allow users to work with the system without advertisement.

That said, tourism and information about tourism is a very competitive market and plenty of information is freely available, so it is questionable how well the market can be penetrated. This fierce *Potential Competition* is evident by the number of examples to which a tourism WiPo can be compared. There are a plethora of websites that offer travel information, booking opportunities and recommendations about activities and sights. This is flanked by an uncountable amount of travel literature available as brochures, books, and e-books. Furthermore, standard search engines such as Google provide a good entry point. This underlines the point made when discussing users that a tourism WiPo has to provide outstanding features in order to be competitive. As a consequence, the tool has to be developed to such a degree that it provides significantly more value than the Web on its own, which is challenging as it requires significant more development.

Therefore, it can be argued that a tourism WiPo would be best organised in a crowd model, in which users and curators become one group. Regarding *Organisational Structure*, this means that there will be a non-commercial organisation providing the platform. The cost of running the same can then be covered through advertisement or through donations similar to Wikipedia.

Environmental Dimension The tourism case is on a much greater scale than the other cases discussed here. Therefore, it potentially has the largest impact on the environment by running alone. However, these *Direct Consequences* are not of particular relevance to WiPo and can, as mentioned before, be covered in the general discussion on large-scale infrastructures and their energy consumption.

In contrast, the *Indirect Consequences* of a tourism WiPo for the environment can be described as potentially dangerous, as it is the aim of this tool to enable tourists to visit places in nature, which they could not otherwise locate. As such, WiPo could increase the number of tourists visiting certain locations, which might be in nature reserves or other protected areas. These tourists might not be mindful of the environment and could destroy sites either by accident, or on purpose. This issue could be addressed by WiPo by providing appropriate information including rules, regulations and potential fines, but ultimately it is out of the control of those running WiPo.

Technical Dimension In this case, *User Input* is first and foremost keywords, potentially enhanced by a query-based set of URLs. However, users may also provide profile data such as personal preferences (e. g., mountainous regions). This input can then be enhanced by keyword extraction and other classifying techniques in order to select appropriate sources and calculate the most relevant result set.

The main task of WiPo, from a technical point of view, is the integration of various sources and databases. Thus, *Source Selection* is most demanding, given the fact that most sources are unstructured text. Sources should initially be seeded by curators. Subsequently, given the sheer amount of source, the source selection has to be algorithmically supported. To that end, the crawling has to be focused and a quality assessment has to be conducted on newly discovered sources. The discovery and selection will have to be guided by user profiles and search queries.

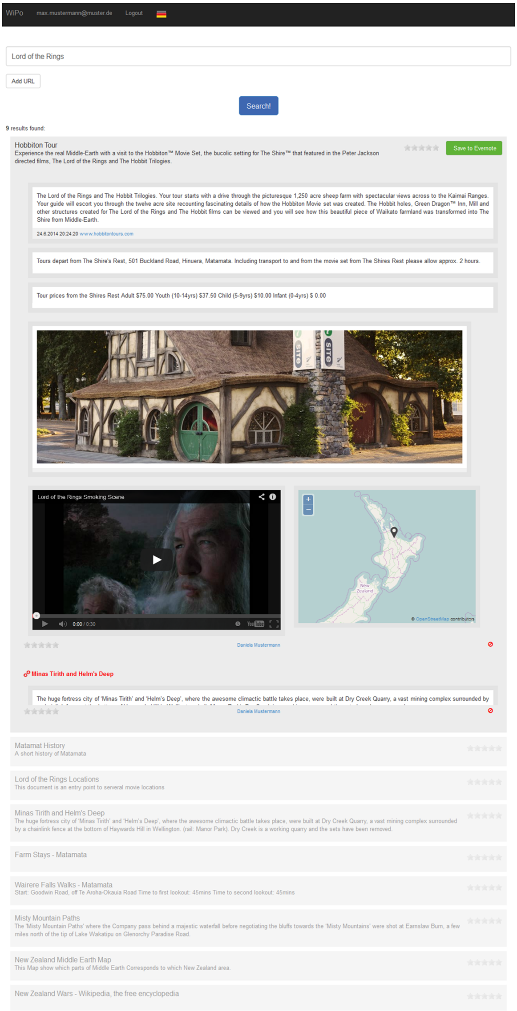
Data Mining in particular for this case, requires integrating different sources in one document, thus, link analysis is very important. Other important aspects are content analysis and clustering to be able to find related topics. In this context, machine learning of how curators have decided on potential documents is promising.

In this case, the main challenge for *Curation* is that most likely a crowd curation approach will be followed. Therefore, a suitable interface for crowd curation has to be found, allowing for communication between curators, discussions, and version histories. Furthermore, the automated updates mechanisms will have to be significantly improved.

Service Response Time needs to be quick given that users today expect answers from Web search engines within fractions of a second. This can be achieved using standard technologies and sufficient hardware resources. However, this only concerns the response to actual queries. Regarding the data mining running in the background and extending the database, users probably do not have particular needs. Thus, this can take longer if it is well communicated that additional information is tailored precisely to their needs and that they will be notified once results are available. Providing an e-mail notification service should suffice regarding long term background data mining. Nevertheless, first results of these steps should be available within several hours at the most.

The *Data Freshness* has to be high in order to satisfy the challenging demands of tourism WiPo users. Most of the data will be static by nature and will not change often. Hence, a medium up-date frequency will suffice. Nevertheless, there will be information such as opening hours, weather data, or tour vacancies that need to be updated more frequently, while locations where films have been shot will most likely not change, although their accessibility might.

Visualization & Output The Output has to be appealing to the users and should incorporate information such as maps and tables, but this is very common in Web applications and is therefore no major challenge. That said, in order to be competitive, this presentation will need to be meticulously planned by usability experts and information specialists to present the condensed and integrated information in the best possible way. As an example, the current implementation can be seen in Figure 4.24.



9 results found:

Hobbiton Tour
Experience the real Middle-Earth with a visit to the Hobbiton™ Movie Set, the bucolic setting for The Shire™ that featured in the Peter Jackson directed films, The Lord of the Rings and The Hobbit Trilogies.

The Lord of the Rings and The Hobbit Trilogies. Your tour starts with a drive through the picturesque 1,250 acre sheep farm with spectacular views across to the Kaimai Ranges. Your guide will escort you through the twelve acre site recounting fascinating details of how the Hobbiton Movie set was created. The Hobbit holes, Green Dragon™ Inn, Mill and other structures created for The Lord of the Rings and The Hobbit films can be viewed and you will see how this beautiful piece of Waikato farmland was transformed into The Shire from Middle-Earth.

24.6.2014 20:24:20 www.hobbitontours.com

Tours depart from The Shire's Rest, 501 Buckland Road, Hinuera, Matamata. Including transport to and from the movie set from The Shires Rest please allow approx. 2 hours.

Tour prices from the Shires Rest Adult $75.00 Youth (10-14yrs) $37.50 Child (5-9yrs) $10.00 Infant (0-4yrs) $ 0.00

Daniela Mustermann

Minas Tirith and Helm's Deep

The huge fortress city of 'Minas Tirith' and 'Helm's Deep', where the awesome climactic battle takes place, were built at Dry Creek Quarry, a vast mining complex surrounded by

Daniela Mustermann

Matamat History
A short history of Matamata

Lord of the Rings Locations
This document is an entry point to serveral movie locations

Minas Tirith and Helm's Deep
The huge fortress city of 'Minas Tirith' and 'Helm's Deep', where the awesome climactic battle takes place, were built at Dry Creek Quarry, a vast mining complex surrounded by a chainlink fence at the bottom of Haywards Hill in Wellington. (rail: Manor Park). Dry Creek is a working quarry and the sets have been removed.

Farm Stays - Matamata

Wairere Falls Walks - Matamata
Start: Goodwin Road, off Te Aroha-Okauia Road Time to first lookout: 45mins Time to second lookout: 45mins

Misty Mountain Paths
The 'Misty Mountain Paths' where the Company pass behind a majestic waterfall before negotiating the bluffs towards the 'Misty Mountains' were shot at Earnslaw Burn, a few miles north of the tip of Lake Wakatipu on Glenorchy Paradise Road.

New Zealand Middle Earth Map
This Map show which parts of Middle Earth Corresponds to which New Zealand area.

New Zealand Wars - Wikipedia, the free encyclopedia

Figure 4.24.: Example Result of a WiPo Search.

4.3.5. Scenario 3: Healthcare

Case Description Similar to other sectors, the domain of healthcare has experienced a tremendous increase in technology usage. This is, for instance, evident by the amount of technology used by hospital administrations and clinical trials, as outlined in numerous studies [For14; FBL+15; FV12]. Technology is also widely used in routine tasks such as ordering medication [Luc13], during paediatric anaesthesia [LC11] and when interacting with patients, for example via patient questionnaires filled out on iPads [FBR+12]. Besides, consumer health apps, such as RUNTASTIC[37] and my MYFITNESSPAL[38], have become more popular than ever before [Mac14].

In this context, the idea of a healthcare WiPo fits well, the aim being to utilise WiPo's strength and provide integrated information on a rare medical condition to support patients. This case was also described in [DSV12]. In this scenario, patients are considered who want to know as much as possible about their illness, integrate this knowledge with their personal experience, and share it with like-minded people. This makes it the ideal case for a crowd-sourced community application of WiPo.

Typically, such a community would form around an illness that at best has a long recovery time or at worst is incurable, such as some forms of cancer. In a way, this application will benefit even more from the involvement of many users (patients) compared to the tourism case. Extending this thought, WiPo could be turned into a dedicated social network enriched with information regarding the disease. Thus, the system would be connected to medical (journal) databases and be fed with blogs of affected people, relevant medical support groups, and research centres to name but a few possibilities. It stands to reason that a medical or scientific expert curates the content in order to ensure it is of the highest quality.

In theory, users could supply their medical records, if available, from a private source such as a national health database or simply upload files containing this information. Also, they can provide helpful URLs such as those of a trust which is funding research into the disease. They can then search through the database to retrieve all information relevant to them. Extending the community model, they can share information and receive advice from others.

Social Dimension *Users* of the system will mainly be (severely) ill people. However, WiPo must not be limited to patients but should also be accessible by care-

[37] https://www.runtastic.com/, accessed: 2015-05-31.

[38] http://www.myfitnesspal.com/, accessed: 2015-05-31.

givers or family and friends who want to support patients as much as possible. As a result, the group of primary users could be from all age-groups and have varying technical skills. Therefore, the tool should be easy to use and provide good quality, clear information. In contrast to other scenarios however, the tool is not addressing a mass market and can thus be slightly more complicated and might require some learning as long as it is worth the effort for the user.

A second user group might be medical staff such as doctors, nurses, or researchers, who become part of the community and in that way are able to a) gain knowledge about potentially new fields, b) share their knowledge if they are already active in that field, c) gain new insights / carry out research in the field in order to incorporate first-hand experiences or recruit patients willing to partake in medical trials. This group actually can be both, *Users* and *Curators*. As users they probably are no different from patients and their relations. However, as curators they may have special requirements regarding the usage which should be similar to known tools of medical practitioners.

Political/Legal/Ethical At first sight, there seems to be nothing unethical with this idea. As long as data is of good quality and is well curated there is nothing unethical, but if data is not vetted correctly and pseudo-scientific information is uploaded then this could be potentially harmful to patients' health. Nevertheless, if strict guidelines are followed, this will actually be a better tool than normal search engines as the information presented to them will be evidence-based medicine and will not include pseudo-science and other harmful pieces of information. However, this issue soon becomes complicated when considering the financing of this service, as there can arguably be big ethical concerns with letting only those people access information that may be of vital importance to them who are willing to pay. This will be discussed later when considering the appropriate business model.

In this context, another legal and ethical question arises, namely the question of responsibility for advice taken from WiPo and whether WiPo has to be registered and approved as a medical tool by authorities. While the eventual decision on that matter has to be taken by solicitors and authorities, on a practical level, including medical researchers and practitioners in the process will help adding to WiPo's credibility and minimise the risk of inaccurate or even dangerous information. Furthermore, including a disclaimer stating that WiPo may serve as information tool but does not replace professional advice, can be a first step.

Copyright, in contrast, is probably less of an issue because data sources are limited and made available only to a small group, which – depending on the organisation – might not even be public. Furthermore, advanced content such as medical journals can be licensed.

Careful consideration will need to be given to *Privacy* and security of the stored data. This is because the stored patient data is potentially highly sensitive and users will not want it to leak. In order to address this need, customers will have to be informed in detail on where and how their data is stored. In addition to providing this information, technical means, such as encryption, are to be used when storing sensitive information.

Economic/Business Dimension From a philanthropic point of view, this use case is one that should not be run for profit but should be available for free[39], due to the fact that it is unethical for patients to have to pay for a service that could help them. Thus, a subscription or even a freemium model is out of the question. However, costs of running have to be covered. While they may be rather small considering that cloud computing provides large scale resources at rather cheap prices, costs can explode if licensing fees for expensive databases or scientific journals have to be covered. Another potential cost is human resources; however, this could be overcome using a private crowd, i. e., a crowd with restricted access to ensure all members share the philanthropic thought behind WiPo similar to medical practitioners working for Doctors Without Borders[40]. The information generated by this use would be of great interest to scientific experts working on the disease. Unlike medical doctors who often work on numerous diseases a scientist – or even a PhD student working on this one specific disease – would probably contribute time to curating the information, and it would be of high quality due to the rigorous scientific method. Since having high-quality information would greatly benefit their research, contributing content would be a reciprocal relationship and therefore highly attractive to scientists. Alternatively, employers might allow their medical staff to contribute to the healthcare WiPo for a limited time during their shifts either because employers share the philanthropic view or because they hope to gain greater reputation.

Running cost will most likely be covered by donations from a third-party or the patients themselves. Advertising may also be an option, but it needs to be

[39] Confessedly, this might not be possible depending on the data that needs to be acquired. Nevertheless, this is the ideal.

[40] http://www.doctorswithoutborders.org/, accessed: 2015-05-31.

obvious that the advertisement is clearly separated from the information and that advertisers have no influence over content, as this could lead to exploitation of WiPo by, for instance, pharmaceutical companies keen on pushing their product to make a profit. A good advertising candidate could be a medical research charity with an interest in the disease in question.

Overall, it seems a better option to cover running cost using donations. Increasingly, charities and individuals are using the viral effect of the Internet to boost donations - this could be a good option for a medical WiPo. For instance, in 2014 the Ice Bucket Challenge went viral [Jul14] and resulted in over 100 million dollars being donated within 30 days to the ALS Association [Car14]. However, virality of Internet phenomena is not predictable and should not be counted on. That said, there are studies such as [WMA13] focusing on predicting virality but they do so based on observations once content is online not beforehand.

Similar to the tourism use case, the *Organisational Structure* of the medical case will be a non-commercial organisation providing the platform. *Potential Providers* are providers of medical databases, scientific literature, and Internet sources from blogs or forums. Those providers might be willing to give away their content for a discount or even for free but that cannot be said for sure. Furthermore, given the arguments presented above, the discussion of *Potential Customers* seems inappropriate and is omitted.

Curation of this use case should be human as the data needs to be high-quality and algorithms may be more error-prone in some case, leading to potentially fatal information. Community building is very important in this use case to provide valuable information, first-hand experiences from affected people, and also to provide the feeling of not being alone. This use case has the most demands towards social features as it is probably the most emotive of the cases looked at. It should be ensured that all content published, even crowdsourced content (provided by community members), has been double-checked by medical staff to avoid potentially damaging mistakes. As a consequence, a new editorial system will have to be added to WiPo that allows for different privileges for different people. Also an authentication method has to be employed to make sure someone claiming to be a doctor actually is one.

In regard to *Potential Competition*, UpToDate[41] should be mentioned. This company provides a curated medical information database on a subscription basis. Their service is targeted at medical professionals, as well as at patients. Besides UpToDate there are numerous databases, such as PubMed[42], which

[41] http://www.uptodate.com, accessed: 2015-05-31.

[42] http://www.ncbi.nlm.nih.gov/pubmed/, accessed: 2015-05-31.

provide access to medical journals. However, these do not provide to-the-point information integrated with a patients' own information nor do they allow inter-patient communication. Therefore, it has to be said that competing products do exist; however, their provided utility greatly differs from WiPo's features.

Ecological Dimension The *Direct Ecological Consequences* for this case can again be reduced to that of green IT. Nevertheless, as the expected data is rather small the ecological impact of such a solution is no more significant than that of running an individual Web server and is, therefore, negligible. *Indirect Consequences* are not expected.

Technical Dimension From a technical point of view, the main task of curation in this use case is the integration of various sources and databases. The number of sources in this case is likely to be far lower than in the tourism case, and therefore will be easier to integrate as most are probably well-known in advance or, if discovered by the system, small in number. A more challenging task would be the integration of many private health records. Patients could be asked to provide their profiles in a standardised format; however, such an interface would still need to be developed.

As a consequence of the above, *Source Selection* is initially done primarily by the curators to provide the best quality possible. Nevertheless, user-supplied URLs will also be checked in this process.

Regarding *Data Mining*, the main challenge will be to analyse health records. Thus, entity extraction can be very important. Furthermore, association rule mining (analysing dependencies in the data) and collaborative filtering (making use of the community) can be applied.

The *Service Response Time* has to take into account that a medical service is intended for continued usage over a longer period of time – for the whole illness. Thus, users might tolerate longer response times, for example, in the range of days. As suggested for tourism, a long term service definitely requires a functionality that informs users of new results that were calculated asynchronously. This is a very important functionality for this use case and it is also linked to the *Data Freshness*. On the one hand, users want to receive new information as quickly as possible but on the other hand they may tolerate longer response times – at least as far as advanced results are concerned – if this means the information is well adapted to their specific situation. Bearing this in mind, as well as the fact that the data is not to be expected to be too large, regular updates on a weekly or, even better, daily basis are appropriate. In contrast, discovery of new sources may be in the range of weeks. As soon as new information is available this should be matched with user profiles; the computation of which should take no longer than 24 hours. Furthermore, queries that can be answered on the curated database should return results within seconds.

Similar to other described cases, *User Input* will be keywords potentially enhanced by a query-based set of URLs. Specific to this case is the possibility of users to provide health records to build their profile data. This input can then be enhance by keyword extraction and other classifying techniques in order to select appropriate sources and calculate the most relevant set or results.

As far as *Visualization & Output* is concerned, this case is focused on textual (medical descriptions) and image data (medical photographs, x-ray images, etc.) that have to be presented adequately. This type of information is rather simple and can be displayed with the current prototypical implementation. However, integrating the necessary social communication aspects requires significant work.

4.3.6. Scenario 4: WiPo for Business Environment Analysis

Use Case Description This case description extends on that of a company wishing to keep abreast of their market situation originally described in [DSV12] and it will be discussed how a business can keep up-to-date regarding their environment. As known from the introduction to this section a viable tool for doing this is the PESTEL analysis [CPT10; CIP13; Rap07; Jur13], which can also be used as a tool not only for ensuring compliance with the law but also acquiescence with the morals and implicit expectations expected of a company, as investigated in [Ras14]. In this respect, this subsection makes use of the fact that PESTEL has been explained already.

It is supposed that a company is wishing to automate or at least computationally assist their scanning of the environment using a PESTEL analysis. Scanning the environment for such information is costly as there is an almost an infinite amount of information available [Agu67]. This is why a centralisation and automation process as proposed here can be really beneficial for a business like a large agricultural company.

Agriculture is a rather complex example as there are many factors influencing production that are out of the control of the business itself. Therefore, it is very beneficial to scan the complex environment for potential changes.

Here, it will be shown that a PESTEL analysis can be automated and meaningfully applied to an agricultural business case study, with the PESTEL dimensions discussed in PESTEL order. In this context, *Service Composition* is addressed in each dimension without explicitly referring to it as such.

The *Political* dimension is of great importance for agricultural businesses as it significantly affects the way they operate. For instance, agricultural businesses in the European Union (EU) are heavily subsidised. If that were to be changed in favour of ecological rather than traditional agriculture, traditional farms would face difficulties. Besides subsidies there are numerous regulations on farming. If, for example, certain cheap pesticides are prohibited for ecological reasons, the harvest and subsequently the turnover might shrink while at the same time costs increase causing a real danger for the viability of the business. Sources to tap for these dimensions are commentaries from specialised solicitors, lawyers, and, most likely, agricultural bodies and their publications, such as Federated Farmers of New Zealand[43] and publications from the former.

Regarding the *Economic* dimension, the market needs to be observed. On the market, a particular important piece of information is the sale price, which is volatile, as crops are traded and speculated on various national and international markets. To this end, the business needs to know up-to-date prices for the grown crops as well as future projections. Therefore, meaningful sources to tap are buyers as well as news agencies, such as Thompson Reuters Agricutlure Professionals[44] for projected prices on a given market.

On a *Social* level it is important to observe consumers and their behaviour. Interesting questions are for instance: What is the overall attitude of society towards traditional farming or ecological small-scale farming? And what do consumers think about wind turbines and biogas plants as sources of electricity? The former question influences the decision on what way to produce crops and the latter influencing the decision on how to best use the land, which is related to the question of opportunity costs. To answer such questions, relevant sources have to be found and analysed. These may be from online forums but also news sites addressing these issues.

Concerning the *Technological* level it can be stated that, like many other industries, agriculture is facing an ever faster technological change. This includes first and foremost the use of IT in production. For instance, it is possible to analyse aerial photographs of fields to determine which areas need more or less fertiliser. Along with this goes a huge market for software as well as hardware such as drones which can save farmers plenty of time by virtually flying over their

[43] http://www.fedfarm.org.nz, accessed: 2015-05-31.

[44] http://financial.thomsonreuters.com/en/products/tools-applications/trading-investment-tools/eikon-trading-software/agricultural-commodities.html, accessed: 2015-05-31.

fields. To include this, relevant blogs and vendor pages have to be tapped to receive the latest information. Furthermore, agricultural bodies and consultancies can be a valuable source.

The *Environmental* dimension is probably one of the most important when talking about farming. On the one hand, there are regulations in place, as previously mentioned in the political section, while on the other hand the weather and effects of global warming are very important as these may change the ability to grow certain crops in the long run. For instance, if the groundwater level decreases and dry periods become longer, the richness of the soil will decrease as well. Another matter currently discussed is the salinisation of soil [Pfl14]. To keep abreast of these trends, actual and historic weather data as well as information on soil quality have to be gathered for instance from meteorological services, agricultural researchers or conservation organisations. This enables forecasting of trends relevant for agricultural businesses.

Finally, the company has to be up-to-date with a number of *Laws* and regulations such as approved seeds, crop sequences and legal fertilisers and pesticides and the conditions under which they are to be used. For this kind of information, it is important to tap authoritative sources which could be governmental organisations, farming associations, and agronomists. Another important topic in this dimension is that of labour law and work safety law which need to be obeyed. Similar to the agricultural regulation an authoritative source such as online law collections (e. g., the German Gesetze im Internet[45] which holds nearly all German legislative texts) needs to be tapped. This is inherently connected to the political dimension discussed above as the mentioned regulations can change quite quickly.

As evident from this, all dimensions are rather interwoven and influence each other. That is why using WiPo can be very beneficial in this context. In this way, sources can be analysed in an integrated manner which ensures the information is the most recent and relevant. Furthermore, offline capabilities allow for carrying information when in the field, where Internet connectivity is often very weak. Having described the use case in detail, the next paragraphs will discuss WiPo in this case using the SPEET dimensions which must not be confused with the use case itself.

Social Dimension In a general business use case, *Curators* of WiPo are internal strategic analysts of businesses who wish to partially automate their data collection when doing strategic planning. As none of the individual use cases in this

[45] http://www.gesetze-im-internet.de/, accessed: 2015-05-31.

use case class are quite alike – simply owing to the fact that no two companies will address strategic planning alike – the tool will have to address a large array of needs. In the general use case, *Users* are likely to be top-level managers who are used to a report-like output which should be addressed by WiPo.

In this particular agricultural use case, the differentiation between *Users* and *Curators* is probably less relevant. It can be expected that there is no such formal organisational structure as for big corporations. Furthermore, curation and usage will probably be done by the same person first, setting the scope and integrating the desired sources and then evaluating the retrieved information using the system. This much is true when WiPo is to be sold on an instance basis. However, anticipating the economic discussion, an alternative is to sell access to a service for an entire industry, e. g., agriculture. In this case curators would be professionals from the farming industry such as specialised solicitors, market experts, and agricultural researchers. In this case the users would be farmers and agronomists. However, their technical capabilities often vary. Thus, the system has to be simple to use, yet informative which is a GUI challenge.

Political/Legal/Ethical Dimensions Since this case is more business oriented and taps professional sources, there will commonly be established licensing models to obtain data. Thus, *Intellectual Property* is less of an issue. Similarly, *Privacy* is less relevant because a business and not a private person is dealt with. However, strategic planning may be highly confidential, hence, measures will have to be taken to prevent search queries or even profiles of companies to be overheard or compromised by a third party. This, however, is only relevant if WiPo is to be deployed as service rather than as an instance.

Economic/Business Dimension As briefly mentioned already, this scenario offers two different *Business Models*. First, WiPo could be licensed on an instance basis, i. e., companies license the software and configure and run it themselves. However, it is quite likely that companies would need technical support. While this allows for a maximum security and flexibility, it also means that there are no economies of scale, as every company does more or less the same thing. As a consequence, in particular for the agricultural case discussed here, it makes sense to provide companies with a service that they can query. This offers the big advantage that every company can consume the most relevant information for them without the need for their own curators. This means determining the *Potential Customers* in this case is simple – the users are agricultural businesses with their specific information needs.

Regarding *Organisational Structure*, this WiPo case resembles a standard company with departments such as development, marketing, and accounting. However, a special department would be the curation department where a number of experts gather and curate information for the target industry. Given the professionalism of this use case these would most likely be expert employees.

As evident from above, *Potential Providers* are any domain specific information providers. Depending on their bargaining power, it might be difficult to negotiate economically viable terms. Then again, if WiPo serves as an intermediary the organisation behind it should be able to find a solution that provides the information cheaper than they would be if the individual customers were to buy it.

Potential Competition would be numerous services offering similar information individually. However, providing information in an integrated manner may give WiPo an advantage over them; in particular if the information provided by WiPo is extremely reliable. However, WiPo may be more expensive as it has to cover the cost of operation and could potentially provide too much information for an individual business to handle.

Environmental Dimension *Direct Consequences* for the environment are basically identical to the previous cases and shall therefore be omitted here. Nevertheless, the *Indirect Consequences* are extremely relevant as agriculture is a nature-based business and information provided through WiPo could have a significant impact on the environment. As an example, WiPo could hold information that a certain legal pesticide, which harms wildlife, is far more effective than others. This might result in an increased usage and a decrease in wildlife. That said, this quickly turns into an ethical discussion for which no ultimate answer can be given here. Yet, it shows that WiPo needs to be aware of its role as a multiplicator.

Technical Dimension *Curation* in this use case is probably more the task of finding, integrating and organising a number of trustful sources than that of conduction quality checks on a couple of thousand Web entries. As a consequence *Data Mining* is not as pronounced in this use case. Nevertheless, it might be of importance with regard to analysing trends and discovering changes in the environment. The key curation process in this WiPo case is taking the mined data, as well as the data directly provided by suppliers and potentially by users (e. g., information on their location) and applying a set of context-specific rules to produce the required high-quality information output.

Similar to the health case, *Source Selection* is initially done primarily by curators to provide the best quality possible. Nevertheless, user-supplied URLs will also be checked in this process.

The *Data Freshness* requirements are difficult to generalise as there is information which is likely to be static and other information which is highly volatile. A daily update frequency of the database should be sufficient for the more volatile information, such as weather data. Laws are probably not updated at a high frequency but it should be ensured that the information stays accurate. Thus, for this type of information even an on-demand update triggered by curators is sensible.

User Input in this case will mainly consist of keywords regarding topics in one of the domains. But it could also be files containing information on their field which can then influence the information provided on the platform regarding soil conditions in a certain region. More importantly, analyses of ideal crops for the given soil could be provided by WiPo. Additionally, users can supply URLs which may extend the body of knowledge on the system.

For *Service Response Time* this means that a query should yield results with a delay of no more than seconds based on the data gathered a priori. However, a notification function should be implemented which notifies users of changes to articles they are interested in or alerts them whenever new documents have been added. As far as additional services such as crop recommendation mentioned above are concerned, the response time might be longer but should not take more than minutes.

At a first glance this use case is similar to many other use cases regarding *Visualization & Output* as WiPo will mainly provide textual and graphical information. However, as soon as trends and developments are included, comprehensive charting becomes necessary.

4.3.7. Case Comparison

WiPo is claiming to be generically applicable and so far analyses suggest that it can indeed be helpful when applied to the cases discussed. Furthermore, the technology has the potential to help in many other cases if tailored appropriately. This case comparison section shall present and discuss the similarities and differences of the presented WiPo cases to allow for some generalisations regarding various WiPo use cases. To this end, this section will again be structured according to the SPEET dimensions preceded by general remarks regarding the *Use Case* attributes *Proposition & Experience* and *Service Composition.* A summary will be given in Table 4.4 at the end of this section.

Case Description While inherently different, the use cases have a similar *Proposition & Experience*. All of them involve integration of some sort and offline availability to some degree. In the tourism and LandSAR cases (see Section 4.3.4 and Section 4.3.2, respectively) the emphasis is more towards offline availability, whereas in the other two cases the emphasis is on integration. The *Service Composition* in all cases comprises public and semi-public Web sources as well as private sources such as specific files or specific databases. As a consequence, use cases can be concerned with structured sources (healthcare, Section 4.3.5, and business, Section 4.3.6) or unstructured sources (tourism, and LandSAR)[46]. Given that all use cases have similar characteristics only with different weighting, the more important characteristics are set in bold in Table 4.4.

Social Dimension In the social dimension *Users* and *Curators* vary from technical experts to laypeople. In the tourism case expert users can be supposed but curators are not necessarily experts, in the healthcare case both are not technical experts. For the business case curators are experts in their fields and employed to curate and hence will be familiar with the system while users are not. Finally, for the LandSAR case both groups can be given extensive training and can therefore be considered expert users. Thus, the tailoring always has to include an extensive target user and target curator analysis in order to provide a system that fits the needs and technical abilities of the user.

Political/Legal/Ethical Dimension It is very hard to generalise regarding ethical aspects of WiPo because ethics are very use-case-dependant. The question of whether people in need should be granted access to information through WiPo irrespective of their liquidity (e. g., the healthcare case) is a completely different question to who should be allowed to use WiPo in general. For instance, is it ethical to sell WiPo to armament companies? Or, is it right to make a chemistry WiPo available to the open public which may give terrorists information they would not receive otherwise? These questions cannot be answered per se but highlight that knowledge and the sharing of knowledge always bear responsibility.

As noted before, it has to be taken into account that this work has been developed in the context of information systems and computer science rather than in the area of law. Hence, the legal aspects will only briefly be discussed without

[46] Having presented cross-references to all cases once, they will be omitted in the following to improve readability.

going into detail. In particular, the focus will be on *intellectual property*, including aspects of copyright and licensing as well as *privacy*. Furthermore, the attribute of *liability*, which was already mentioned in the healthcare case, will be discussed more generally. More precisely, it will be investigated how relevant it is in each case that the operator of WiPo can potentially be held liable for incorrect information.

The assumption underlying WiPo is that content providers publish their work online with the intention of it being read and used by interested people, which is facilitated by WiPo. However, over the past couple of years there has been a disagreement between search engine providers, such as Google, and various publishing companies and organisations over the *intellectual property right* of publishers in Germany. Publishers wanted to receive some compensation for providing content or at least parts of content to search engine providers. This eventually resulted in a change of law being passed in summer 2013. The change concerned §87f to §87h UhrG [Urhed], which give publishers the right to compensation for snippets of texts and images. At first publishers permitted the use for free [tag13]. Later, some publishers tried to enforce monetisation but those who did were simply removed from a result set [Nig14] or reduced to a minimal inclusion [tag14b]. This has led publishers to demand the Federal Cartel Office take action, but it refused on the grounds that there was no evidence for an abuse of a dominant market position [tag14a]. This lead most publishers to recant [tag14c] and make their content available for free.

In contrast to search engines, it is WiPo's declared aim to build a database from existing texts and deliver them. In light of intellectual property rights, this is unlikely to be legal for open public services, such as tourism, without the consent of the content owners. A legal solution has to be found to address this issue. However, in the case of close group usage as for a company, LandSAR, and even the medical support groups (if closed) this seems to be less problematic. That said, these cases will also have to be examined by expert solicitors.

A starting point, avoiding these pitfalls, could be focusing on sources publishing their content under open licenses such as CC BY-SA 3.0, or GNU FDL which Wikipedia is using [Wiked]. However, this has considerable organisational overheads. Furthermore, there are many licences used on the Web, such as the widely used Attribution-NonCommercial-ShareAlike 2.0 Generic (CC-BY-NC 2.0)[47] license, which only allows non-commercial usage. This would be sufficient for some WiPo use cases but not for others. Thus, it is advisable to seek permission of the content owner to use their content in a commercial

[47] https://creativecommons.org/licenses/by-nc-sa/2.0/, accessed: 2015-05-31.

way. In some cases, such as the tourism case, some providers (e. g., tour guides) would probably want to share their content because it can be seen as free advertisement.

Regarding the use cases and *Intellectual Property*, it can be concluded that tourism has probably the biggest difficulties because a number of Web sources should be integrated. All others are either not open to the public or incorporate content that can rather easily be licensed such as medical databases.

Interestingly, WIPO – all capitalised – is also the abbreviation for *World Intellectual Property Organization* a United Nations agency which aims at leading the development of an international intellectual property system [WIPed]. This was unknown to the authors when creating the acronym WiPo for Web in your Pocket.

Similarly, *Privacy* is very difficult because no guidelines exist that are internationally consistent. However, the overall relevance for WiPo is little, given that WiPo will fully disclose how profile data is used and what information is stored. That said, privacy is of different importance depending on how sensitive the stored information is. Regarding the four cases looked at here, it can be stated that healthcare and LandSAR are the most demanding in terms of *privacy*. While the business case can be attributed with a medium sensitivity, the tourism case is least demanding.

Finally, the issue of *Liability* shall be briefly touched upon. As discussed previously, this is particularly relevant for the healthcare case because in this sector very strict rules exist by which products and services used in medical treatment need to be approved. The liability in the tourism case can be compared to that of a regular travel guide book which is practically close to zero. In the LandSAR case curation is done by the organisation, hence, liability risk of such a tool should be minimal, too. In the business case, in particular in that with a service provided by professional curators, liability might be an issue and specialists should be involved to minimise the risks for WiPo.

Economic Business Dimension Generally, it can be said that plenty of information available on the Internet is free. However, this goes along with a high cost for searching and evaluating the quality of the information. Thus, some information seekers are willing to pay for domain-specific high-quality content [GRC05]. Therefore, regarding *Potential Customers*, WiPo can operate in three modi. First, it can be run as a service for customers, which involves active curation by WiPo (the business case). Secondly, it may be run as a service in which curation is done by the customer and only the infrastructure is provided (the

LandSAR case). Thirdly, it may be operated for consumers but financed by a third party (tourism and healthcare cases). In this last case the reasons for financing might vary and the examples altruism (healthcare) and profit increase through advertisement (tourism) probably form the extremes. Consequently, there are three types of customers, those who are interested in information, those who want a platform for their own use, and those who want to benefit[48] from supporting the platform. This has to be kept in mind when realising the platform. In particular if users and customers are not the same, careful consideration has to be given to the question of to whom the focus should be given: the customers or the users.

As a result of the above, there are three well-fitted *Business Models*. For those customers seeking information, a subscription model seems appropriate if they seek it more than once and rely particularly on the freshness of information. This can possibly be combined with a freemium model in which a number of requests may be free or certain additional function may only be available to paying users. Out of the cases discussed, this is true for the business and tourism case. The second model is a classical software licensing model, in which customers pay a one-off fee for usage – or more likely in the WiPo case – a subscription fee to continuously use the software. This makes sense if the infrastructure is provided by WiPo. Out of the cases looked at, this is most applicable to the LandSAR case. Finally, an advertisement or donation based strategy could be used by tourism and healthcare cases, respectively.

While it is not possible to list all individual sources, *Potential Providers* can be classified into two categories. They can either be established suppliers of information, such as database providers, or they can be operators of websites with public information. However, most use cases – as all of the discussed – are likely to interact with both categories. From a business point of view, it is important to keep in mind that these two types exist and that a WiPo operator will have to discuss use-case-dependent terms.

Similarly, no general statements can be made regarding the *Potential Competition*; instead, a WiPo business has to look at its competition on a case-to-case-basis. In some cases there is hardly any competition and other areas have established services which may be hard to compete with. Other than that, for the cases under investigation it can be said that tourism has the highest expected competition whereas business as well as LandSAR will face less competition. In contrast, the healthcare case has the lowest expected competition.

[48] It is supposed that having a platform is the ultimate goal, i. e., benefit, for an altruistic sponsor.

Finally, the organisational model may vary but can roughly be divided into two categories, in-house curation and curation conducted by a third party. However, this differentiation may not always be entirely selective. The tourism case would need to be classified as third party, but then again crowd curation might also include some in-house experts; the same is valid for the healthcare case. In contrast, the other two cases can be clearly assigned to one of the two (business → in-house; LandSAR → third-party).

Environmental Dimension As stated when defining the *Environmental* dimension, the *Direct Consequences* are not yet the main focus of standalone IT systems. That said, this issue is of equal relevance to all use cases, not only those discussed here. In contrast, the *Indirect Consequences* vary and are strongly use-case-dependent. Out of those cases discussed the healthcare case has probably hardly any effect. All others are not really predictable. Based on the assumption that people do not inherently care for the environment, the tourism case and the business case will potentially have a negative impact. However, if the right information is communicated well enough it may also have a positive influence, which is most likely in the LandSAR case.

Technical Dimension From a technical point of view, the main task of WiPo is the integration of various sources and databases. This is in contrast to conventional algorithmic search engines, which perform well in terms of automated information gathering but are limited in regard to quality assessment and do not provide for integrated information presentation which WiPo achieves by curation.

The use cases show that expert *curation* and crowd *curation* can be differentiated. All of the cases are supported by algorithmic extension such as automated updates. The curation tasks also vary between use cases. Whereas healthcare, business, and LandSAR focus on the integration of a relatively small number of high-quality sources and quality checks thereof, the tourism case focuses more than the others on collecting a very large body of knowledge in one place. That is why this use case requires crowd curation. While healthcare may be applicable to both, the other two will most likely be curated by an individual or a small group of experts.

Closely related to the topic of curation is the area of *Source Selection*. Sources should initially be seeded by curators. For those cases in which a number of pre-known sources have to be integrated (business, LandSAR, and partially healthcare), the selection will mostly be done manually. Given the sheer number of sources in the tourism case and also partially in the healthcare case, the source selection has to be algorithmically supported by focused crawling. That said, even crawling has an ethical dimension to it. Koster [Kosed] claims to have developed the Robots Exclusion Protocol in the early 1990s. This protocol, which is based on a text-file (robots.txt), allows server administrators to tell robots whether they are welcome to access a given page or not. However, there is no need for a robot to obey this request and the content can simply be crawled anyhow. Nevertheless, it is good practise to obey these files but this it is an ethical decision whether to do so.

Regarding *Data Mining*, a number of mechanics have been identified for the different use cases. Among them are Link Analysis, Clustering, Entity Recognition, Collaborative Filtering, Association Rule Mining, and Trend Analysis. From this it can be seen that there are various tools available. However, each use case has its specific needs. While all use cases can make use of Link Analysis to determine general dependencies, the tourism case can particularly benefit from clustering in order to centre documents around topics and collaborative filtering to recommend appropriate items to users. Healthcare and LandSAR can make good use of association rule mining and pattern matching and the business case benefits from trend analysis. Nevertheless, it should be kept in mind that these are only starting points; an informed decision has to be taken when actually implementing the use case.

As far as *Service Response Time* is concerned, it can be said that in any case an instant reply from the database can be expected. Differences can be observed for the duration of additional calculations with the LandSAR case being the most time critical. Response time is also critical for the tourism case because users expect instant replies. For the other two cases it is less relevant how quickly new information are retrieved, as long as the returned information is accurate. The same is true for *Data Freshness* which has to be high in all cases to ensure information is not outdated.

User Input has been restricted to keywords, URLs, and files. The difference between the use cases lies in the types of files that are submitted. In the tourism and healthcare examples files are used to extend the profile and potentially extend the knowledge base. In the business case user-supplied documents are meant to be analysed and returned with enhanced information, while in the LandSAR case documents are submitted to be reviewed by search managers (curators).

Table 4.4.: Summary of Comparison.

Dimension	Attributes	Tourism	Healthcare	Business	LandSAR
Use Case Description	Proposition & Experience	**offline availability**, integration	offline availability, **integration**	offline availability, **integration**	**offline availability**, integration
	Service Composition	**unstructured sources**, structured sources	**structured source**, some unstructured, files	**structured sources**, few unstructured, files	**semi-structured files**, (un)structured sources
Social	Users	technical	non-technical	non-technical	technical
	Curators	non-technical	non-technical	technical	technical
Political/Legal/ Ethical	Intellectual Property	high	medium	medium	medium
	Privacy	low	high	medium	high
	Liability	low	high	medium	medium
Economic/ Business	Potential Customers	seeking publicity	n.a.	seeking information	seeking a platform
	Potential Providers	private & public	private & public	private & public	private & public
	Potential Competition	high	low	medium	medium
	Organisational Structure	in-house & third-party	in-house & third-party	in-house & third-party	in-house & third-party
	Business Model	advertisement	donation	subscription	licensing
Environmental	Direct Consequences for ENV	green IT	green IT	green IT	green IT
	Indirect Consequences for ENV	potentially negative	neutral	potentially negative	neutral/positive
Technical	Data Mining	clustering, collaborative filtering	association rule mining, pattern matching	trend analysis	pattern matching, association rule mining
	Service Response Time	high/medium	low	medium/low	high
	Data Freshness	high	high	high	high
	Visualization & Output	multi media	medical imagery	charting	maps
	User Input	files for profiling & new knowledge	files for profiling & new knowledge	files for analysis	files for new knowledge
	Source Selection	focused crawling	manual	manual	manual
	Curation	crowd	crowd/experts	experts	experts

Visualization & Output can be described as similar but different for all cases. More definite is the fact that all cases have the same output basis, such as textual and image data but all have different special requirements. Tourism requires advanced multimedia usage for films, healthcare requires the ability to appropriately show medical images, business requires advanced charting, and LandSAR has very special map visualisation requirements.

From all this, it can be seen that while use cases can be inherently different, they can be gauged by the same sets of attributes and it stands to reasons that this framework is generally applicable to all WiPo use cases. While it is likely that attributes may have further manifestations, the cases looked at here can be considered rather comprehensive because of their diversity. Thus, the cases presented can serve as references when new use cases are to be implemented.

4.4. WiPo Conclusions and Future Work

In the preceding sections the concept of WiPo has been introduced and the initial implementation has been described. This was followed by an extensive discussion of four use cases. It was shown that WiPo, as an information platform and single point of truth that pushes information to the user, can be particularly beneficial if a number of sources for a well-specified topic have to be integrated, kept up-to-date, and possibly persisted offline on a mobile device – all enabled by means of manual curation. Comparing WiPo to currently existing tools, it can be said that WiPo is a manually curated information repository, similar to Wikipedia – but WiPo is also self-updating –, with enhanced search functionality, as provided by common search engines, such as Google.

4.4.1. Beyond Manual Curation

Currently, curation relies on human experts and is only marginally supported by algorithms. This is the bottleneck of the system as curation is a labourious and time-consuming task. This issue is particularly severe in cases in which time plays an important role. Using crowds, as suggested for the tourism case, is an alternative to entirely manual curation and could, therefore, be the first task to extend the WiPo infrastructure. Previous published work by the author [SV12] has already suggested a move towards collaborative curation, borrowing ideas from Wikipedia and other collaborative online knowledge organisation platforms such as the dmoz and Delicious. However, collaborative or crowd curation goes along with plenty of organisational overheads. Furthermore, it can be difficult to motivate people to contribute to such a tool in order to achieve high-quality content and to resolve possible conflicts between curators – often the right incentives need to be given.

These potential pitfalls have been well studied on Wikipedia, probably one of the best known knowledge repositories built and maintained by a crowd. In [SPM11] a brief review of literature on these issues is presented. As an example, Wagner and Prasarnphanich [WP07] state that altruistic motivation occurs more often than selfish motivations among contributors to Wikipedia. Kittur and Kraut [KK10] describe the ideal group structure based on an analysis of over 6800 publicly available Wikis as having a core of leaders who do most of the work. This shows that it is not clear how well a crowd-based WiPo can function and, something which should be subject to future research.

When crowd curation is to be employed, it is important to provide the right incentives. According to Hammon [Ham12], who presents an analysis of forces driving crowds in her PhD-thesis, crowdsourcing should address creative and innovative Internet users. Overall extrinsic but non-monetary factors such as appreciation of work had the strongest driving force. This could for instance be achieved by gamification. In order to address intrinsic motivation, the platform should be developed in a user-friendly manner [Ham12]. Monetary motivation provides a positive effect which, however, is not very strong. Both, monetary as well as other extrinsic factors have a positive influence on the intrinsic motivation. From this it can be inferred that WiPo should provide an easy to use platform that allows users to enjoy their work. Furthermore, extrinsic motivation should be provided through a gamified interface which allows, for example, the achievement of virtual badges as this addresses the motive of gaining social recognition [BL13]. Whether money should be paid has to be decided on a case to case basis [Ham12]. Regarding WiPo this is also a question of economic viability and is probably not feasible in the cases discussed here.

Furthermore, a natural consequence of crowd curation would be to take social media into account. Thus, the crowd functions should be developed in a way that allows for sharing information or curation needs in one's social networks such as Facebook in order to gain wider visibility and acquire possible curators that would not be aware of WiPo otherwise.

Besides manual curation another means of curation is possible, namely a complete automation using ontologies such as *YAGO2* [HSBW13] which is sourced from Wikipedia. At first this might seem to be a renunciation from the idea that manual labour is superior to algorithms with regard to quality assessment. However, the ontology has to be built in the first place. This usually happens at least in a supervised manner, i. e., the ontology is filled by a human supervised algorithm. This thesis proposes to set up a new ontology for a new WiPo based on human expert supervision. Furthermore, it is also possible to employ any combination of manual curation, crowd curation, and ontologies, resulting in the seven combinations depicted in Figure 4.25.

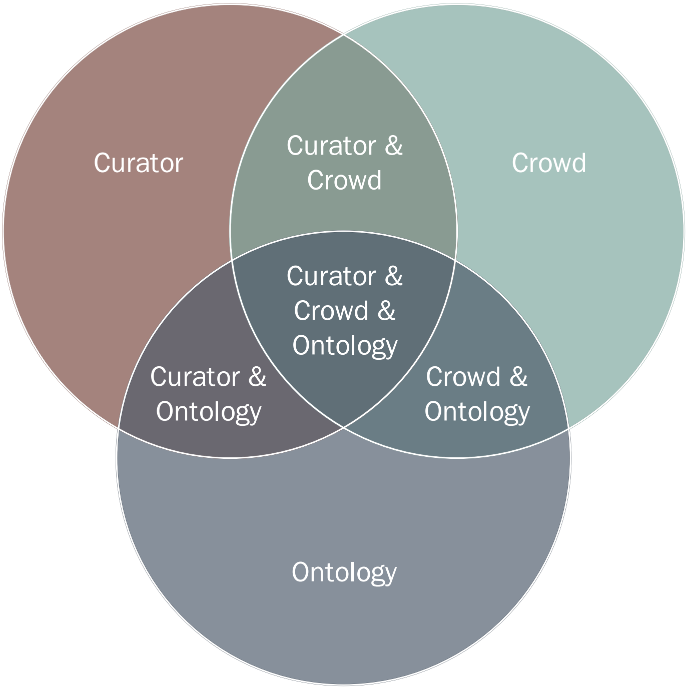


Figure 4.25.: Possible Combinations of Curation Types.

Self-learning algorithms, which are based on expert or crowd decisions, could be used to pre-assess the quality of retrieved documents and discard obvious garbage. Furthermore, they could be used to cluster documents as well as extract keywords and categories automatically, allowing for the linking of results as a means to build ontological knowledge which can then be used to automate curation further. It should be noted that any given WiPo instance may pass through more than one (maybe all) combinations during its lifetime. For instance, it could start as curator only and be then extended by crowd and ontology support. To this end, use cases will have to be analysed in order to determine which of the combinations is applicable and to define transitions from one to another.

4.4.2. Future Technical Developments

Possible future extensions are manifold but can roughly be divided into back-end and front-end developments. Regarding the front-end, there are two things to work on; the first being usability. While the prototype is in principle well-operable the devil is in the detail necessitating improvements. To this end, usability experts should be consulted in order to do an in-depth analysis of the current state and extend on this.

Furthermore, the newly introduced modi operandi such as crowd curation will have to be reflected appropriately. This includes in particular the implementation of communication channels, including the possibility to submit documents for revision, to track changes, and to revert to previous versions, to name a few options.

Additionally – highly depending on the use case – a mobile app will have to be developed in order for use cases to be able to make full use of all offline features. From those use cases discussed, this affects tourism and LandSAR. In healthcare and in a business context this is of lower importance, although it can be useful in those cases, too. For instance, patients could read relevant papers on the train while on their way to their specialist and also carry this information to their medical doctor. A farmer might want to take the information out in the fields. A dedicated app has the further advantage that it can be tailored to the exact needs of a user and that it can improve relevant off-line functionality for specific use cases. Also, it would be a great improvement to both front-end and back-end alike to make it easier to include non-Web sources, such as external databases or user files.

Regarding the back-end, there is even more room for improvement. The main aim being to reduce the burden on the curators' shoulders. The implemented solution can be seen as a skeleton or proof of concept of what is possible.

Starting at the point of gathering data, focussed crawling is a main challenge. This means that adaptive crawling will have to be implemented in such a way that at crawl time it can be decided whether a document is relevant or not. For instance, language could be a criterion. As of now the crawler also picks up Web pages in languages other than English, which are of no use to the currently employed curators. Also, topic determination could be helpful. The idea of focused crawling was discussed as early as 1999 [CBD99]. More recently, the issue of low recall owing to Web structure was addressed [PT10], as was the use of sentiment analysis in focused Web crawling [VCK14].

All of this concerned crawl time. However, there is plenty of algorithmic work that can be applied after document collection and before manual curation to reduce a curator's workload. This includes automatic content analysis and categorisation for instance by means of cluster analysis (e. g., [LRU14]). Further steps in the future should include the application of link analysis as described in [LRU14]. In this way the natural link structure of the Web can be exploited to gain information on which articles to integrate and how. However, this would probably also require a more complex GUI. Further research could also go into the recommendation part, including building privacy-conform user profiles which will yield better search results. This is also elaborated on in [LRU14].

Once fully developed, WiPo should make use of the fact that the architecture was built in a way that it can be run on a distributed system architecture, in particular a Hadoop[49] cluster. In this regard, it will be interesting to see how much performance in particular regarding crawling, analysing, and query time can be improved.

4.4.3. Future Work

Thus far, the conclusion has mainly been concerned with possible technical and organisational enhancements and developments that are useful to improve the utility of WiPo. However, these have all been theoretical in nature. The design science approach suggested by Peffers et al. [PTRC07] and followed here also requires some practical evaluation. This section is dedicated to showing how the initial steps of design science have been fulfilled and how the less intensively discussed parts can be addressed. To this end, the six steps shall be repeated. Furthermore, information is given on how this work addressed them.

Identify Problem & Motivation In Chapter 3 it was pointed out that while algorithmic search engines suffice in most cases, there are particular, topic-centric cases in which they do not. It would thus be helpful to have a better solution.

Define Objectives of a Solution From this, in Section 3.5.2 the objective has been derived to develop a search tool that delivers integrated high-quality information by exploiting curation.

[49] http://hadoop.apache.org/, accessed: 2015-05-31.

Design and Development The design has very extensively been discussed in Section 4.1 based on Petri Nets. The developed implementation has then been introduced in Section 4.2.

Demonstration While not a demonstration in the proper sense of a design science approach, Section 4.3 has demonstrated how WiPo can be applied to different use cases theoretically. A practical demonstration to the scientific community has been done at EC-Web [SGH+14a] regarding the concept and at BTW 2015 [SGH+15] presenting the actual tool.

Evaluation One demo case has been evaluated in a small scale interview series, described in Section 4.3.3. Consequently, this has created a starting point for future research.

Communication This has started with a number of publications describing the concept, the implementation and the demonstration [DRSV14; SV13; DSVR13; DSV12; SGH+15; SGH+14a]. Furthermore, this work is part of communicating the ideas behind WiPo. However, given that the evaluation part will have to be iterated further, additional communication is to be expected.

In summary, this work has laid the foundations for a search process reintegrating humans for quality checks and tailoring results towards users' needs. It was elaborated on a first prototypical implementation, which served as a proof of concept and as a basis for discussion with scientific peers.

An overall evaluation is out of the scope of this work. However, pointers will be given as to how research on WiPo can be continued – following the introduced design science approach. Firstly, a use case has to be chosen because a tool can only be evaluated seriously with a given purpose which is likely to be different for different use cases. Given the experience of this work, the LandSAR case is appropriate because interested partners could be identified, which allows for a concrete application despite the competition.

Once this is settled, three iterations through the design science phases *Design & Development, Demonstration,* and *Evaluation* (*Communication* too if appropriate results are achieved) are suggested. This is in line with the approach by Peffers et al. [PTRC07], whose process explicitly provides a feedback loop. For convenience sake, the graphical illustration of their process is repeated in Figure 4.26.

Then, WiPo has to undergo the improvements identified in the first round of interviews (*Design & Development*). Secondly, the results should be presented to a number of experts conducting a small-scale expert interview series similar to the first one (*Demonstration*). An interview study is recommended as it is of the highest importance to get first-hand information form eventual users. To this end, individual qualitative interviews appear to be the ideal method. Qualitative interviews are commonly unstructured, resulting in a topic centric conversation, or are semi-structured, i. e., following a guide of predefined open questions or topics. The latter allows for flexibility while ensuring that all important aspects are covered and the interviews are comparable. Some authors explicitly encourage departing from the interview guide and discussing tangents in order to get as much insights as possible [BB07]. Evaluating this series will possibly lead to the identification of necessary improvements (*Evaluation*).

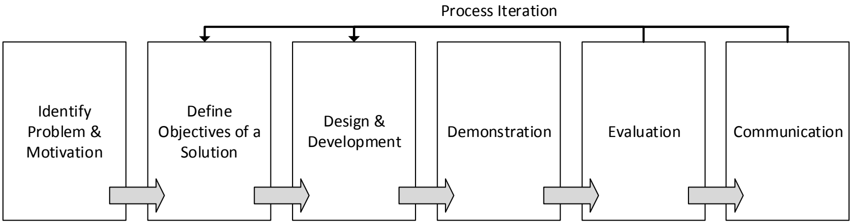


Figure 4.26.: The Design Science Research Process, Adapted from [PTRC07].

After that, a next iteration could then be to implement the identified features (*Design & Development*). Next, the tool should be presented to its later users and they should be allowed to test its functionality over a set period of time – ideally a couple of months (*Demonstration*). The following evaluation step could make use of focus group interviews, i. e., a larger number of group interviews, as studying group context is "probably the most natural method for gathering knowledge [..], especially in an organisational context." [Chr97]. In contrast to individual interviews, in a group setting different opinions are dynamically discussed. This allows for a more complete picture to be gained. However, this dynamic has to be well controlled by the interviewer in order to prevent the interview heading in the completely wrong direction. Regarding group sizes, 5 to 6 participants seem to be a good group sizes. This is also referred to as mini focus groups [Edm99]. It is probably appropriate to form two sets of mini focus groups, i. e., users and curators.

Following this refreshed *Evaluation*, the feedback gained during the overall process will be compiled and implemented in the last iteration proposed here (*Design & Development*). Based on the resulting tool, the *Demonstration* can be renewed by rolling WiPo out to beta testers, i. e., WiPo is run under real conditions and is used by the eventual users. Finally, a large-scale survey based on questionnaires is proposed to evaluate the tool at large. Only then, final conclusions regarding the applicability of WiPo can be drawn (*Evaluation*).

Part II.

Data Marketplaces and Quality-Based Data Pricing

5. Data Marketplaces

Research has shown that companies which use data analysis extensively are commonly market leaders in their domain [Dav06]. Particularly, so-called data-driven decision making is associated with higher productivity and profitability of firms [BHK11]. Consequently, it has been recognised that data has value [Mil12b]. While it is a first step to build a data-driven enterprise that collects and analyses data as discussed by Davenport [Dav06], there is a recognition that external data is also relevant [Ros14; BHK11; MSLV12]. As a result, data marketplaces have emerged that can be seen as an advancement of established information services such as Bloomberg[1] to the data level [Mil12c].

This chapter is dedicated to defining and describing the phenomenon of data marketplaces. Markets and marketplaces for information as well as the marketability of information goods have been explored within the field of economics for some time, e. g., in [MYB87; Bak97; Bat90; LS11]. More recently (in 2011), Balazinska et al. [BHS11] put the topic of data marketplaces on the research agenda of the database community. They identified two main research challenges: firstly, understanding how the value of data is determined and modified on data marketplaces as well as what pricing schemes and services facilitate data marketplaces and secondly understanding the behaviour of market participants and the underlying rules. Balazinska et al. [BHS11] attribute the second challenge to the economic community and recommend the first to be addressed by the database community. The first challenge – as far as understanding value creation on information markets is concerned – has, however, been partly addressed by economists already, e. g., in [LS11]. The problem of value generation on electronic platforms intended to facilitate trading of data, and the technical implementation of such facilitating factors, have mainly been addressed by two research groups; the research group who initiated the topic, of which Balazinska et al. et. al. are members [BHS11; KUB+12a; KUB+12b; KUB+13] and a second group with Tang as a lead author [Tan14; TSBV13; TWB+13; TASB14]. However, open problems in this area remain [BHK+13].

[1] http://www.bloomberg.com/, accessed: 2015-05-31.

In the reminder of this work, the focus will be on the aspect of value creation, more precisely pricing of data as a specific information good. Consequently, a hybrid approach is used, combining database knowledge with economic theory as pricing is inherently rooted in economic theory and cannot be explored thoroughly without an economic understanding. This section will start with a recapitulation of basic microeconomics in Section 5.1. In the proceeding Section 5.2 data marketplaces will be defined based on economic theory. Extending the theoretical discussion, data marketplaces will then be examined from a practical point of view in Section 5.3. Finally, pricing of information goods is discussed in Section 5.4.

5.1. Basic Microeconomics

In order to lay the foundation for the remainder of this chapter, some basic microeconomic terms and concepts of neoclassic economic theory need to be introduced. This section focuses on concepts relevant to this work; readers interested in further information are referred to [SN10a; SMS11; MT12; PR13], which also build the basis of this section.

A basic concept in economics is the supply and demand of goods [PR13]. It is supposed that consumers (demanders) act in such a way that they maximise the utility they receive from goods they consume. Analogously, it is supposed that producers (suppliers) act in a way to maximise their profits. Both parties do so only in their respective interest [SMS11].

Every consumer has individual preferences that describe the individual utility values they derive from the consumption of different goods. This results in an individual demand curve, showing how much of a given good is consumed at a given price. An overall demand curve is achieved by aggregating all individual consumer demand curves within the economic zone under investigation [SMS11]. For most goods, the demand decreases when prices increase and vice versa.

Suppliers use productive factors (e. g., machines, resources or human labour) to produce goods in such a way that they maximise their profits. Usually, suppliers have fixed costs, i. e., costs that do not change depending on the output, e. g., rent for office buildings, and variable costs, i. e., costs that depend on the output such as resources used for production. As a result, average costs per output unit usually decrease with increasing output, i. e., the fixed costs are covered by more units. However, at some point this typically changes as the variable costs increase again, for example when processes become inefficient owing to spatial constraints [SMS11].

Demand and supply meet on markets which will later be extensively discussed in Section 5.2.1. For the purpose of this introduction, perfect markets are supposed, which have a number of properties:

1. Comparable and homogeneous goods providing a similar utility [RN02; PR13]
2. A sufficiently large number of demanders and suppliers to allow for perfect competition, which implies that either have about the same market share and, thus, cannot set prices but have to accept the market price [RN02; PR13]
3. No market entry or exit limitations [PR13; NDH02]
4. Complete market transparency (i. e., perfect information regarding availability, quality, and prices of goods) [RN02; NDH02; PR13]
5. Immediate adoption to changes in the former [RN02; NDH02]
6. No personal preferences regarding certain suppliers or time of purchase [RN02; NDH02]

It has been established that suppliers try to maximise their profit which equals the difference of revenue and costs: $P = R - C$. Costs and revenue both depend on the amount x of sold units. Furthermore, revenue depends on the price p that can be realised. In this basic concept, as pointed out above, it is supposed that producers cannot set prices but have to deal with a given market price. Therefore, the profit equation can be formulated as: $P = px - C(x)$ [SMS11]. In order to maximise the function $P(x)$ the first derivative $P'(x)$ has to equal zero:

$$P'(x) = p - C'(x) \tag{5.1}$$

$$P'(x) \overset{!}{=} 0 \tag{5.2}$$

$$0 = p - C'(x) \tag{5.3}$$

$$p = C'(x) \tag{5.4}$$

As shown by Equation 5.4, the profit is maximised if the price equals the *marginal cost* (the cost of producing an additional unit). However, in order to be a maximum, $P''(x) < 0$ must hold true, which depends on the exact form of

$C(x)$ because $P''(x) = -C''(x)$. It therefore follows that production increases if p increases as it is profitable to produce at higher marginal cost. This relation is described by the individual supply curve. Similar to demand, the overall supply curve is constructed by aggregating all individual supply curves [SMS11].

As mentioned above, supply and demand meet on markets. In a theoretical framework, this is achieved by equating the supply curve and the demand curve. This means, eventually, the market will arrive at a market price or equilibrium price p^* [SN10b; Sta51; PR13]. The process of arriving at p^* can be made clear by supposing there were a $p_1 > p*$, which would imply an excess of supply. Thus, providers have to decrease prices in order to sell their goods until they reach p^*. A decrease in prices inevitability leads to reduced production and the excess of supply is reduced. The same argument can be made vice versa, i. e., in light of excess supply production is reduced and the price adapts accordingly, the result being the same. Supposing there were a $p_2 < p^*$, which would imply an excess of demand, then producers would increase prices and production until p^* is reached. At the price of p^* all goods are sold, therefore, the equilibrium price is also referred to as *market clearing price* [PR13]. Figure 5.1 shows how p^* can be determined graphically for simple demand and supply curves.

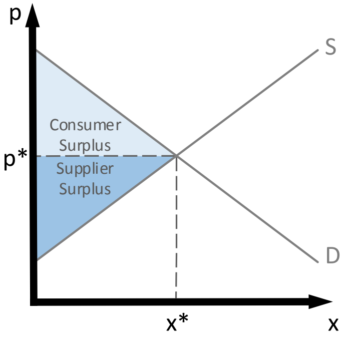


Figure 5.1.: Market Equilibrium, Adapted from [PR13].

Obviously, demanders are not willing to buy a good at any price and have an upper limit for the price they want to pay. This limit is referred to as the *reservation price* [PR13]. If demanders are able to buy a good at a price lower than their reservation price, they gain what is called *consumer surplus* [PR13]. Analogous to consumer surplus, suppliers gain what is referred to as *supplier surplus* if they are able to sell a good for more than their marginal cost of production. The sum of both is called *economic surplus* or *total welfare* [PR13]. This is illustrated in Figure 5.1, too.

In the larger context of data marketplaces, market structures are also relevant. In reality, perfect markets rarely exist, in particular, it is likely that one or a few market participants are larger than others and therefore have more power. In the context of a data marketplace study, Muschalle et al. [MSLV12] discussed the following market structures: *monopoly, oligopoly,* and *strong competition*. Having dealt with strong competition (perfect markets) already, monopolies will be discussed in detail as they are highly relevant in the context of data marketplaces.

In a *monopoly* situation sole suppliers (monopolists) of a good can dictate prices to maximise their profit as they do not face any competition. This is achieved by selling less of a good than under perfect competition at a higher price [PR13; MT12]. Generally, it is supposed that the monopolist knows the demand function $x(p)$. This implies monopolists also know the price they can achieve depending on the amount $p(x)$ [SMS11]. From this, the revenue can be described as $R(x) = xp(x)$. Similar to the case presented above, monopolists want to maximise their profits formulated as: $P = R(x) - C(x)$.

$$P'(x) = R'(x) - C'(x) \tag{5.5}$$

$$P'(x) \overset{!}{=} 0 \tag{5.6}$$

$$0 = R'(x) - C'(x) \tag{5.7}$$

$$R'(x) = C'(x) \tag{5.8}$$

As evident from Equation 5.8, the profit of a monopolist is maximised if the marginal revenue (the revenue generated by selling an additional unit) equals the marginal cost. Furthermore, in order to be a maximum indeed, $P''(x) < 0$ must hold, which depends on the exact forms of $R(x)$ and $C(x)$. Commonly, the resulting price of a monopolist supplier p^*_M is greater than the optimal price under strong competition p^* [SMS11]. This results in an increased producer surplus at the expense of consumer surplus. However, the total welfare also diminishes. All this is presented in Figure 5.2.

As evident from Figure 5.2, it would be ideal for monopolists if they could ask exactly the reservation price from each individual customer as this would maximise their surplus and not lead to a loss in welfare. Asking different prices of different customers is referred to as price discrimination and will be extensively discussed in Section 5.4.1. In contrast, customers do not want to reveal their true reservation price to avoid being exploited. Thus, complete price discrimination does not happen in practice [PR13].

When more than one provider dominates a market, this is termed an *oligopoly*, i. e., the market is dominated by a few [PR13]. The behaviour of these oligopolistic markets is hard to predict; effects can range from price fights to pooling of interests. To analyse this scenario a thorough understanding of the specific industry is necessary. Muschalle et al. [MSLV12] suggest that game theory analysis is a means of forecasting the behaviour of the various parties involved. A special form of an oligopoly is a *duopoly*, i. e., two equally strong, dominant suppliers.

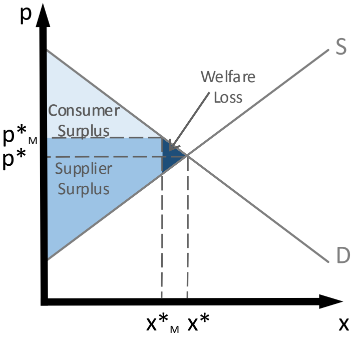


Figure 5.2.: Market in a Monopoly, Adapted from [PR13].

Finally, for completeness' sake, *monopsony* and *oligopsony* shall be briefly explained. As their names suggest they are similar to *monopoly* and *oligopoly* structures but applied to the demand side. The difference is that in a monopoly (oligopoly) a sole provider (a limited number of) faces a large number of customers, while in a *monopsony* (*oligopsony*) a single (a limited number of) demander(s) face a large number of suppliers. Thus, the same implications apply vice versa [PR13]. However, as of today, the market for data is supplier-driven [Mil12c] and no strong demanders have emerged. Thus, these structures are of less importance.

5.2. A Theoretical Perspective

This section firstly discusses markets and marketplaces from a general economic perspective, expanding on the above. Then, this will be transferred to electronic markets and marketplaces. Subsequently, data as a digital good will be established and its marketability will be discussed. These first three parts are loosely based on [Vom14; VSSV15]. Last in this theory section, **the marketability** of digital goods will be discussed.

5.2.1. Markets and Marketplaces

In everyday language, the terms market and marketplace are commonly used as synonyms without taking into account their differences. However, in order to understand data marketplaces, it is important to define the terms for the purpose of establishing a common understanding. In economic theory, as evident from the previous section, *markets* are an abstract construct where actors (first and foremost suppliers and demanders of a given good) meet and exchange a good and determine the price of a good [SN10b]. This implies that a market commonly focuses on one product [BH99]. In contrast, the term *marketplace* for a given good describes the actual, physical or virtual place where the good is traded, i. e., it provides the infrastructure for trades [Gri03]. This means, the difference between a market and a marketplace can be attributed to the level of abstraction. Marketplaces are the infrastructure that enables the abstract concept of markets. Indeed, the sum of all market-based transactions, e. g., selling and buying a specific good in a specific region constitute a market [Sta51; PR13]. For instance, one could speak of the book market in Germany, which is constituted by all book transactions in the country through various channels, such as online and offline marketplaces.

A market as such serves three functions [Vom14; VSSV15; Bak98]. From an organisational perspective, the functions *Institution* and *Transaction* are important [Sch00; Bak98]. From an economics perspective, *Matching Buyers and Sellers* including the pricing mechanism [Bak98] is of utmost importance as it coordinates market participants [SN10b] and – as shown in Section 5.1 – determines the amount of a good produced.

The *Institution* function describes the rules that underlie the market, the behaviour of market participants – such as laws and established communication channels [Sch00]. The second function, *Transaction*, is the actual exchange of goods which can be subdivided into four phases [RN02]:

1. Information phase, in which information on the good to be purchased is acquired/provided

2. Negotiation phase, in which the contractual terms (including the price) are agreed upon

3. Transaction phase, in which goods are exchanged

4. After-sales phase, in which possibly additional services are performed

Regarding matching, it can be said that prices coordinate the actions of buyers and sellers [SN10b; Sta51; PR13]. This is because a price serves as an indicator to other market participants. For instance, a low price could encourage consumers to purchase more of a good, while a higher price motivates producers to increase production [SN10b; Sta51; PR13].

Under *strong competition*, i. e., a market close to perfection as described above, the market price will approach the marginal cost of production. This means, suppliers are no longer capable of setting a profit-maximising price. As a consequence, providers may sell their goods at the market price or not sell at all. Eventually, this is desirable as it maximises the overall surplus [PR13].

5.2.2. Electronic Markets and Marketplaces

The concept of markets and marketplaces can be transferred to the digital world, where it is referred to as electronic markets and electronic marketplaces. Even more than with their offline counterparts, the terms are used inconsistently and sometimes even synonymously [BH99]. However, here the differentiation will be treated analogously to the above; electronic markets can be considered the abstract concept that comprises all rules and transactions and pricing mechanisms that build the electronic market for a good [Sch00]. Most importantly, for a market to be considered an electronic market, the negotiation phase, at least, has to be electronically supported [Sch00]. Likewise, an electronic marketplace is an online infrastructure through which market participants interact [RN02].

In summary, an electronic market comprises all electronic market-based activities. As a result, all electronic marketplaces are part of the electronic market. The electronic market in turn is part of the overall market [BH99]. Thus, the overall market is instantiated by electronic and offline marketplaces. At this point, it should be further clarified that speaking of electronic markets or marketplaces does not imply that digital goods are traded. Indeed, on electronic marketplaces both, physical as well as digital goods, can be traded.

However, there is one major difference between markets and their electronic equivalent. The usage of IT decreases transaction costs significantly as it becomes easier to find relevant information and, therefore, markets are brought closer to perfection [Bak97], which affects pricing. Whilst on imperfect markets providers may have the ability to set prices, this ability can be reduced by electronic marketplaces.

Given that the terms electronic market and electronic marketplace are confused even more frequently than their non-electronic counterparts, Wang and Archer [WA07] have analysed a number of definitions and identified two concepts that exist in parallel. According to their classification, an electronic marketplace can either take the form of a *Governance Structure* or of a *Business Model*.

While the governance structure definition of an electronic marketplace is basically equivalent to the electronic market definition given above, i. e., a market in the abstract sense, the business model definition reflects a marketplace as a concrete institution. In this definition a marketplace is understood to be a virtual place that brings together supply and demand. However, any organisational form falls into this category irrespectively of who drives the platform suppliers, demanders, or an independent third-party.

As a result of the various forms that a data marketplace as a business can take on, several attempts to classify data marketplaces have been undertaken. Vomfell [Vom14] provided a unified classification framework based on [WA07; RN02; Luo12]. The framework is organised along the three dimensions *Orientation, Type,* and *Ownership*, and categorises, different *Business Models* along these dimensions.

Orientation operates on the scale of hierarchy to market. While market forces are allowed to operate freely on the market-end of the spectrum, in a hierarchy model, they are skewed towards either suppliers or demanders (cf. the argument of oligopoly versus oligopsony in Section 5.1). More concretely, this dimension shows whether the electronic marketplace is run in someone's interest.

Next, the *Type* dimension differentiates between vendor-based and marketplace-based electronic marketplaces. While marketplaces as platforms are inherently unbiased, marketplaces driven by vendors (or buyers) are likely to be biased in their respective favour.

Finally, marketplaces are categorised based on their *Ownership*, which can be a) private, i. e., owned by a single company (seller or buyer); b) consortia-based, i. e., owned by a small number of companies (sellers or buyers); and c) independent, i. e., the marketplace is run as platform without any connection to sellers or buyers. Since *Ownership* and *Type* are functionally dependent, here, this classification has been simplified and *Type* has been omitted. This insight is also reflect in [VSSV15].

Vomfell [Vom14] identified six types of relevant electronic marketplace *Business Models* depicted in Figure 5.3. Firstly, there are highly biased electronic marketplaces which are privately owned hierarchies and, thus, favouring an organisation. As such they appear both as seller-driven as well as vendor-driven.

Next, there are consortia-based marketplaces that serve a particular organisation or a number of organisations. However, they are commonly less hierarchic because they allow for multiple vendors or buyers to participate. Finally, there are true marketplaces which are independent and allow for the market forces to flow comparatively freely. However, there is one restriction. In order to be considered a true marketplace, the operator must not sell data they own, as this might bias their behaviour. Therefore, this type of marketplace has been classified as a consortium because of a similar bias.

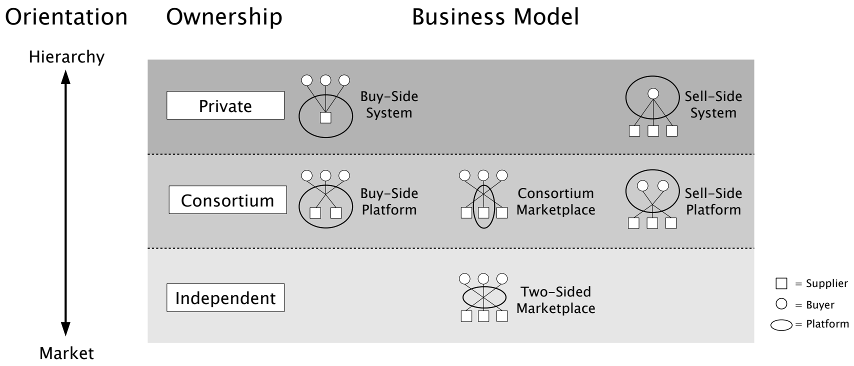


Figure 5.3.: Classification of Electronic Marketplaces (simplified from [Vom14], published in [VSSV15]).

5.2.3. The Marketability of Data

For non-digital goods, the mechanisms of markets do not change tremendously because they are traded electronically. However, it is generally acknowledged that electronic marketplaces reduce the transaction costs. Digital goods are inherently different; in order to understand the market for data and the according data marketplaces, one has to look into the particulars of data as an information good.

As should be evident from Chapter 3, data is the basis of information, which requires context. Therefore, most of the time, data is actually traded when speaking of information goods [Lin09]. In order to avoid confusion, the term information good will be used here because most of the literature refers to digital goods as information goods, e. g., [Lin09; SV99; WHCA08; BC08; BS03; Cho10; CY07]. The term information good has been defined by Shapiro and Varian [SV99] as "everything that can be digitalized". This is a very broad but fitting definition and will be applied here.

While it has been discussed for some time whether information goods can indeed be considered economic goods, it is now widely accepted that they are because they fulfil the basic requirements of a good: transferability, utility, and the existence of a demand [Jür97]. Nevertheless, they have been described as a *clearly peculiar* economic good [Bat90].

One peculiarity of information goods is their very special cost structure. They are initially very costly to produce and require high upfront investments. However, once produced, information goods are cheap to reproduce, i. e., the fixed costs are very high, while marginal cost is close to zero [SV99; SF08; HHS06]. This implies high economies of scale [SF08]. Furthermore, the low marginal cost and special cost structure also allow for information-based services to compete with traditional services, for instance, Skype[2] can be seen as competitor for traditional telephony [SF08]. Similarly, the film streaming provider Netflix[3], who proposed a partnership with the movie rental chain BLOCKBUSTER[4] in 2000, is now a successful company, while BLOCKBUSTER went bankrupt in 2010 [Sat14].

However, the distinctive cost structure also leads to some difficulties. Unusually, the up-front investments are generally sunk cost [SV99], i. e., cost that cannot be recovered through stopping production (e. g., computation time and labour cannot be undone). This is in contrast to traditional industries, where upfront investments such as machinery can be monetised once production is stopped. This cost structure makes information goods special in that they resemble public goods [SF08; Lin09]. Public goods have two distinct properties. They are non-excludable, i. e., it is not possible to exclude someone from consumption who has not paid for it. Furthermore, they are non-rivalrous in consumption, i. e., the utilisation of the good by one consumer does not hinder the usage of another consumer [PR13]. Private goods, in contrast, are both excludable and rivalrous. If only one of the two criteria is satisfied, one speaks of natural monopolies (only excludability), for example, toll roads, or common resources (only rivalness), such as the environment [MT12; LS11].

[2] http://www.skype.com/, accessed: 2015-05-31.

[3] https://www.netflix.com/, accessed: 2015-05-31.

[4] http://www.blockbuster.com/, accessed: 2015-05-31.

At first sight, it may be hard to exclude people from information available online. However, it is possible to exclude people from using information goods by employing technical and legal means. While this does not guarantee excludability per se, it allows for a legal enforcement of excludability. The application of the rivalness criterion is difficult because information goods can be copied at no cost and do not wear, even after usage. This implies that the usage of an information good by one consumer does not prevent another consumer from using it, too [LS11]. Thus, Linde and Stock [LS11] have suggested substituting rivalness with network effects for an information good classification depending on whether it is positive if the information is distributed or not. This results in four types of information goods [LS11]:

- **Private Information** excludable and negative network effects, e. g., trade secrets
- **System Information** non-excludable and negative network effects, e. g., insider information
- **Market Information** excludable and positive network effects, e. g., Encrypted Pay TV
- **Public Information** non-excludable and positive network effects, e. g., information on Wikipedia

That being said, it is rather a matter of personal taste whether one speaks of network effect or rivalness because the argument can be made that negative network effects lead to rivalness, for example, one company may not want another company to have the very same information good. Considering the value of business information, the point can be made that some information goods are indeed rivalrous as they potentially provide a competitive advantage, e. g., exclusively knowing about an oil deposit on a premise before buying it can be such an advantage.

Generally, it can be argued that the value of information is hard to estimate as it is very different to different people [SV99; SF08]. Bates [Bat90] states that the value of information is probabilistic in that it depends on the returns of its (future) use. He introduces the term stock value to refer to the value of a piece of information at a single point in time. As the stock value potentially decreases with every sold unit of information, he compares the decrease in stock value to marginal cost. If the good underlies network effects (such as communication software which is more valuable if more people use it) the stock value increases with its distribution [Bat90]. Having established the stock value, it can be argued that the consumption of information goods is indeed rivalrous whenever the stock value decreases with an increase in distribution of the information good.

Considering data, all four cases are possible. It can be established that it is an information good that can be made excludable by implementing appropriate technical and legal means. Thus, data can fall into any of the four categories as the following examples show. All data that eventually influences business decisions, such as market research data, can be considered *Private Information* as it can be a competitive advantage if nobody else has access to it. This of course is only true if means to exclude others from using this information are in place. For instance, stock market information can be a private information good. However, if the very same information is freely available, it counts as *System Information*. An example of data made available to the general *public*, which also is non-rivalrous, is Wolfram|Alpha. Again, if this data were protected by technical means it could be considered *Market Information*. Admittedly, there is a pro version of Wolfram|Alpha; however, it provides more comprehensive features rather than more information.

While for traditional goods the general assumption is that more of a good is better, sometimes less information can be more valuable than abundant information [SV99]. As an example, a condensed report can be mentioned, which would otherwise only be accessible through a number of database tables with data that has not been aggregated. Similarly, the value lies less in the information, which is omnipresent on the Web, but in the way it is presented through curation [SV99]

Another major challenge in dealing with information goods is that they are experience goods [SV99; SF08]. This results in a paradox identified by Arrow [Arr62], who states the value of information can only be judged by consuming (experiencing) it. If a provider of data hands it to consumers in order to evaluate the value it has to them, then they have, in fact, acquired it free of charge. This is why usually only samples are given out for evaluation. However, these samples do not provide full insights and may therefore not be suitable for judging the overall relevance [HC02]. Furthermore, it can be said that buyers commonly trust information goods to be of value if they have proven to be so in the past [SV99].

Regarding marketability, i. e., through which structures information are sold, it can be stated that products that are well standardised and have a low complexity are ideal goods to be sold through markets [MYB87]. The process of standardising products is also referred to as commoditisation – the process of becoming a commodity. At this point, price becomes the only differentiating factor [Lan12].

This phenomenon occurs in stable industries when homogeneous products are offered to customers, who are price-sensitive and have relatively low switching-cost [RST10].

In the technical domain it is argued that data will become, or in fact already is, a commodity [TWB+13; Mil12b; Koled; Haq08; LK14], which from an economic point of view is not preferable for data vendors. Given the low marginal cost of information goods, commoditisation is likely to lead to a fierce price competition leaving little space for profits as prices converge to marginal cost, i. e., zero. This can be illustrated by an example from Shapiro and Varian [SV99]. They showed that in a market of digital telephone directories prices decrease to zero because all providers must lower their prices below their competitors' prices in order to gain a market share. As long as marginal cost is not reached, every copy sold contributes to the revenue. Nevertheless, with marginal cost being practically zero, the price will eventually be zero. This is problematic as fixed costs have to be covered, too. Decreasing prices inevitably lead to a decrease in profit and may eventually lead to a point where fixed costs cannot be covered any longer.

However, it is rather unlikely that data will soon become a true commodity as hardly any movement in this direction can be observed [Vom14; VSSV]. In contrast, it has been argued that information good providers can be viewed as monopolists in their domain as they offer unique products [FOS97; WB10]. This is in line with Hui and Chau [HC02] who argue that monopolistic situations occur because of differentiation of information goods. This phenomenon is also known as monopolistic competition and was first described by Chamberlin [Cha33] (as stated in Stahl [Sta05]). It is owing to the fact that in reality perfectly homogeneous goods are seldom observed. Most of the time similar, but not identical goods, are sold, i. e., goods that serve the same purpose and belong to roughly the same price category [Sta05]. In such a situation, participants can still easily enter and leave the market which leads to competition. However, given that goods are not as homogeneous as necessary for perfect markets (see Section 5.1) and also that people do in fact have personal preferences regarding certain suppliers or time of purchase, providers have the ability to set the price within a given price interval [Gut84]. One simple fact of differentiation may be branding [PR13]. For instance, Dr.Oetker[5] are able to sell their baking soda at a premium price on the German market even though the actual ingredients are not significantly different from other providers of baking soda. In context of information goods, it can be stated that while there are plenty of

[5] http://www.oetker.de/, accessed: 2015-05-31.

newspapers, there is only one New York Times[6] [FOS97]. Consequently, the New York Times can act as monopolist in some limits if customers are reluctant to switch for whatever reasons. However, this monopoly is not perfect as readers may change to another newspaper.

5.3. A Practical Perspective

Based on the observation that the analysis of freely available data, together with commercial data and in-house data on centralised cloud infrastructures, is an increasing market segment, a series of overall 12 interviews with data experts has been conducted by the author and other, published in [MSLV12; SLV15]. The following section is largely based on these interviews. Methodologically, the authors followed a seven-step approach to semi-structured interviews, as suggested by Kvale and Brinkmann [KB08], covering the topics experience of interviewees, data-related products, and business models as well as the question of what an ideal marketplace for data would look like. Interview partners were top-level managers and experts from European and U. S. companies providing data marketplaces or data-related services. The experts had different backgrounds, of which data marketplace operation, social media monitoring, text enrichment, and consultancy were most important. The studies sought to answer the following questions:

1. What are common queries and demands of participants on a data market?
2. Who are participants and beneficiaries of data marketplaces?
3. Which pricing models use beneficiaries for data and data-related products?
4. Which technical challenges arise from the combination of data and data-related services on data marketplaces?

While the technical details (question four) are less relevant in the context of this work, the other questions are and will be addressed here. Regarding common queries, Muschalle et al. [MSLV12] identified two major scenarios. The first scenario is estimating the value of a *thing*. In this scenario, a data marketplace is merely a data warehouse that is used to aggregate and evaluate facts about *things* gathered from various sources including in-house private data. As an example the authors mention estimating the value of a publication outlet by

[6] http://www.nytimes.com/, accessed: 2015-05-31.

asking the following question: "*What are top-10 journals for 'my' life science publications?*" [MSLV12]. In this scenario, the indicators that show what the value of a *thing* is, are stored in the data warehouse and must not be confused with the value the answer of the query provides to the querying party.

The second scenario is retrieving all known facts about a *thing*. In this case, too, the data market resembles a data warehouse. In contrast to the first case, it is not the value of a *thing* (alone) that is stored but rather factual information about a plethora of *things* integrated into a universal relation. While this problem has been known for decades in various computer science disciplines, such as information integration [BN08], text mining [DRV06], and information retrieval [Mar06; LAF12], it is still considered to be a big technical challenge to resolve and reconcile logical objects across a set of heterogeneous sources.

While both application scenarios are technically challenging and at first seemingly separated, both use cases address a scenario in which facts – be it value or other facts – about *things* are stored and made accessible through a central infrastructure. Accordingly, data marketplaces can be seen as virtual equivalents to marketplaces for goods and commodities on which the aforementioned facts or collocations of facts can be purchased and sold. Additionally, data marketplaces do not only provide customers with data but also with data-related services such as cleansing, integration, and analysis. Moreover, these marketplaces act as integration platforms for data from various sources and as such also as single point of access. As a result, data marketplaces enable new business models providing information and analysis tools as goods and services. Muschalle et al. [MSLV12] identified finance, healthcare, and business intelligence as major areas of application.

It can be stated that these goods and services are very costly to produce as they require an enormous computing infrastructure, for instance, when analysing large amounts of Web data. On data marketplaces collection, storage, and analysis of data are provided through a central platform and running costs are covered by a number of users, therefore, data marketplaces can be particularly beneficial for Small and Medium-sized Enterprises (SMEs). As a result, individual companies in general and more specifically SMEs do not have to independently carry the cost of implementing and running such an expensive infrastructure [MSLV12; SLV15].

The remaining two questions, namely the organisation of and pricing on data marketplaces, will be addressed in the following two subsections. Thereafter, an empirical overview of the market for data will be given. This section is then concluded by highlighting important related work on specific data marketplaces as well as on surveys of the market for data.

5.3.1. Infrastructure of and Actors on Data Marketplaces

In 2012, Dumbill [Dum12] identified three major characteristics of data marketplaces: a) they allow for comparison of data on offer regarding scope and quality, b) they cleanse, prepare, and integrate data to make it ready for use, and c) they allow for broad access of data. This could be refined by a series of interview studies [MSLV12; SLV15] which identified a number of high-level architectural layers of a data marketplace, illustrated in Figure 5.4 and described thereafter.

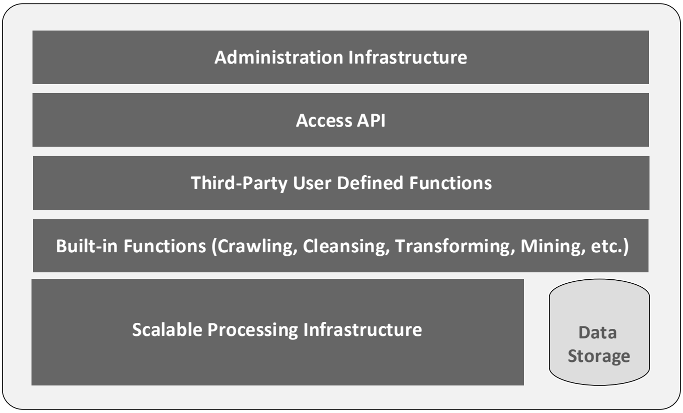

Administration Infrastructure
Access API
Third-Party User Defined Functions
Built-in Functions (Crawling, Cleansing, Transforming, Mining, etc.)
Scalable Processing Infrastructure
Data Storage

Figure 5.4.: Architecture of a Data Marketplace (condensed from [MSLV12; SLV15]).

Scalable Processing Infrastructure to process large data quantities such as excerpts of the Web.

Built-in Data Storage to persist the provided data.

Built-in Functions to integrate data from public sources (e. g., the Web), private sources (e. g., stock exchange data), and in-house sources (e. g., ERP Systems[7]) and additional functions to process the data through cleansing, transforming, mining, aggregating, etc.

Third-Party UDFs layer which allows the execution of User Defined Functions (UDFs) that have been developed by third-party developers to complement the built-in functions.

Access APIs provide functionality to offer and access the built-in functionality, data, and UDFs through an Application Programming Interface (API).

Administration Infrastructure which allows for managing all administrative tasks such as providing offers, concluding contracts, billing, etc.

Furthermore, the studies identified seven types of beneficiaries and actors in the broader data marketplaces environment. Understanding their needs is important when building a data marketplace, as it is the declared aim of such a system to solve their problems and provide them with pricing mechanisms that work for them. The actors can be categorised into two groups: *direct* and *indirect actors*. Whereas direct actors make use of data marketplaces directly (by selling, buying, etc.), indirect actors provide supporting services such as consultancy or quality assurance. Figure 5.5 gives a short overview of all actors, before they are described in detail.

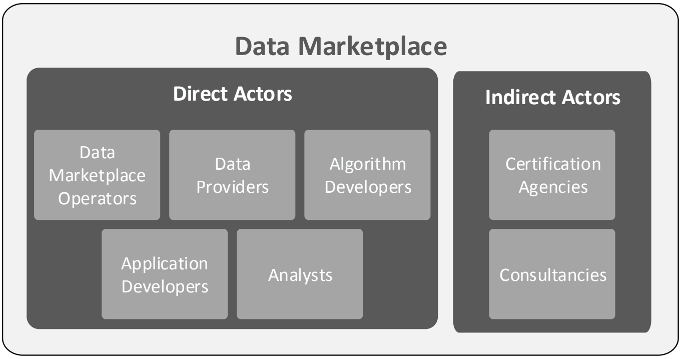


Figure 5.5.: Overview of Actors on a Data.

[7] Enterprise Resource Planning Systems.

Data Marketplace Operators are the economic entities that run data marketplaces. As such, they carry the risks of operations, provide and maintain the hardware infrastructure, and receive usage fees from participants. Data marketplace operators face a variety of challenges. In this work, there will be a focus on the economic challenges but it is acknowledged that there are many more including legal, ethical, and technical challenges.

Data Providers can be differentiated in commercial and non-commercial data providers. The first group consists of established data and information providers, THOMPSON REUTERS[8] or BLOOMBERG, for instance, and also search engine operators such as GOOGLE or BING. Besides these well-known data providers, operators of online forums or providers of linked data who seek monetisation are also considered to be commercial data providers. Non-commercial data vendors are mainly governmental organisations which want to provide their data for the greater good or are required to do so by law. Data providers use data marketplaces mainly to store their data and to make it available to a greater community, i. e., achieve visibility. Thus, data providers deliver data to the data marketplace and receive monetary compensation in return if they charge for their data.

Algorithm Developers provide data marketplaces with UDFs, which provide an additional value to the data marketplace. Typical application areas of UDFs are data mining, cleansing, relevance computation, and provenance. Most commonly, these UDFs are focused on a specific source, language, or industry. Data marketplaces provide algorithm developers with the functionality to offer their UDFs as black-box-functions which can then be used by other actors for a fee. In turn, algorithm developers receive a share of the turnover generated by their algorithms.

Data Analysts are typically domain experts who want to use the plenitude of data available on data marketplaces for their own analyses and reports. They hope that the quantity and quality of the available data will improve their analysis results, which will eventually give them a competitive advantage. Commonly, this group runs structured ad-hoc queries against the data marketplace. By doing this, they structure their (often domain-specific) knowledge about data integration which they may also provide as UDFs. In this case, they take on a second roll as algorithm developer.

[8] http://thomsonreuters.com/, accessed: 2015-05-31.

The data marketplace operator benefits because data analysts pay fees for retrieving data.

Application Developers use data from data marketplaces to feed their applications with data. Often, analysts are not able to express their information needs in a formal or technical way. Thus, application developers provide them with applications that satisfy their information needs without having to formulate structured queries. To this end, application developers translate the information need into pre-compiled queries which are then run, whenever less technical analysts use their application. In essence, application developers run pre-compiled queries on data marketplaces and pay a fee for the usage of the data.

Certification Agencies certify the security of a data marketplace, or the quality of data or UDFs. These certificates help platform operators to signal trustworthiness to their customers. As a consequence, the data marketplace may gain market shares. This signalling is important because of the information paradox discussed in Section 5.2.3, i. e., that it is hard to try data in advance of a purchase. As this actor does not directly trade data or data-related services, it has been classified as an indirect actor within this work.

Consultancies support users of data marketplaces who themselves are not capable of unleashing the full potential of data marketplaces. Common tasks include support in choosing sources, algorithms to apply, or developing products based on available data. In a way, data marketplace operators generate customers for consultancies. However, consultancies in reverse provide customers with support that does not have to be delivered by the data marketplace operator and might raise awareness for the fact that data marketplaces exist. Again, consultancies have been classified as indirect actors because data marketplaces are commonly not their main focus.

Figure 5.6 shows a simplified schema of interaction between all direct actors on a data marketplace, where unlabelled arrows represent data flows; in exchange for data, money is transferred. This is done in such a way that money is always sent and received by the data marketplace operator, which is sensible as they are an intermediary which participates on all transactions by means of revenue sharing. Furthermore, data and algorithms interact in such a way that algorithms use data to create new data. Obviously, individuals can take on multiple roles on the market, e. g., an analyst who is also providing data.

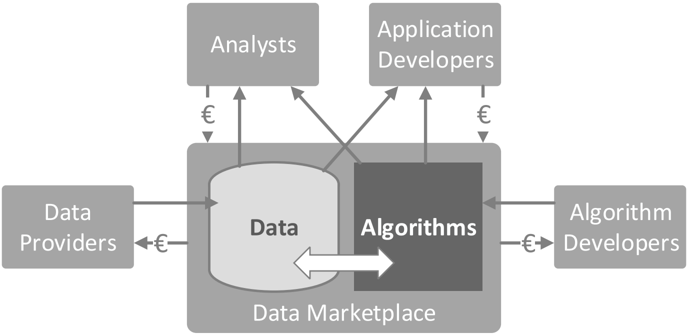


Figure 5.6.: Actors on a Data Marketplace (inspired by [MSLV12]).

5.3.2. Observed Pricing Mechanisms on Data Marketplaces

In this section, common pricing mechanisms used by data marketplaces, as well as factors influencing them, will be presented based on the interview studies [MSLV12; SLV15]. The authors stated that pricing models can largely be categorised as either atomic or hybrid models, the latter being combinations of the first. The following six atomic models have been found: *Free, Trade, Pay-per-Use, Flat Rate, Tiered Pricing,* and *Progressive Pricing.* It will be discussed in due course how (if at all) they allow for discrimination of customers through mechanisms of personalised pricing, versioning, or group pricing which is an extension to [MSLV12; SLV15].

Free data is commonly provided by authorities, governmental organisations, or NGOs. While this data may not be monetised on a data marketplace, it may still help a data marketplace to gain customers, which in turn attracts data vendors. Furthermore, free data can be used for data mining or be integrated with other commercial data to result in new insights and potentially new valuable data. Since the data is provided for free, no price discrimination applies. However, in the context of a data marketplace it may well happen that data providers such as NGOs want to publish their data but lack the infrastructure to do so. In this case, they might be willing to or indeed have to (if they must publish) pay someone to publish their data.

Trade as a business model is based on the observation that sometimes data providers allow a third-party the usage of their data for free if they receive additional, value-added data in return. Given that this type of trade is usually formalised through individual contracts, this can be considered personalised pricing.

Pay-per-Use charges for the actual consumed unit of data or service. However, so far it was also stated that this pricing model was only observed in the area of consultancy but never in the area of actual data marketplaces. If it is applied in its purest form, pay-per-use does not discriminate customers. However, if providers offer a discount when larger quantities are consumed, they essentially offer different versions (different average prices) of the product and thus discriminate customers. That said, such an offering actually counts as *Tiered Pricing*.

Flat Rate models – sometimes referred to as *subscription* models – are based on the usage time only. They allow a certain time of usage of data or services indifferent of the actual usage intensity. This pricing model is common for software licenses as well as for news agencies. A special case of this is one-off purchases of data for which a fixed fee is paid for unlimited usage. This special case, however, does not necessarily imply updates to the sold information good. If a continuous service is considered, it can be argued that users select their own average price based on their intensity of usage [SSW05]. Thus, this model can also be considered a form of second order price discrimination.

Tiered Pricing consists of different pricing tiers or packages. Basically, these contain the same service in different quantities such as the number of API calls or allowed data usage. Furthermore, higher tiers may include more comprehensive services, e. g., additional quality checks. In contrast to the flat rate model, this model is influenced not only by time but also by the amount of usage. As a result, this price model is an implementation of either second or third degree price discrimination depending on the exact configuration. If the tiers explicitly target different audiences such as universities, businesses, or NGOs, but provide essentially the same product, it can be considered group pricing, i. e., third order price discrimination. In contrast, if tiers offer the same product in different versions (quantities or qualities), it is a case of second degree price discrimination or versioning. Moreover, a combination of the two is also possible.

Progressive Pricing refers to a model in which the price depends on the time of purchase; it increases the later a good is purchased. This model is applied if the distribution of data is to be limited, for instance on stock photo websites. This model not only limits the distribution of a good, providing exclusivity, but it also allows for price discrimination based on exclusivity and personal appeal. More concretely and staying in the context of digital images, if a customer desperately wants an image which has been sold some times already they have to pay a high price. However, if they search for a particular type of image but do not mind the specific representation, they might as well purchase a cheaper image that has not been highly distributed yet. Overall this can be viewed as a form of versioning, the versioning criterion being exclusivity.

Regarding hybrid models, *Two-Part Tariff* and *Freemium* could be identified as being currently employed. The first combines a basic fee (that may be progressive) with a usagedependent model such as *pay-per-use* or *tiered pricing*. In this model the fixed component serves as a contribution to cover fixed costs and the variable part to discriminate customers by their willingness to pay. As with *tiered pricing*, depending on the actual arrangement, this may either be a form of versioning or group pricing. Finally, *Freemium* – a portmanteau combining free and premium – offers basic goods or services for free and charges for additional services. The charged part can take on any of the pricing models discussed above. As a consequence, all that has been stated with regard to price discrimination above also applies here. Additionally, the overall arrangement can be viewed as versioning in itself, providing a free and premium version, potentially even more than one for the latter. Simon and Fassnacht [SF08] encourage this strategy to penetrate a market in particular for goods with strong network effects.

Besides describing these observed pricing models, Stahl et al. [SLV15] discovered three major categories of factors influencing the price for data and data-related services: a) *economic*, b) *licensing*, and c) *data-related* factors.

Economic factors are composed of factors such as the *type of market, payers* – as evident from the reviewed business models, these do not have to be identical with the one acquiring the data –, the *exclusivity* of the data, and the *subjective* value. The licensing-related factors are to some degree a sub-group of the former group, as they are negotiated by two business parties concluding a contract. To be more precise, these licensing factors usually restrict the usage of data in respect to *time* or *re-selling*. Data-inherent factors include *type of data, amount of data, timeliness*, and probably most important, the *quality* of data. Table 5.1 provides an overview of the identified factors.

Based on the criteria presented in Table 5.1, Figure 5.7 presents a decision support flow chart for choosing an appropriate pricing model. If users are not willing to pay for the data but a third party wants it to be published, it should be offered for free and the third-party should be charged. If no-one wants to pay, *free* is the only option. However, it is possible that this will attract more customers. Obviously, trade makes only sense if the trade partner has appropriate data to offer.

If customers want fair pricing and are not overly interested in planning security, *pay-per-use* is a good choice. *Tiered pricing* is sensible if the market can be discriminated by tiers. Alternatively, if exclusivity is important, *progressive pricing* might make sense. Else, only a *subscription* model remains.

Regarding the more complex pricing models, it has to be gauged whether potential customers are willing to pay a basic fee, in which case a *two-part tariff* can be considered. Also, if it is deemed economical a *freemium* model is an option to attract customers. Although not explicitly addressed in [SLV15], Figure 5.7 essentially presents a guide on how to discriminate customers based on their preferences.

Table 5.1.: Factors Influencing the Price of Data and Related Services based on [SLV15].

	Factor	Manifestation	Influence
Economic	Payer	Demander, Supplier	Demanders usually have a higher willingness to pay.
	Market Form	Polypoly, Oligopoly, Monopoly	Monopolists can usually achieve higher prices.
	Subjective Value	— — —	The higher the perceived value, the higher the achievable price.
	Exclusivity	— — —	The more exclusive the data, the higher the achievable price.
Licensing	Type of usage	Self, Reselling, Any	Commonly increasing rights increase the achievable price.
	Usage Time	Temporary, Unlimited	The longer the usage right, the higher the achievable price.
Data-Inherent	Timeliness[9]	Static, Up-To-Data	Up-to-date data is commonly more expensive.
	Amount of Data	Complete Data Set, Excerpts	The more comprehensive the data, the higher the achievable price.
	Type of Data[10]	Structured, Unstructured	The more structured, the higher the achievable price.
	Data Quality	— — —	Qualitative data (e. g., complete, accurate) is commonly more expensive.

Based on an interview study with 20 paid content[11] managers, Theysohn et al. [TPS05] claim that price differentiation commonly uses volume discounts and versioning. Furthermore, they state that in Germany two-part tariff (45%), pay-per-use (30%), and flat rate (25%) are used for paid content.

[9] In the original study this was called *Type of data*

[10] In the original study this was called *Form of Data*

[11] Goods that are created, sold, and billed electronically.

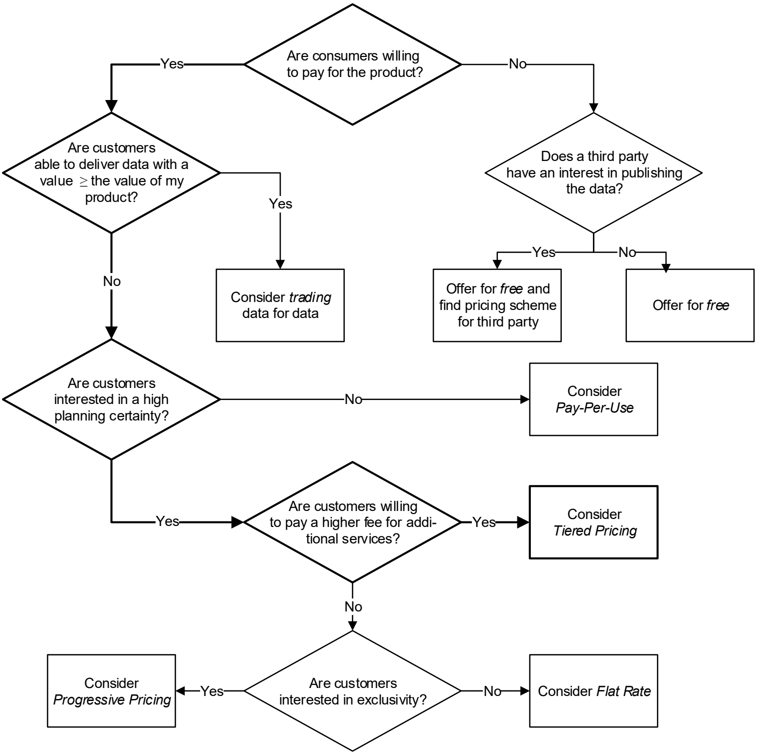

Are consumers willing to pay for the product?
Yes
No
Are customers able to deliver data with a value ≥ the value of my product?
Yes
Consider *trading* data for data
Does a third party have an interest in publishing the data?
Yes
No
Offer for *free* and find pricing scheme for third party
Offer for *free*
No
Are customers interested in a high planning certainty?
No
Consider *Pay-Per-Use*
Yes
Are customers willing to pay a higher fee for additional services?
Yes
Consider *Tiered Pricing*
No
Are customers interested in exclusivity?
Yes
Consider *Progressive Pricing*
No
Consider *Flat Rate*

Figure 5.7.: Decision Support for an Appropriate Pricing Model; Translated from [SLV15].

5.3.3. The Market for Data and Data-Related Services

Between 2012 and 2014 the author and others conducted a series of three Web-based surveys to study an increasing number of data marketplaces and data vendors. This was in order to determine the current state of the market for data and data-related services, which will be briefly summarised here from [SSV13; SSV14a; SSV14b; VSSV; Vom14]. Based on [MSLV12], a number of data marketplaces as well as data providers were derived and extended by means of Web search. This sampling resulted in a total number of 46 companies or organisations providing data or platforms for trading data in [SSV13].

Following an iterative approach, 12 dimensions were developed to categorise data marketplaces and data providers. The dimensions were divided into objective dimensions, which are easily verifiable such as applied *Pricing Model*, and subjective dimensions, which are dependent on the researchers' judgements such as *Trustworthiness*. An overview of all dimensions used is given in Table 5.2, in which each dimension is annotated with a number of daggers to indicate in which of the studies it was first introduced (1 = †, 2= ‡, 3 = ‡‡). Most of the dimensions are not mutually exclusive, i. e., a provider may fall into more than one category. If a dimension consists of mutually exclusive categories, it is annotated with an *m*.

Laying the foundation for the future studies, the first iteration did not state any trends but provided a picture of the English and German data market as of 2012. In 2012, out of overall 46 companies under investigation, 16 classified as raw data vendors and 7 as data marketplaces [SSV13]. Regarding *Pricing Model* the following four pricing models were observed which were about equal with regard to usage (absolute figures in parentheses), flat rate (15), pay-per-use – including tiered pricing – (12), freemium (15) and free (12).

One year later, 42 of the initial 46 companies were still in business and five additional participants were added. Furthermore, the study was extended by two new categories, namely *Pre-Purchase Testability* (as objective category) and *Pre-Purchase Information* (as subjective category) [SSV14a; SSV14b]. These categories are important with regard to the fact that information is an experience good.

Results in the two relevant categories show that raw data vendors have decreased by one and that data marketplaces have increased by three. Regarding pricing models, a clear increase in pay-per-use (+6) could be observed. While free (+2) and flat rate (+3) also increased, freemium decreased slightly (-1).

Overall the second study [SSV14b; SSV14a] concluded that raw data offerings were decreasing in number, favouring high-quality content. This was supported by findings from an interview study [MSLV12] which also predicted that data-enhancing services, such as associated algorithms and data visualizations, were due to be offered. Regarding the new *pre-purchase* dimensions the study concluded that most providers provide enough information in order for customers to take informed decisions on which product to purchase. Overall, the market was not settled in 2013.

The most recent study [VSSV; Vom14], conducted in late 2014, looked at 72 data marketplaces and data providers according to the definition presented in Section 5.2.2, restricted by the following limitations: a) the provided data had to be in a machine-readable format, b) the data had to be hosted by the provider (excluding directories of data), and c) if the tool was an analysis tool, it had to use

Table 5.2.: List of Data Marketplace Dimensions Compiled from [SSV13; SSV14a; SSV14b; VSSV; Vom14].

	Dimension	Categories	Question to be answered
Objective	Type†	Web Crawler, Customizable Crawler, Search Engine, Pure Data Vendor, Complex Data Vendor, Matching Vendor, Enrichment Tagging, Enrichment Sentiment, Enrichment Analysis, Data Market Place	What is the type of the core offering?
	Time Frame†	Static/Factual, Up-to-Date	Is the data static or real-time?
	Domain†	All, Finance/Economy, Bio Medicine, Social Media, Geo Data, Address Data	What is the data about?
	Data Origin†	Internet, Self-Generated, User, Community, Government, Authority	Where does the data come from? Who is the author?
	Pricing Model†	Free, Freemium, Pay-Per-Use, Flat Rate	Is the offer free, pay-per-use or usable with a flat rate?
	Data Access†	API, Download, Specialised Software, Web Interface	What technical means are offered to access the data?
	Data Output†	XML, CSV/XLS, JSON, RDF, Report	In what way is the data formatted for the user?
	Language†	English, German, More	In studies one and two it was looked at what the language of the website and the language of the data were. Study three examined the meta data language, i. e., what language does one need to understand in order to make sense of the data?
	Target Audience†	Business, Customer	Towards whom is the product geared?
	Pre-Purchase Testability^{‡,m}	None, Restricted Access, Complete Access	Do providers give consumers the chance to "try before they buy"?
	Ownership^{‡‡,m}	Private, Consortium, Independent	How independent is the provider?
Subjective	Trustworthiness†	Low, Medium, High	How trustworthy is the vendor? Can the original data source be tracked or verified?
	Size of Vendor^{1,m}	Startup, Medium, Big, Global Player	How big is the vendor?
	Maturity^{1,m}	Research Project, Beta, Medium, High	Is the product still in beta or already established?
	Pre-Purchase Information^{‡,m}	Barely Any, Sparse Media Information, Rich Media Information	Do providers provide enough information to take informed buying decisions?

proprietary data for the analysis. Besides these inclusion criteria, also exclusion criteria were applied, for instance all governmental providers were excluded as they do not focus primarily on business. Furthermore, financial institutions were excluded because of their sheer number. Based on this – in comparison to its predecessor surveys – more profound definition of survey subjects, an extensive keyword Web search was conducted resulting in 72 survey subjects. Regarding dimensions, an ownership dimension was added to investigate potential bias (see Section 5.2.2).

Although the numbers cannot be compared to the data obtained in previous years, because of the significant extension, the two dimensions discussed in the previous studies shall be presented here for the most recent study. In the overall set, raw data vendors were by far the most prominent providers (27), followed by data marketplaces (15), followed by enrichment analysis (9), to name just the first three. Regarding pricing models, free (15), freemium (18), and pay-per-use (23) were far less often used than flat rate (39). The new dimension ownership yielded an impressive result revealing that 54 providers were private companies, whereas only 6 were consortia and only 12 could be considered independent. As overall trends, the 2014 study identified aggregated and matched data, an indicator that simply selling data is less common. The shift in pricing models was somewhat surprising but a possible explanation could be that vendors feel comfortable in their competitive situation to apply pricing models in their favour [Vom14]. Also, the ownership dimension suggests that the market as a whole is still limited and that products are very much differentiated. Muschalle et al. [MSLV12] mention that only some companies are facing a strong competitive environment, whereas others face no (or just a limited number of) directly competing products or services which is evident by the fact that these interviewees were not able to name any competitors. While all interviewees considered competition as an adequate means for innovation as well as for welfare and market size growth, they also considered the market as of 2012 to be big enough that there is no fierce competition. Furthermore, they predicted that data marketplaces will not only provide data, but will also offer data associated services such as enhancing algorithms or data visualisations.

5.3.4. Related Work on Data Marketplaces

Since data marketplaces aim to provide data through computer accessible means, data services – a special form of Web services – can be seen as the basis of data marketplaces. Carey et al. [COP12] provide a general data service architecture, and examine concepts and example products for *service-enabling data stores, in-*

tegrated data services, and *cloud data services.* In particular the last concept is relevant when speaking of data marketplaces as these also build on cloud infrastructures.

GE ET AL. [GRC05] were among the first to study electronic information marketplaces. However, they restricted themselves to question and answer websites (e. g., ASKJEEVES[12]), on which users can ask questions that are then answered by other users or experts. Also, they only investigated five websites.

Research on data markets is still in its infancy – at least as far as overviews are concerned. In 2012, DUMBILL [Dum12] named FACTUAL[13], INFOCHIMPS[14], DATAMARKET[15,16], and MICROSOFT AZURE DATA MARKETPLACE[17] the most mature marketplaces for data.

Out of those, INFOCHIMPS can be considered very close to a pure data marketplace because they focus on the integration of various different sources to provide added value to customers. Similar in their approach are FACTUAL who focus on integration of geographical data [SV13]. MICROSOFT AZURE DATA MARKETPLACE can be seen as addition to MICROSOFT'S[18] cloud services with a very strong selection of data suppliers and a unique combination of application and data [Mic11]. DATAMARKET – who were recently acquired by QLIK[19], a company focusing on data visualisation [Gis14] – targeted their offer on consumers and less data-affine professionals by enhancing their data with additional visualisation.

Also in 2012, MILLER [Mil12a] published an interview series with ten data marketplace providers available as a series of Podcasts. Among others, he discussed DATAMARKET, INFOCHIMPS, KASABI[20], and MICROSOFT AZURE DATA MARKETPLACE. However, these interviews are only available as raw podcasts and no transcripts or similar documentation are publicly available which makes it difficult to analyse the contents. Nevertheless, he also published a very condensed report [Mil12b] of his findings. In this report he states that the purpose of data marketplaces is to provide individuals and companies with data. He adopts the major characteristics of [Dum12]. As there was no commonly accepted list of features a data marketplace has to provide, he compiled a set of features that

[12] http://askjeeves.com, accessed: 2015-05-31.

[13] http://factual.com, accessed: 2015-05-31.

[14] http://infochimps.com, accessed: 2015-05-31.

[15] https://datamarket.com/, accessed: 2015-05-31.

[16] Acquired by Qlik available at: http://www.qlik.com/, accessed: 2015-05-31.

[17] https://datamarket.azure.com/browse/data, accessed: 2015-05-31.

[18] http://www.microsoft.com/, accessed: 2015-05-31.

[19] http://www.qlik.com/, accessed: 2015-05-31.

[20] Formerly available at: http://kasabi.com/.

is representative of the definitions given above. He mentions the following criteria: 1) gather data from multiple public and private sources, 2) offer individual data sets for download, 3) harmonise data, 4) offer an API, 5) accept contributed data, i. e., allow for third-party data supply, 6) aggregate data sets, 7) offer on-site tools for manipulation and visualisation, 8) nurture a community, 9) offer a market, i. e., offer a choice, 10) hook directly into other tools (in fact a resulting possibility owing to APIs), and 11) segment the space, by focusing on certain domains.

While he acknowledges that there is a lot of interest (and money) in data marketplaces, he also states that data marketplace providers and customers "are still struggling to understand what anything is worth." [Mil12b]. As challenges he identifies figuring out what can be charged for data in an environment where plenty of offers are free, explaining the huge cost of transforming raw data to clean data, and translating these costs into a viable pricing structure. This will be addressed in this work. Furthermore, he states that the market for data is very heterogeneous and advocates that data should become a commodity. This, as was elaborated before, is something that is very difficult (and not desirable from a providers point of view) to achieve for data.

Individual data providers that have been investigated and shall be described here include Kasabi [MD12], Freebase[21] [BEP+08], FactForge[22] [BKO+11], and Microsoft Azure Data Marketplace [Mic11]. Furthermore, the Marktplatz für Informationen und Analysen (MIA)[23] is to be mentioned which is a research project founded by the German Federal Ministry for Economic Affairs and Energy[24] with the aim of providing an information marketplace infrastructure focusing on the German Web [MIAed].

Kasabi wanted to bridge the gap between supply and demand of data using the Resource Description Framework (RDF) as a means to store data [MD12]. However, they went out of business in July 2012 in order to refocus [Dod12]. This potentially implies a lack of user acceptance or willingness to pay as it seems unlikely that a viable business would shut down. The problem of scale is also addressed in interviews with former managing personnel of Swivel[25], a data exploring and visualisation platform that went out of business in 2010 [Kos10]. Former CEO Brian Mulloy said Swivel had only very few customers because the company did not fully understand the needs and demands of

[21] http://www.freebase.com/, accessed: 2015-05-31.

[22] http://factforge.net/, accessed: 2015-05-31.

[23] http://mia-marktplatz.de/, accessed: 2015-05-31.

[24] http://www.bmwi.de/EN/, accessed: 2015-05-31.

[25] Formerly available at: http://Swivel.com.

their customers and, thus, had problems generating revenue. Nevertheless, he expressed confidence that it is generally possible to turn data into profit. However, he emphasised that this is likely to be achieved by a very big company [Kos10].

Freebase consider themselves to offer "a database system designed to be a public repository of the world's knowledge" [BEP+08]. To achieve this, Freebase aims at combining the advantages of structured databases with collaborative wikis. Concerning the community aspects, Freebase could also be considered to be similar to the Web in your Pocket (WiPo) which is discussed in Chapter 4. In contrast to WiPo, it also provides public read and write access through an HTTP-based API and, thus, can also be considered a data marketplace even though it rather aims on storage of information than on trading of data.

FactForge wants to serve as an entry point to a number of linked open data resources [BKO+11]. To this end, FactForge integrates eight central linked open data sets into one view. Consequently, it is built following the Linked Data paradigm established by Berners-Lee [Ber09].

The Microsoft Azure Data Marketplace is a global online platform enabling customers to search for, buy, and sell public domain and commercial data as well as Software-as-a-Service (SaaS) applications [Mic11]. It is targeted at consumers and providers of data alike as it tries to simplify processes for both parties. Microsoft [Mic11] mentions the following key features: 1) a global marketplace for information and application extending the reach of data providers, 2) a unified billing infrastructure, 3) diverse content types, 4) robust security and availability, 5) integration with other Microsoft products, 6) a rich set of tools, and 7) analytic features.

The Microsoft Azure Data Marketplace is special among those data marketplaces discussed here because it allows for a combination of applications and data sets in order to provide the best value to customers who might not be able to make sense of the raw data alone. Thus, it can be stated that it is the most complete data marketplace with regard to the functionality presented in Section 5.3.1.

Having highlighted that a number of data marketplaces exist which aim at providing value to their customers, some authors argue that while third-party data has some value to businesses, in-house data is more important, e. g., [Ros14]. As a result, it is recommended to closely observe what benefits third-party data provides to a business [Ros14].

This thesis focuses on data marketplaces that only offer business data. However, there is also research into the value and tradability of personal data. In this context, [GHH13] and [PPG+13] shall be mentioned as examples. The first investigates how Facebook data can be of value in the context of recommender engines. The latter states that even though people are generally concerned about their personal data, they do not value it highly enough to be willing to pay for protection or control of their personal data. Li et al. [LLMS13] suggest a framework for pricing private data based on its accuracy and compensate data owners for their loss of privacy.

5.4. Pricing of Information Goods

As pointed out in Section 5.2.3, data as an information good has some peculiarities with regard to its cost structure. Therefore, established pricing models, such as cost-based pricing, do not work for digital goods [HHS06]. The wide-spread misconception – in particular in less economic domains – that prices should be based on costs [MSLV12], can therefore be refuted. As a consequence, other pricing mechanisms have to be found. The topic is particularly important because Theysohn et al. [TPS05] found pricing to be the second most important factor in selling information goods after branding, based on an interview study with 20 paid content managers. Admittedly, this may not be a representative study, but it supports the assumption that pricing is a comparatively important factor when selling information goods.

This section first elaborates on why cost-based pricing alone does not work for information goods. Then, a classification of information goods will be provided and data as well as data marketplaces will be classified accordingly. Subsequently, appropriate pricing strategies for data and data marketplaces will be discussed. Finally, this section is concluded by outlining related works on information good pricing in the economics domain.

5.4.1. The Theory of Pricing of Information Goods

Economic intuition supposes that a market for information goods should not work because it is supposed that on perfect markets goods are sold at marginal costs which are practically zero for information which should lead to all prices tending towards zero [LS11]. More detailed, the low marginal costs of information goods can lead to price fights because the overall costs are irrelevant for short term decisions. As long as suppliers realise positive margins on each unit sold, it makes sense to keep on selling. In the long run, this approach is dangerous because overall costs have to also be covered, which is hard to achieve with minimal profit margins [SV99]. Nevertheless, a market for data and data-related services exists, as of now. One reason for this could be that data has not yet undergone tremendous homogenisation. This is also evident from the argument that information good providers can (to some degree) be viewed as monopolists because their offerings are unique in their domains [FOS97; WB10].

As pointed out previously, cost-based pricing, which is usually done based on variable costs, does not work for information goods [HHS06]. Similarly, pricing based on the competition's prices does not work if products are very similar because this leads to price fights [HHS06]. Applying these pricing mechanisms is among the top ten pricing mistakes [Sjo14]. Thus, pricing for information goods should consider information value rather than the cost of production. However, this is complicated by the fact that different people have very different value attributions. Therefore, it makes sense to differentiate products and prices as well as to offer different product price combinations for different customers [SV99; HHS06].

Supposing – as argued in Section 5.2.3 – data providers can be seen as operating under monopolistic competition and, thus, a company can indeed set prices to some degree – it can consider discriminating customers based on their willingness to pay in order to maximise the company's profit. This is owing to the fact that the producer surplus increases at the expense of consumer surplus. However, the overall welfare shrinks, too (see Section 5.1).

Economists differentiate three degrees of price discrimination, first introduced for monopolists by Pigou [Pig20], in 1920. Nowadays, these are referred to by descriptive names that illustrate the three degrees by Shapiro and Varian [SV99], which will be given in parentheses in the following description.

1. Degree price discrimination (*Personalised Pricing*) means charging exactly the reservation price of each individual customer leaving no consumer surplus.

2. Degree price discrimination (*Versioning*) means grouping customers by their willingness to pay in such a way that customers in each group pay a little less than (or exactly) their reservation price. This leaves a little surplus for each group.

3. Degree price discrimination (*Group Pricing*) means grouping customers according to an a priori identified criterion and charge groups separate prices. This leaves a little surplus for each group, too.

As of 1920, Pigou [Pig20] stated that only third degree discrimination was to be found in practice. However, in 1999, Shapiro and Varian [SV99] gave real-life examples for all three. As an example for personal pricing they mention the analysis of click streams of online retailers to make personalised offers and mailings that offer the same product to different customers at different prices. This individualisation, also referred to as mass customisation, is facilitated by the Internet [SF08].

Addressing versioning, as described by Shapiro and Varian [SV99], it should be made clear that it can be argued that this is only one possible form of second order price discrimination. While it achieves the result that customers can be categorised into groups with different reservation prices, it also implies different products. However, often product differentiation paves the way for price differentiation [NRRS05]. Thus, versioning is not quite the same as described in [Pig20]. Additionally, the groups are not built by the provider knowing which price to ask of which customer but by offering different versions of a product. With the help of a principle referred to as *self-selection* customers reveal their preferences and classify themselves into one of the groups [SV99]. In an interview study, it has been found that paid content managers prefer self-selection over price discrimination without self-selection [TPS05]. The fact that information goods can be modified rather easily also facilitates versioning [SF08; HHS06]. Sometimes versioning is also referred to as vertical differentiation, e. g., in [BC01a].

Regarding group pricing, Shapiro and Varian [SV99] mention four relevant points: *price sensitivity, network effects, lock-in,* and *sharing*. Price sensitivity refers to differentiating consumers based on their social status, which is also linked to their ability to pay and, accordingly, to their reservation price. As an example, student discounts can be mentioned. Network effects apply to information goods for which a higher distribution increases the utility of the good, such as communication software like Skype. Regarding pricing, it is recommended to adjust prices to the number of users within an organisation. Lock-in refers to getting customers to buy a product that has high switching cost at a rather cheap price and then gradually move them to another group. This, for instance, applies to discounts for new customers in the mobile communications market. Furthermore, this effect is even stronger for products which also have network effects. Finally, they mention sharing arrangements, i. e., building customer groups based on how many individuals have access to the information sold, such as libraries versus individual consumers.

Indifferent of which price discrimination model is appropriate in any given situation, it is quite important to differentiate information products to maximise profits. In contrast, commoditisation should be avoided because it potentially leads to price fights which are fierce in markets with marginal costs close to zero. Thus, it is always important to differentiate the product. To this end, it is among many other things important to know the customers and the market, personalise information and prices where possible, introduce versions that address different preferences of customers, and understand how much it costs to produce information [SV99]. This will be dealt with more extensively in the next section.

These insights notwithstanding, it should be noted that this behaviour also has its limitations. Shapiro and Varian [SV99] point out that in the U. S. the *Robinson-Patman Act* of 1936 prohibits price discrimination if it "effectively lessens competition". Similarly, in the *Treaty on the Functioning of the European Union* [EU12] Article 102 (c) prohibits "applying dissimilar conditions to equivalent transactions with other trading parties, thereby placing them at a competitive disadvantage". Nevertheless, price discrimination is only illegal if it indeed lessens competition [SV99].

5.4.2. Classifying Data and Data Marketplaces as Information Goods

In order to price information goods, one needs to know the potential customers, their needs, and their usage behaviour [HHS06]. In other words, it is important to understand a) who actually generates the income (i. e., who is the customer) and b) what the actual source of income is (i. e., what is their need or the product) [SF08; SSW05]. This is important as the source of revenue defines the market and the competition [SSW05]. Regarding the first, three categories can be differentiated [SF08]:

- Buyers, e. g., on film streaming platforms such as Netflix buyers pay a usage fee.
- Vendors, e. g., on ebay[26] vendors of goods pay a brokerage fee.
- Advertisers, e. g., Google services are mostly free to use but are indirectly paid for by advertisers.

[26] http://www.ebay.com/, accessed: 2015-05-31.

With regard to data marketplaces, it is sensible to suppose that vendors of data pay a commission for using the data marketplace infrastructure. With regard to the actual information good data, it is most likely that it will be paid for by purchasers.

The second important thing to consider is the actual source of revenue. To this end, several works having done this will be reviewed and a categorisation of data and data marketplaces will be conducted. The results are illustrated in Figure 5.8. On a very high level, the sources of revenue can be divided in into four categories [SF08; Wir10]:

- Content as a business model comprises collecting, selecting, packaging and provisioning of content.
- Commerce as a business model is concerned with providing a platform for trading goods and services.
- Context as a business model refers to navigational and aggregation services, such as search engines.
- Connection as a business model provides platforms through which people can connect, such as social networks.

Initially, one might think that data marketplaces follow a *context* business model because they aggregate data and provide navigation to data that might partially be available spread over the Internet. However, more than that, they provide a platform on which third parties may trade data. Therefore, it is supposed that they follow a *commerce* business model. For data, the case is less complicated as it is intuitively clear that data providers follow a *content* business model because they either collect or create data themselves and offer it in different packages.

Wirtz [Wir10], which will build the baseline of this comparison, provides subcategories for each of the models presented above. However, as only content and commerce have relevancy in the context of this work only these will be described more extensively. For content, he names a) E-Information, e. g., news and factual content such as SPIEGEL ONLINE[27], b) E-Entertainment, e. g., games, music, and film providers, such as Netflix, c) E-Education, e. g., the Virtual University[28], and d) E-Infotainment, a combination of E-Information

[27] http://www.spiegel.de/, accessed: 2015-05-31.

[28] http://vu.org/, accessed: 2015-05-31.

and E-Entertainment, such as sports news websites, e. g., KICKER[29]. From this – as well as from the definition of information in Chapter 3 – it should be clear that data can be categorised as *E-Information*.

Regarding commerce WIRTZ [Wir10] mentions the sub categories a) E-Attraction businesses serving the initiation of business relations, such as banner ads, e. g., GOOGLE ADWORDS, b) E-Bargaining / E-Negotiation businesses providing platforms to negotiate prices, such as auction platforms, e. g., EBAY, c) E-Transaction businesses simplifying transactions between two parties such as payment providers, e. g., PAYPAL[30] and, d) E-Tailing combining all of the above, for instance online retailers, e. g., AMAZON[31].

HUI AND CHAU [HC02] classify digital products into three categories: a) Tools & Utilities, i. e., software products that serve a given purpose such as creating documents for a word processor and are commonly downloadable, b) Online Services, i. e., services that provide users with access to a network, such as online telephony, or that serve a purpose such as providing search, and c) Content-Based Digital Products, i. e., products that provide value simply by their content, such as online newspapers. While *Tools & Utilities* is a new category, their *Online Services* are approximately equal to *Context* and *Connection* described by WIRTZ [Wir10]. Similarly *Content-Based Digital Goods* correspond to the *Content* category.

Another classification is provided by THEYSOHN ET AL. [TPS05] who distinguish a) consumable digital goods, which can only be used for a certain period, b) durable digital goods, which can be used indefinitely, and c) digital services. In contrast to [HC02], their *Digital Services* also comprise trading platforms. However, their distinction between *Durable* and *Consumable Digital Goods* adds a new dimension to *Content* [Wir10] because it differentiates the durability rather than the type of content. Furthermore, *Consumable Digital Goods* may also include software.

SKIERA AND LAMBRECHT [SL07] classify sources of revenue in the digital context into a) products, i. e., physical or virtual goods, such as books or search services, b) contacts, i. e., contacts with customers are used to gain revenue through advertising for instance, and c) information, in this context used synonymously to user behaviour data mostly gained as by-products. Furthermore, they point out that often more than one source is used and that all are interdependent. Compared to the here used baseline [Wir10], *Product* comprises services which

[29] http://www.kicker.de/, accessed: 2015-05-31.

[30] https://www.paypal.com/, accessed: 2015-05-31.

[31] http://www.amazon.com/, accessed: 2015-05-31.

are provided by *Context* and *Connection* but also physical goods which are commonly sold through *E-Tailing*. However, the source of revenue in this case is not the platform but the physical good. *Information* is special in that only user behaviour data and panel data are considered. Relaxing this restriction, it can be considered *E-Information* as defined by Wirtz [Wir10].

Similar to [TPS05], Herrmann et al. [HHS06] focus on the way in which information goods are used. They differentiate between, a) durable digital goods (e. g., encyclopaediae), b) consumable digital goods (e. g., communication services), c) event goods (e. g., news), d) experience goods (e. g., music), e) network goods (e. g., auction platforms), and f) applications (e. g., software or online banking). Similar to [HC02] they include *Applications* which are approximately equal to *Tools & Utilities*, while their category of *Durable Digital Goods* is similar to that of [TPS05]. However, Herrmann et al. [HHS06] diverge in what they consider to be *consumable digital goods*. In contrast to [TPS05], they include services but do not include software. Furthermore, they introduce the concept of *Network Goods*, i. e., goods or services that benefit by an increasing number of users. This concept is basically applicable to all products and services in the area of *Context, Connection,* and *Commerce*. This means the goods in these categories can be, but do not have to be, *Network Goods*. Similarly, they introduce *Event* and *Experience Goods*. While the first can be interpreted as sub-group of *E-Information*, the latter corresponds to *E-Entertainment*. At this point, it should be made clear that *Experience Good* as a category name is unfortunate, as mentioned previously, most information goods are experience goods in that they have to be consumed in order to judge their value. Hence, the term *E-Entertainment* will be used here which has basically the same definition. All of the definitions listed above are illustrated in Figure 5.8.

While this list is comprehensive in terms of revenue sources relevant for this work, it should also be mentioned that further revenue sources for digital goods exist. In the context of blogs for instance, Karla [Kar08], while mentioning some known revenue sources – advertisement (E-Attraction), paid content (Content), and Business-Intelligence (BI) (Information) –, names donations, sponsoring, merchandising, and syndication, i. e., re-usage of digital content by further providers, as possibilities. Out of these, donations and sponsoring might only be relevant when data has to be offered for free. Syndication mainly applies to E-Entertainment or E-Infotainment goods, which makes it less applicable to data although not inapplicable. Merchandising, in contrast, is not applicable, at least as long as no strong brand has been established.

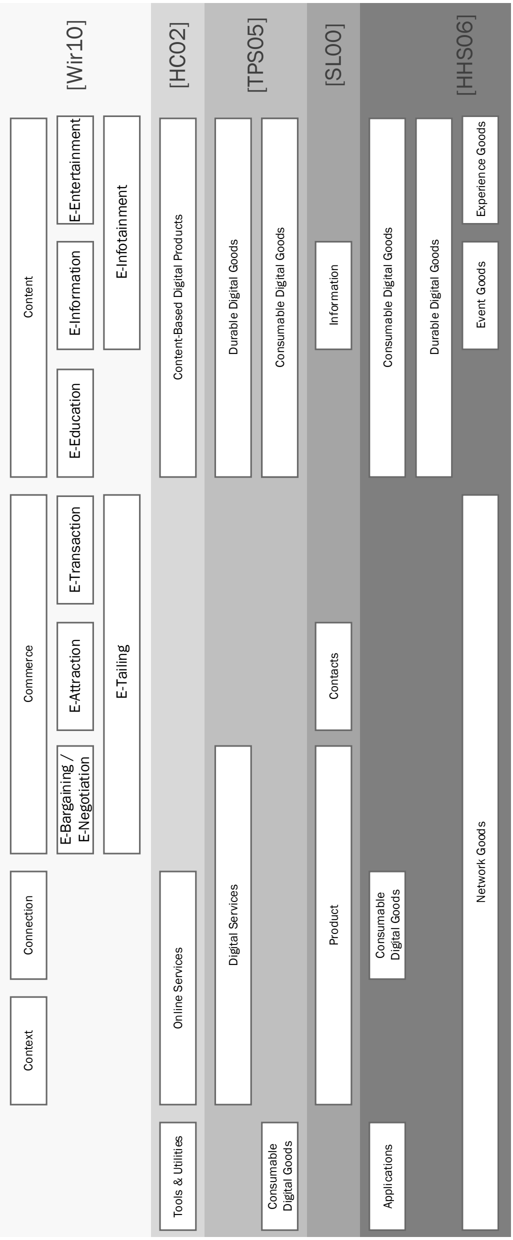


Figure 5.8.: Comparison of Information Good Classification.

These elaborations can now help to classify data as an information good. It is intuitively clear that data falls into the *Content* or *Content-Based Digital Goods* category and will – most of the time – be categorised as *E-Information*. Regarding the sub-forms, data as such is mostly a *Durable Digital Good* because it does not wear. However, depending on the context, some data may be *Consumable Digital Goods*. For instance, weather forecast data can only be used so long as the date it forecasts has not been reached yet. As evident by this example, in particular *Event Goods* are likely to be *Consumable Digital Goods*. However, it has to be mentioned that it may also be used for analysis afterwards, but this is not its intended use and as such this type of data classifies as *Consumable Digital Good*.

At this point, it should be added that in contrast to other digital goods, such as software, data as a *Content-Based Digital Product* cannot be tried before it is bought, which was previously identified as Arrow's paradox (see Section 5.2.3). Indifferent of what subset or excerpt of the data is provided to a customer, it will always be a sample that does not convey the overall quality [HC02]. However, regarding flexibility, *Content-Based Digital Products* have an advantage as they are very flexible and, thus, allow for various forms to be offered. Therefore, versioning can be applied particularly well to data. This implies that the market is less likely to be competitive and, as a result, monopolistic competition is to be expected [HC02].

Data marketplaces fit best into the category of *E-Bargaining / E-Negotiation* as they provide platforms for trading data. However, depending on the specific implementation – in particular if data marketplace operators trade themselves – data marketplaces may fall into other *Commerce* categories and may even be considered *E-Tailing* businesses. One might also think that data marketplaces can be seen as *Tools & Utilities*. However, as these shall by definition be downloadable this category does not apply to data marketplaces.

5.4.3. Pricing Strategies

Having established that for information goods traditional cost-based pricing models do not work [SF08; HHS06], this section is dedicated towards outlining pricing models that can cope with the special demands of information goods. Furthermore, it has been mentioned that in order to price information goods it is important to understand the source of revenue as well as to answer the question of who pays for the information good [SF08; SSW05]. Based on a discussion of potential revenue sources, the whole scope of information goods could be narrowed down to *Content* and *Commerce* as relevant for this work.

Commonly, it is suggested to exploit the possibilities that the Web has to offer in order to build suitable more fine-grained pricing mechanisms [SF08; SSW05; Kun12; HHS06; HHS11]. This explicitly includes pricing models that are otherwise seldom to be found in the offline world [SF08]. However, there is hardly any guidance for firms [HHS11]. Overall, information goods sold over the Internet facilitate price differentiation for a number of reasons. Simon and Fassnacht [SF08] have collated the following reasons:

- Costs of implementing price differentiation are rather low compared to the offline world.
- Product differentiation can be achieved with very low marginal cost.
- Using *Digital Rights Management (DRM)*, for instance, can ensure that purchased goods cannot be sold on. However, this mechanism is not very much liked amongst consumers.
- Establishing different distribution channels is simpler than in the offline world. For instance, users can be differentiated based on their IP addresses' geolocation but served from the same country.

Very broadly speaking, differentiating prices is advantageous to reach customers with heterogeneous willingness to pay, while at the same time it is easy to implement in online sales. As a consequence, it is simple to provide products that diverge in a number of dimensions and are priced accordingly. Not segmenting the market in such a way is also one of the top ten pricing mistakes that can be made according to Sjofors [Sjo14]. Thus, it is little surprising that this is currently applied by a majority of paid content providers [TPS05]. In order for price differentiation to work, Reinartz [Rei02] identified five criteria [Rei02; NRRS05]:

1. Customers must be heterogeneous in their willingness to pay, otherwise this would render the idea of price differentiation useless.
2. Markets must be segmentable, i. e., it must be possible to recognise customers who might be willing to pay more. He suggests using account numbers and similar identifiers, which is facilitated by Web technologies.
3. The potential for arbitrage should be minimised, i. e., customers should be hindered to resell goods they purchased at a low price with a profit.
4. The cost of segmenting must not be greater than the increase in profits.
5. Customers should perceive the pricing as fair.

At a high level, pricing models can be differentiated depending on whether customers participate in the process of setting prices (*interactive pricing*) or not (*non-interactive pricing*) [SF08; SSW05]. Some authors use the term *participative pricing* instead of *interactive*, e. g., [Kun12].

Non-interactive pricing can be considered a "take-it-or-leave-it-offer", the simplest form of which is one consistent price [SSW05]. Differentiating prices, however, has the advantage of generally leading to an increase in profits as they allow for the better tapping of the willingness to pay and even reach customers with a willingness to pay lower than the market price would be [SSW05]. For non-interactive prices, Skiera et al. [SSW05] mention a) characteristics of users, b) characteristics of usage, and c) characteristics of products as possible factors for price differentiation.

Regarding *characteristics of users*, they mention individual pricing (first order price discrimination) versus group prices (third order price discrimination). However, in the latter case it has to be verifiable whether an individual belongs to a group. The first requires extensive knowledge about the individual which even Internet technology cannot provide entirely to date [SSW05].

Regarding *characteristics of usage*, they mention time as a differentiating factor. This can be done either *static*, i. e., prices vary in pre-set intervals such as special weekend tariffs for telecommunication, or *dynamic*, when prices vary depending on the amount of products sold or the novelty of a product [SSW05]. Another factor in this category is *quantity*. In this case it can be differentiated whether one product is to be sold and bulk discounts are offered or whether more than one product is to be sold in bundles, which is commonly beneficial for products with marginal costs close to zero [SSW05]. Last in this category is *search cost*. This exploits the fact that different customers have different search costs. It supposes that customers with high search costs have a higher willingness to pay if they can avoid the search costs [SSW05; SF08].

Regarding *characteristics of product*, versioning is suggested, the implementation of which is facilitated by Internet technologies [SSW05; SF08] and the fact that information goods can be easily modified [HHS06]. This allows for the provisioning of custom-tailored information products to customers, which is also referred to as pointcast information goods as opposed to the traditional broadcasting [Kar07]. Furthermore, easy versioning enables the exploitation of the long tail phenomenon identified by [And07], i. e., serving customers with individualised products who would not be served if only a limited number of standardised versions could be offered because of too special requirements [Kar07].

Before a more extensive discussion of versioning, Figure 5.9 provides an overview of non-interactive pricing models.

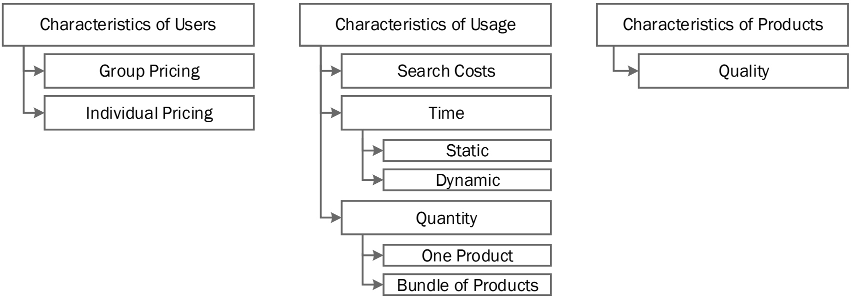


Figure 5.9.: Non-Interactive Pricing Models, Adapted from [SSW05].

Herrmann et al. [HHS06] identify extent (e. g., in context of factual knowledge), number of units (e. g., for communication services), recentness (e. g., news, weather), compression quality (e. g., films), number of users (e. g., auction platforms), and effort of learning (e. g., software) as factors which provide utility and can, therefore, be used to create different versions of a product. Similarly, Shapiro and Varian [SV99] name a number of product dimensions that allow for differentiation (delay, user interface, convenience, image resolution, speed of operation, format, capabilities, features, comprehensiveness, annoyance, support). A collocation of both is represented in Table 5.3. While largely fairly self-explanatory, it is also obvious that the categories are not mutually exclusive. For instance, delay (in delivery), features, and annoyance (unnecessary dialogues) could be seen as sub-factors of convenience. Also, the given likely uses and users have to be understood as examples because other options are possible, too. Nevertheless, this overview serves its purpose well to indicate that there is a plethora of options to create different versions of an information product. In Table 5.3, those attributes relevant for data as such and possibly for data marketplaces, too, are annotated with ✔, those relevant for data marketplaces only are annotated with ✔/✖, those not applicable to either are annotated with ✖.

Shapiro and Varian [SV99] also mention quality in general as another important concept in the context of versioning. In this respect, they recommend ensuring that low-price and high-price products have an appropriate difference in price and quality. Otherwise, when versioning is used, the problem of cannibalisation may occur, i. e., customers who would have bought the high-end

Table 5.3.: Overview of Possible Versioning Factors, Compiled from [SV99] and [HHS06].

Product Dimension	Likely User/Uses	Relevancy for Data	Source
Extent	Light / Heavy Users	✓	[HHS06]
Number of Units	Light / Heavy Users	✓/✗	[HHS06]
Recentness	Patient / Impatient Users	✓	[HHS06]
Quality	(Less) Demanding Users	✓	[HHS06]
Number of Users	Goods with Network Effects	✓/✗	[HHS06]
Effort of Learning	Casual / Intensive Users	✓/✗	[HHS06]
Delay	Patient / Impatient Users	✓	[SV99]
User Interface	Casual / Experienced Users	✓/✗	[SV99]
Convenience	Business / Home Users	✓/✗	[SV99]
Image Resolution	Newsletter / Glossy Uses	✗	[SV99]
Speed of Operation	Student / Professional Users	✓/✗	[SV99]
Format	On Screen / Print Uses	✓/✗	[SV99]
Capabilities	General / Specific Uses	✓/✗	[SV99]
Features	Occasional / Frequent Uses	✓/✗	[SV99]
Comprehensiveness	Lay / Professional Users	✓	[SV99]
Annoyance	High- / Low-Time-Value Users	✓/✗	[SV99]
Support	Casual / Intensive Users	✓/✗	[SV99]

version may switch to the low-end version if it provides the better benefit-to-cost ration to them [BC01a].

Furthermore, Shapiro and Varian [SV99] point out that low quality information goods are sometimes more expensive to create than high-quality goods, which also underlines the fact that it is not always sensible to set prices based on cost. For instance, delivering information with a delay requires additional storage and processing to make sure the information is only provided with a delay and not as soon as it is available. Moreover, it has to be ensured that customers cannot easily turn low-value into high-value product themselves, e. g., by using a hidden API or changing API parameters to receive real-time data instead of delayed data.

In contrast to this stand *interactive* pricing models in which customers and sellers participate in the pricing process [SF08; SSW05] and which are particularly well supported by Internet technologies compared to their traditional counterparts [SF08]. Narahari et al. [NRRS05] go even further and proclaim *dynamic* pricing, which means "computing the right price to the right customer at the right time," i. e., dynamic adjustment of prices depending on the value that customers attribute to the good. While this can include *interactive pricing*, it can also be achieved without customer participation. For instance, airlines use dynamic pricing over time when selling seats on particular flights. However, determining a given customer's willingness to pay at the point of purchase is all but trivial [NRRS05].

Interactive pricing models can be further categorised into three types of dynamic pricing depending on who eventually determines the price – buyers, sellers, or both [SSW05]. If negotiations, which are commonly unstructured, or exchanges, i. e., highly organised markets that allow for efficient trading of specific homogeneous goods are used, both parties have an influence on the final price.

If the vendor is to set the final price, reverse auctions and power shopping are adequate means. In the first model, buyers define a good or a service they want to acquire and providers underbid each other. In the second model, also known as co-shopping, it is supposed that providers offer bulk discounts. As a reaction, a number of demanders unite to form a virtual buying cooperation to achieve a high bulk discount when buying the good.

Finally, when auctions or reversed pricing is used, buyers determine the final price. Using auctions, sellers hope to achieve the reservation price from buyers and buyers hope to acquire the product as cheap as possible, i. e., at the cheapest price a vendor accepts [ST06]. Auction of various forms exist, most notable are online auctions over a period of time and a fixed minimum increment and Dutch auctions in which prices are reduced from a high start price within specific time intervals until someone accepts the price [SF08]. Auctions have been extensively studied which is evident by works such as [Kri09] and have even been modelled for digital goods [GHW01]. However, selling the same digital good twice was named an open problem by Goldberg et al. [GHW01]. Thus, data auctions are not an option if the same data is to be sold more than once.

Reversed pricing has been defined by Bernhardt et al. [BSS05] as a pricing scheme in which buyers bid for a good and the deal is made if the price exceeds a secret threshold set by the seller. Even after the transaction, prices are usually not communicated. Buyers do not compete directly, and are dealt with on a first-come-first-served-basis. However, the latter is not applicable to information goods which can be replicated. Reversed pricing has the potential to realise higher prices from customers with a higher willingness to pay while allowing for serving customers with a lower willingness to pay, too [BSS05]. However, buyers may face significant bid cost (e. g., searching, creating the right bid, and waiting for a reply) and prices are likely to be not transparent [BSS05]. Other authors such as Kim et al. [KNS09] or Kunter [Kun12] differentiate between *Pay What You Want (PWYW)* and *Name Your Own Price (NYOP)* pricing schemes. While the latter essential fits the definition of reversed pricing [BSS05], the first does not imply a threshold. NYOP is particularly interesting because it has been applied to data in an abstract form by Tang et al. [TASB14; TSBV13]; however, overall there is little literature discussing NYOP [ST06].

Analysing PWYW pricing Kim et al. [KNS09] found in three field studies (buffet meal in a restaurant, hot beverages at a delicatessen, and movie screenings at a cinema) that prices paid were significantly greater than zero which contradicts economic rational. Similarly, Spann and Tellis [ST06] discovered that bidders do not behave as expected of a rational price-minimising bidder in the case of a NYOP experiment in the context of airline ticket sales. It can be stated that most customers pay based on a perceived fairness in the deal and their internal reference price in PWYW [KNS09]. Whereas for the restaurant and delicatessen cases revenue was greater than the baseline, this was not the case for cinema tickets. Furthermore, they found that in the cinema case customers had a high variance in estimating the variable costs of a ticket, resulting in the prices being perceived as unfair. They concluded that PWYW pricing can be beneficial in particular if variable costs are low because then the major risk of this pricing model, namely that prices paid are below cost, can be minimised. Thus, they do not recommend it for high-priced products as customer then might see their advantage over the fairness. Furthermore, PWYW can be seen as an activity for sales promotion and marketing which help acquire new customers [Kun12; KNS09].

Conducting three field experiments (ride photos in a theme park, photos on tour boats, and restaurant buffet) Gneezy et al. [GGRN12] found that identity and self-image are important factors influencing the amount people pay in PWYW pricing. However, all studies cited above look at consumers rather than at businesses. Thus, it remains questionable whether businesses have the same idea of fairness or whether it is more likely that they would act more economically rational.

Given modern technology, it is even possible to change the threshold value depending on the user. While an adaptive threshold can increase profits and sellers can react based on bids, this has not yet been implemented because it seems to be less fair to consumers, as well allowing for bid shading, i. e., trying to convince the seller that the willingness to pay is less than it actually is. Furthermore, for NYOP it can be argued that the positive effects of participation outweigh the negative effects of price discrimination [HHS11].

Investigating different information goods, Herrmann et al. [HHS06] present seven price strategies for different goods. They will be briefly recapitulated, aggregated where appropriate, and analysed regarding their applicability to data.

Firstly, they suggest using network and lock-in effect as well as providing introductory discounts. For data, this might not be simple to achieve because data as such is not a network good. However, for data marketplaces this may be a viable option. One particular promising strategy, which has been successfully applied by ebay, is called *follow the free*. Using this strategy, at first the service is provided for free and once a critical mass has been reached, customers (data providers in the case of data marketplaces) are charged for the services.

Versioning, as extensively elaborated on before, has a vast potential of differentiating customers by their willingness to pay for certain criteria. In this context, all product attributes relevant for customers should be considered. The demand for higher-priced goods can be increased by utilisation of non-linear pricing models, for example, if prices increase linearly the quality should increase over-proportionally. In order to tap the willingness to pay as much as possible, timeliness of information goods should be considered.

In the context of versioning, Goldilocks pricing is often suggested [CY07; SV99]. This means that it should be considered that customers commonly choose mid-range-products rather than high-end or low-end if they have no further information. This phenomenon is also known as *extremeness aversion*. Thus, suppliers are well-advised to consider this when determining prices of low-end, mid-range, and high-end versions [SV99].

When bundling is an option, quantitative differentiation is suggested in order to increase turnover [HHS06]. However, this is not applicable to raw data because it is only one good. Nevertheless, it may be a viable strategy for vendors on data marketplaces providing different sorts of data, e. g., weather and stock exchange data. Furthermore, Herrmann et al. [HHS06] recommend optimisation of the applied billing method, i. e., gauging unit pricing against flat rates. However, this is only discussed in the area of online service and cannot be applied to data as such.

5.4.4. Related Work on Pricing of Information Goods

The optimality of pricing mechanisms for information goods has been researched very broadly but – to the best of the author's knowledge – has never been applied specifically to the situation of data. For instance, Wu and Banker [WB10] investigated monopolistic information service providers, using the term very broadly to include social network providers and TV broadcasting companies alike. They rectify the monopoly assumption by highlighting that for instance Facebook and Twitter have a unique offering, granting them a monopoly in their domain. This is a reasonable assumption and shall be adapted here.

Moreover, Wu and Banker [WB10] investigated which of the pricing models mostly used in practice – according to them flat rate, pay-per-use pricing, and two-part tariff – is the most profitable for a monopolistic information provider. They found that pay-per-use is dominated by the other two schemes which are on a par if homogeneous consumption is assumed. In contrast, when consumption is heterogeneous, a two part tariff is the most profitable.

Choudhary [Cho10] examined how pricing schemes alone can be a differentiating factor when selling information goods and help avoid zero profit competitions. Furthermore, he showed that applying different pricing models can increase profits for all involved producers in a duopoly situation. This shows how important it is to choose a pricing model wisely.

Similarly, Fishburn et al. [FOS97] investigate a duopoly situation in which one provider charges a fixed fee and the other uses a pay-per-use model. Using mathematical modelling, they reached the conclusion that most of the time price wars will result and identify some conditions under which equilibria can be reached in their model. This, too, underlines the fact that pricing is highly important.

Bhargava and Sundaresan [BS03] look at the viability of contingency pricing, i. e., pricing based on performance, in the context of information goods to address the problem of quality uncertainty. They found that contingency pricing is attractive when a company is able to perform better (regarding quality) than perceived by the market. Given that the authors look at delivery quality, such as download speed, this sort of contingency pricing is not applicable to the plain data looked at in this work.

Wu et al. [WHCA08] analysed, using a nonlinear mixed-integer programming approach, which bundle size and price combination is optimal for a monopolistic provider of multiple information goods and compared it to individual sales. They found that individual sales can be enhanced by using customised bundling in the case of music providers. However, in contrast to their paper, this work considers only one good and, thus, versioning is more relevant here than bundling.

Given a monopoly assumption, Bhargava and Choudhary [BC01b; BC08] investigated the conditions under which versioning (i. e., second degree price discrimination) is optimal for pricing information goods. They came to the conclusion that it is when the market share of the low quality version alone would be larger than that of the high-quality product alone.

Chang and Yuan [CY07] present a unified framework of information good pricing models which contains far more pricing schemes than relevant for this work. However, it provides a good overview of possible pricing schemes. Similarly, Narahari et al. [NRRS05] provide a classification of dynamic pricing models that is beyond the scope of this work.

6. Pricing Data

In the previous chapter, it has been shown that there is a market for data providers. However, it also has been argued that there is little knowledge amongst providers regarding the pricing of their information goods [BHS11; MSLV12; Mil12b].

From the interviews [MSLV12; SLV15] and the relevant economic literature, such as [SV99], it is obvious that the value of data is highly domain specific. This is owing to the fact that data has no inherent meaning, as argued in Chapter 3. Given that data only becomes information if meaning or context is added to it [LLS10; Nor11; RK96], i. e., it is brought into a form meaningful to humans, it is obvious that the value of data highly depends on the context and the buyer of the data. These factors, however, cannot be determined completely automatically. Thus, it can be established that pricing on data marketplaces in general is still an unsolved issue. In particular the fact that it is difficult for providers to gauge the willingness to pay of customers as they do not know the purpose the data is bought for. Furthermore, no pricing model exists that considers two providers offering similar information goods [BHK+13].

The remainder of this chapter is structured as follows. Firstly, previous work on pricing of data on data marketplaces is outlined. Based on this, the focus of this part will be established to be quality-based pricing. Next, data marketplaces will be formally defined in Section 6.3. Subsequently, Section 6.4 will extensively discuss data quality and its dimensions, introduce measures to gauge data quality, develop a quality scoring model for data marketplaces, and apply these to an example use case, namely that of weather data providers. Thereafter, Section 6.4 introduces a quality-based pricing model base on the Multiple-Choice Knapsack Problem (MCKP); this too will be demonstrated by an example. Eventually, this chapter is concluded in Section 6.6 by summarising the main points and outlining possible future work.

6.1. Pricing of Data on Data Marketplaces

In 2011, Balazinska et al. [BHS11] put the topics data marketplaces and data pricing on the research agenda of the database community. They discussed two pricing schemes: subscriptions with a query limit – effectively tiered pricing – and a pay-per-use scheme in which the unit of consumption is a tuple. Furthermore, they identified four problems with pricing on data marketplaces:

1. Current pricing schemes allow(ed) for arbitrage.
2. Pricing is based on the assumption that all data sets are of equal value.
3. Customers have to store purchased data themselves or have to pay for it again.
4. No guidance is given to data providers on how to set their prices.

It can be argued that the first two weaknesses only occur because of the last [MSLV12], in particular the second weakness can be attributed to lacking guidance, while the third weakness can be considered a mere technicality. The first weakness has been addressed in [BHS11], where the authors have suggested attaching prices to data cells or tuples and selling data on a per-query basis rather than whole data sets. When a query is issued, these basic price queries are aggregated according to predefined rules. As a means they suggest using data provenance techniques to achieve this goal. However, the authors acknowledge that computing these prices is potentially complex [BHS11].

Later, this method is referred to as *query-based data pricing* [KUB+12a]. Koutris et al. [KUB+12a] present a framework that allows data providers to set prices for some (sets of) views and computes prices for queries automatically. Furthermore, it is ensured that the resulting price function is arbitrage-free and discount-free. A prototype has been described as a demonstration in [KUB+12b]. It provides guidance to sellers in that it highlights if the set prices violate the arbitrage-free criterion. In a next step, presented in [KUB+13], the group introduced *QueryMarket* a middle layer software that can be run atop of any database management system that supports the Structured Query Language (SQL). Besides the arbitrage-free criterion the system also allows multiple sellers and shares revenue fairly between them.

While their system is advanced in its capability to calculate prices for individual queries based on an overall price, it does not really address the problem of how much data is actually worth. Also, it does not take into account that the same query may have a different value to different people, which can be exploited to maximise profits.

In [KMKed] current pricing models (pay-per-use and tiered pricing) on data marketplaces are reviewed and an algorithm is proposed that shows that a pricing model is free from arbitrage. While the studies mentioned before have some restrictions on what queries can be offered, Lin and Kifer [LK14] investigate the arbitrage-free criterion for arbitrary queries. They find that pricing one query can lead to undesirable interactions regarding the price of other queries, which needs to be investigated further.

In 2013 Balazinska et al. [BHK+13] presented a discussion of pricing on relational data, arguing that views can essentially be interpreted as versions of the information good data, a suggestion that will be applied in this thesis. Furthermore, they identify three open problems. Firstly, they name the pricing of data updates, i. e., what price to charge if a consumer has purchased a data set that has been updated in the meanwhile and the consumer only wants to pay for the new data. Secondly, they mention the pricing of integrated data and present a complex value chain in which provider A generates data, provider B conducts data mining, and provider C integrates the mining result with other data sets. Finally, they discuss the pricing of competing data sources that provide essentially the same data but in a different quality.

The first challenge can be addressed by calculating the difference between the full price of the new and the old data product. This is similar to the approach suggested by Tang et al. [TASB14] for buying samples of XML data. The second problem can be addressed by introducing intermediary pricing for all providers refining the raw data. This means the raw data vendor operates using established means. Furthermore, all vendors following in the value chain have to deal with the output price of the lower level vendor as cost and build their prices accordingly. As it has not been solved yet, this makes the last challenge an interesting question to address in this thesis.

Another group that investigates data pricing is formed around Tang who has written a PhD thesis entitled *'Quality and Price of Data'* [Tan14] in a very similar domain to this work. In contrast to this thesis, he focuses solely on databases and data quality and the topic of information provisioning is not discussed. The findings relevant to this work have also been published as individual papers ([TSBV13; TWB+13; TASB14]) which built the basis for the argument below.

Tang et al. [TWB+13] argue that using views to attach prices is too coarse and adopt the idea of attaching prices to tuples and use a pricing model that is based on minimal provenance. However, computing prices in this model is $\mathcal{NP}$-hard. Therefore, they also present and evaluate heuristics to approximate the prices. In [TSBV13], the authors introduce the concept of trading data quality for a discount. This is in contrast to all other works looked at so far in which buying data is a "take-it-or-leave-it-decision" – a customer may buy the data at the advertised price or not buy it at all. Aiming at improving this situation, the authors propose to offer buyers the option of naming their own price in order to address customers with a willingness to pay below the full price.

From an economic point of view, this can be seen as a form of PWYW or NYOP pricing as discussed in Section 5.4.3. In this case, the speciality is that the threshold value is known and that for prices below it quality is adjusted. Furthermore, the provider receives exactly the reservation price of the customer who in turn receives a personalised (i. e., quality degraded) offer, as long as a customer's willingness to pay is below the ask price. If it is above the ask price, the demander receives some surplus. From this, it can be concluded that the profit with NYOP is greater as it reaches additional customers. However, this only holds true if it is supposed that customers do not change from the high-end product to a cheaper version.

Tang et al. [TSBV13] propose to trade data accuracy for a discount, with cheaper data being less accurate. To this end, they present a framework in which – if less than the full price is offered – values are randomly drawn from a probability distribution. For the model to be fair, the distance between the probability distribution and the real distribution correlates to the discount. This approach is similar to [LLMS13] in that both approaches offer data with lower quality at cheaper prices. While Tang et al. [TSBV13] allow customers to name a price and decrease the quality accordingly, Li et al. [LLMS13] use the standard deviation as user input and calculate a price based on that. Therefore, the approach suggested in [TSBV13] seems more viable as customers do not need to experiment with input values.

Later, Tang et al. [TASB14] presented a framework to price XML data, keeping the idea of allowing users to trade quality for a discount. In this work, the authors focus on completeness as a quality criterion, i. e., users can decide to get an incomplete sample by proposing a price lower than that advertised by the vendor. To this end, an algorithm collocates a sub-tree of the overall XML tree at random that matches the price offered by the customer. They suppose two application scenarios for this framework. Firstly, users might be on a restricted budget and, thus, satisfied with a subset of the data. Secondly, users might want to get a sample to explore the data set. While exploring the data is a reasonable argument, bearing in mind that data is an experience information good, the first use case has a minor weakness. Customers with limited funds are usually particularly price-sensitive; therefore, it is questionable whether they are willing to buy random data as they would not know what they get for their little budget. That said, the idea of asking customers to reveal their preferences by naming a price, is further explored here.

Also relevant in the area of pricing data is pricing in the technical domain in general. For instance, Upadhyaya et al. [UBS12] investigate how a cloud data service provider should price optimisation that benefit multiple users of the system. To this end, they used a game theory approach in which users actively bid for optimisations and users are then charged according to their bid. In order to provide an incentive for users to reveal their true preferences the approach excludes users without a bid from the optimisation.

Using a stochastic model, Kantere et al. [KDGA11] present a way of predicting the time and number of queries that are necessary to amortise the initial cost of creating the data structure as well as the cost of running the query. Ipeirotis et al. [IAJG06] suggest a cost estimation model that enables adaptive shifting from crawling to querying and vice versa in the context of text-centric tasks. While both works estimate the cost of the respective services, they do not consider the value of their offering to the consumer.

In the context of cloud computing Püschel and Neumann [PN09] have introduced a dynamic pricing model based on utilisation. If a consumer requests data, the seller calculates a price based on client category, resource status, economic policies, and predictions of future job executions and sends it to the consumer who may then choose whether to accept it. This can be considered a form of dynamic pricing. In the same context, Dash et al. [DKA09] have developed an economic model for cloud caching, which was later used by Kantere et al. [KDF+11] to investigate optimal service pricing for cloud caches.

Chau et al. [CWC14] discuss to what extent digital services, such as video streaming or even network access, can benefit from Paris Metro Pricing (PMP), a form of pricing which has *n* service classes with different prices and different levels of congestion. They state that in order for PMP to be applicable, services should not use multiplexing. Furthermore, different classes should generate a monotone preference perceived by the customer. While this generally supports the idea that different service classes are important, it has been applied to delivery of services where congestion plays a role, which is not investigated in this work.

6.2. Focus of this Part

It has been shown that online retailers can employ much more fine-grained pricing models compared to offline alternatives. However, there is little guidance for firms [HHS11]. This lack of guidance has also been identified as an issue for data marketplaces [BHS11]. Despite initial attempts to solve this [BHS11; KUB+12a; KUB+12b; TSBV13; TWB+13; TASB14], the challenge of lacking guidance remains, particularly if multiple data providers are concerned [BHK+13].

Furthermore, to date, there is no sense of value for data [Mil12b]. Both research groups addressing data marketplaces and pricing – those of Balazinska et al. and Tang et al. –, while providing technical means to model prices, do not address the more pressing question of how to price the data initially. It is true that their models already provide guidance on how to price data, as far as arbitrage is concerned. However, the pricing challenge that data has no value *per se* remains. They suppose that the sellers have an idea what their data is worth to the consumer. Hence, this work will particularly focus on a pricing scheme that supports data providers in setting appropriate prices which will also incorporate the idea that data sets are not of equal value.

Building on the information good classification, it should be mentioned that the information good to be studied in this part will be plain structured data, this is durable and consumable E-Information including event goods as defined in Section 5.4.2. This data will be provided through a data marketplace infrastructure where a data marketplace will be an electronic platform that allows for the exchange of data as defined in Section 5.3.1. It is further supposed that this data marketplace is an independent data marketplace fitting the framework presented in Section 5.2.2 in order to exclude any potential bias. While the pricing of services provided through such a platform is also interesting, the focus here will be on the pricing of the good data, as the relationship of data marketplace to data provider can be described as service provider and user which has been studied in other contexts. In contrast to this, data pricing has gained little recognition and has therefore be chosen as subject of this thesis. After all, data marketplaces as platforms are a very specific type of E-Bargaining platform which have been extensively investigated.

Two studies have suggested that prices should be modelled in such a way that they are attached to cells or tuples and compute the final price when the data is queried [BHS11; TWB+13]. In contrast, here an approach will be pursued that does not require sellers to have an idea of the value of their goods. This will be done by adopting the NYOP idea of Tang et al. [TASB14; TSBV13] and combining it with a hidden threshold and *Goldilocks Pricing* as described by

Shapiro and Varian [SV99]. This means data sets – views to be more precise – will be offered at an exorbitant price. Then customers can name the price they are willing to pay. If it is greater than an undisclosed threshold price, customers get the full quality product at their suggested price. If it is smaller, however, the view – or version for that matter – is transparently adjusted extemporaneously to meet the customers price. This is particularly promising if customers have a very heterogeneous willingness to pay. While basically similar to previous studies by [TSBV13; TASB14] here more than one quality dimension will be looked at. This is in line with the observation that the time for dynamic pricing has come which shall compute the right price at the right time [NRRS05].

In order to make use of versioning of products, which has been established to be beneficial [SF08; SV99], some assumptions are made to satisfy the prerequisites for versioning identified in [Rei02; NRRS05]:

1. Customers are heterogeneous in their willingness to pay.
2. Customers are identifiable.
3. Customers are not allowed to resell products.

Throughout the discussion of pricing models in Section 5.4.2 a number of factors have been introduced that allow for versioning of information goods. Since this work focuses on data, only those factors identified as applicable to data will be looked at. These are *Extent, Recentness, Quality, Delay,* and *Comprehensiveness.* Given that these were not mutually exclusive, *Extent* can be mapped to *Comprehensiveness* and *Recentness* to *Delay* and all can be seen as subordinate to *Quality.* This is illustrated in Figure 6.1.

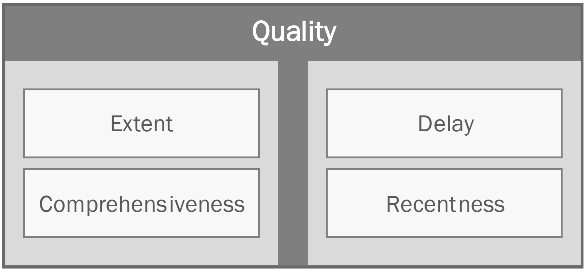


Figure 6.1.: Overview of Relevant Versioning Criteria.

Given that *Quality* is also data-inherent it builds a perfect starting point for versioning. Furthermore, it also allows for an objective value comparison[1] of two data sets that have similar content. To this end, an approach suggested by Stahl et al. [SLV15] is pursued further. The authors suggested modelling the utility U of a data set as follows:

$$\mathrm{U} = \sum_{i=1}^{n} w_i e_i + \sum_{j=1}^{m} v_j d_j$$

Here, U depends on 1 to n economic and licensing-related factors (e_i) and on 1 to m data-related factors (d_j), each with a specific weight w_i or v_j, respectively. In the original publication, [SLV15], e_i, d_j were supposed to be functions of a number of input variables x. However, this seems not to add any benefit at this point. The aim of this part of the work can be stated as follows:

> *Provide a fair pricing scheme to be utilised by data providers on data marketplaces that allows for quality-based versioning and according price discrimination for custom-tailored relational data goods.*

More precisely, a reversed pricing mechanism is proposed that builds on the idea of NYOP pricing incorporating quality as a versioning factor, incorporating a possibility for users to express their preferences for certain quality criteria in order to receive a custom-tailored data product. To this end, a framework will be presented that is capable of deriving a data quality score for data traded on a data marketplace adjusted with users' preferences. Given that data quality is inherent to all data the method can be used domain-independently and data quality criteria have only to be adjusted with domain-specific weights.

6.3. A Relational View on Data Marketplaces

Data marketplaces host data for a number of providers who sell tabular, i. e., relational, data. This section will define data marketplaces and data providers based on the formalisms developed in Section 2.2. The focus on relational data is necessary in order to define precisely the scope of this work. However, the model developed herein may be extended to other types of data in the future.

Regarding providers, it is supposed that they sell data in a tabular format with given column names or attributes. This data can be described as a relation r with

[1] Supposing that quality is *the* value-bearing factor for customers.

n_a unique attributes A_i of domain $dom(A_i), 1 \leq i \leq n_a$. The set of attributes is denoted as: $X = \{A_1, \ldots, A_{n_a}\}$. Consequently, the data (relation) may be described as an instance r of the relational schema $R = (X, \cdot)$.

Most of the time, data providers will not only sell one relation but many of them. Let a provider sell m_r relation instances r_j, then, one provider can be considered providing a database instance $d = \{r_1, \ldots, r_{m_r}\}$ with the according schema $D = (\boldsymbol{R}, \cdot)$, with $\boldsymbol{R} = \{R_1, \ldots, R_{m_r}\}$.

While, in some cases, it might be appropriate to view an entire data marketplace as a database, in this work every provider is supposed to provide an individual database instance. This is a practical presumption based on the assumption that it would be very difficult to enforce a common schema, including *inter-relational* constrains, across providers. Furthermore, this would require data vendors to adapt their data to the schema of any data marketplace they are selling on, which seems an unnecessary burden. Moreover, this thesis considers data pricing from a vendor's point of view, thus, there is no benefit in viewing a data marketplace as a database.

Even viewing data offerings as databases can be complicated as customers will often require data from different relations, which then have to be joined upon request. To simplify this and to add clarity, in this work, data providers' offerings will be treated as a *universal relation* (u), a tool which has a long history in database theory. It has for instance been described in [MUV84; Vos86]. For the purpose of this work, a universal relation is created by joining all $r_j \in d$ in such a way that no data is lost, using a full outer join. It can be argued that joining could be done only when necessary – an approach that might be followed in an implementation; however, using only the universal relation has the advantage that no further joins are necessary during the formal elaborations in the remainder of this chapter, which improves understandability. Furthermore, any original relation r_j may be arrived at by appropriate selections and projections over u. Formally, the universal relation u over a database instance d can be defined as:

$$u = r_1 ⟗ \ldots ⟗ r_{m_r}$$

It should be noted that this approach requires attributes to be unique within each single database. However, this is a minor technicality and can be achieved by renaming affected attributes[2].

[2] The process of renaming is formally defined, for instance, in [Vos08].

Similar to a general relation schema R and a database schema D, the universal relation schema is defined as:

$$U = (X_U, \cdot) \text{ with } X_U = \bigcup_{j=1}^{m_r} X_j$$

As a next step, a data marketplace comprising k_d data providers shall be defined. A data marketplace is mainly characterised by the offerings of data providers. Thus, for the purpose of this work, it formally consists of the set of universal relations by different providers $\boldsymbol{u} = \{u_1, \ldots, u_{k_d}\}$ and the set of according schemas $\boldsymbol{U} = \{U_1, \ldots, U_{k_d}\}$. Furthermore, a data marketplace is likely to impose some technical and administrative restrictions Σ_M, such as supported data types, possible access methods, storage capacity, and supported billing schemes, to name but a few. In concordance with previous elaborations, this explicitly does not include any constraints regarding the question of how data of one provider has to relate to another provider's. The reason for this is that providers are treated as independent actors. To summarise, a data marketplace M is specified as follows:

$$M = (\boldsymbol{u}, \boldsymbol{U}, \Sigma_M)$$

Given that the universal relations are derived from databases, an equivalent formulation may use the set of databases $\boldsymbol{d} = \{d_1, \ldots, d_{k_d}\}$ and the set of database schemas $\boldsymbol{D} = \{D_1, \ldots, D_{k_d}\}$:

$$M = (\boldsymbol{d}, \boldsymbol{D}, \Sigma_M)$$

Having defined data marketplaces and the offerings of data providers, it is further necessary to define operations that are supported on offered universal relations by relational algebra. For the purpose of this work, only basic set operations (i. e., union, intersection, difference) and the basic relational algebra operations presented in Section 2.2.2, namely, projection, selection, natural join, and full outer join, are allowed when formulating queries. This has the advantage that the computational complexity of queries is $\mathcal{O}(n \log n)$ at the most, for selections even only $\mathcal{O}(n)$ [Vos08]. This presumption has the advantage that, when calculating prices, the time for collocating the data is negligible. Notwithstanding this, extending the allowed operations is an interesting field of future research.

6.4. Data Quality

In previous chapters, it has been shown that data, as basis of information, has a value that is highly dependent on the context as well as on potential buyers. Furthermore, it has been shown that it is sensible to create different versions of information goods in order to tap the willingness to pay of heterogeneous customers. Out of a number of factors that potentially allow for price discrimination, data quality has been identified as promising. In particular the fact that first steps in this direction were taken in [TSBV13; TASB14] supports this. Nevertheless, studies by Tang et al. [TSBV13; TASB14] only use one data quality dimension each – accuracy for relational data [TSBV13] and completeness for XML data [TASB14]. Given that there are more criteria in the context of data quality that are relevant for customers and consequently allow for price discrimination, the approaches described in [TSBV13; TASB14] can be considered somewhat limited. Furthermore, data quality, if expressed in a meaningful score, allows the comparison of two offers from different providers. However, it is intuitively clear that considering more than one quality dimension complicates the pricing approach.

This section will first give an overview of data quality criteria, which are applicable in the context of data marketplaces and data trading. Next, a quality scoring model is derived that allows for a comparison of different offerings in Section 6.4.2. Subsequently, Section 6.4.3 will discuss some measures that capture how well individual criteria are fulfilled. This will then be applied to an example in Section 6.4.4. Finally, more quality scores are introduced to underpin the general applicability; the quality measures and scores of this section also build the basis for automated versioning of relational data products, discussed in Section 6.5.

6.4.1. Data Quality Dimensions

Many works regarding data quality and how to measure it have been published over the last decades. Focusing on data quality in a Web context – precisely the context of most data marketplaces and data providers –, Naumann [Nau02] aggregated several works on data quality, namely [Bas90; Red96; WS96; JV97; CZW98; Wei99], into four sets of quality criteria. Therefore, the following elaboration builds on Naumann [Nau02], rather than on the individual works.

The criteria sets are a) *content-related,* i. e., directly rooted in the data; b) *technical,* i. e., related to the organisation and delivery of the data; c) *intellectual,* i. e., related to the knowledge of eventual users; and d) *instantiation-related* i. e., related to the presentation of the data. In the following a brief description of each measure is given along with an elaboration of how it is relevant in the context of data marketplaces, i. e., whether it a) can be used for versioning, b) is automatically calculable, c) can be used for an automated intra-marketplace comparison, and d) can be used for an automated inter-marketplace comparison.

The following content-related criteria are mentioned in [Nau02]:

- *Accuracy,* the percentage of correct values in the data set
- *Completeness,* the percentage of non-null values in the data set
- *Customer Support,* the amount and usefulness of available human help
- *Documentation,* the extent of available meta data regarding the data sets
- *Interpretability,* the match between a user's technical understanding and the data
- *Relevancy,* the degree to which the data satisfies a user's information needs
- *Value-added,* the value the use of the data provides to its users

As previously argued, the value of data is highly customer-dependent. The same is true for the *value-added* by using a data set. It is the aim of this work to approximate this through the other quality dimensions. Hence, it will not be further analysed to avoid recursion. Most of the other criteria cannot be calculated fully computationally, the exemption being *completeness*. All others require knowledge that goes beyond the actual data but most of them can be calculated at least partially automatically. Regarding *customer support* and *documentation*, the existence and the extent can be evaluated. While this is a first step that allows for versioning and comparison, it does not say anything about the actual quality. Other criteria may only be evaluated given external data. For instance, *accuracy* can be compared to verified accurate data. Moreover, some criteria remain that cannot be automatically examined, namely *interpretability* and *relevancy*, as both require an in-depth understanding of users, which cannot be achieved in an automated way. Consequently, they cannot be used for comparisons. All other criteria in the content-related group can be used for an inter-marketplace comparison as well as for an intra-marketplace comparison as far as they can be automatically assessed.

For *completeness* – which can be considered one form of *extent* identified in Section 6.2 – computational evaluation is possible and differently complete versions can be created easily. At this point, it should be mentioned that in particular regarding completeness there are two different assumptions, *Closed World Assumption (CWA)* and *Open World Assumption (OWA)* [BS06]. In the CWA case, it is supposed that only the values available in a relation represent facts about the real world, i. e., if a value is missing, the data is incomplete. In contrast, under the OWA, for a null value, one cannot state if a missing value is really missing or whether it simply does not exist [BS06]. While the CWA has the limitation that it restricts the model, it has the advantage that all information about the data is already contained within the data. As argued in Section 2.2, for the purpose of this work, it is not particularly relevant why a value is missing. The fact is, it cannot be delivered to the customer. Hence, the Closed World Assumption will be made. Having defined the CWA, which is commonly used for completeness only [BS06], it will be used as general assumption in this work. This means, it is always supposed that all necessary information for pricing is available on the data marketplace(s) under investigation. As stated above, this abstracts from reality but allows for simpler modelling of a rather complex construct.

Thus, the CWA has also an effect on *accuracy*. In this work – similar to Tang et al. [TSBV13], who suppose that the available *accuracy* is worth the full price – it is supposed that the accuracy is data inherent. This leads to the conclusion that it can be used for versioning. For comparison between different offerings, in contrast, *accuracy* has only limited applicability. More generally, all measures that can only be partially assessed automatically have also only limited applicability to intra-marketplace and inter-marketplace comparison.

Naumann [Nau02] mentions the following technical criteria:

- *Availability,* the probability that a query is answered within a given time period
- *Latency,* the time between issuing a query and receiving its first response
- *Price,* the amount of money a user has to pay for the data
- *Quality of Service,* the error rate when transmitting (mainly relevant in streaming)
- *Response Time,* the time between issuing a query and receiving its full response
- *Security,* the degree of protection through encryption and anonymisation
- *Timeliness,* the freshness of the data

Following the argument of value-added, price will be excluded. The remaining criteria can all be influenced by providers in the same way *accuracy* or *completeness* can, i. e., while there is a (technical) upper limit, they can be lowered. Thus, they all can be used for versioning. Furthermore, an automated calculation is also possible for most of them. The exemption is *quality of service* for which it has to be defined what quality specifically means. Thus, it is overall not very precise and has, therefore, been excluded from further examination. Moreover, *availability* requires multiple measurements over time in order to be properly evaluated. *Timeliness* can be considered equivalent to *recentness* introduced in Section 6.2. Supposing that all data is delivered by the same infrastructure, it is reasonable to argue that the technical criteria are rather irrelevant for intra-marketplace comparisons. However, if they are used for versioning, i. e., used to purposefully differentiate products, in a non-random way, they can also be used for an intra-marketplace comparison. For instance, a data provider could discriminate customers based on their preference for security and ask a premium price for encrypted transmission. Nevertheless, the marketplace operator has to provide the possibility for doing this to the actual provider, which in turn raises the question of pricing for this possibility. However, this thesis focuses on data providers and their customers. In conclusion, all measures discussed in this group are partially applicable in an intra-marketplace comparison and all are highly important in an inter-marketplace comparison as these are likely to run on different infrastructures.

Next, the intellectual criteria shall be outlined:

- *Believability,* the expected accuracy
- *Objectivity,* the degree to which the data is free of any bias
- *Reputation,* the degree of high standing of the source perceived by customers

All of these criteria are value drivers; however, all but *objectivity* cannot be assessed automatically without any further user input because they are inherently dependent on the users. *Objectivity*, in contrast, could be partially automatically calculated if there are technical means to verify the data (supposing verifiable data is also objective). Regarding versioning, none of the criteria in this group can be used as it is very difficult to influence them in the short run because they are perceived by the user, rather than actively created. *Reputation* and *objectivity* may both be used for intra-marketplace comparison to some degree. The first requires some infrastructure on the data marketplace to measure the reputation such as a rating system for customers on the platform. *Objectivity* could be used as a measurement if the requirements for an automated assessment are fulfilled. In this case it could even be used for an inter-marketplace comparison, which seems very unlikely for the other two because it is hard to measure *reputation* objectively automatically across platforms. For *believability* the argument can be made that it is inherently unmeasurable given its subjectivity.

Finally, the group of instantiation-related criteria will be discussed:

- *Amount of Data,* the number of bytes returned as a query result

- *Representational Conciseness,* how well the representation matches the data
- *Representational Consistency,* how well the representation matches previous representations of the same data
- *Understandability,* the degree to which a data set can be understood by a user
- *Verifiability,* the degree to which a data set can be checked and verified

While the *amount of data*, which also can be seen as manifestation of extent, as described in Section 6.2, and the *representational consistency* can be assessed automatically, the other three cannot. *Representational conciseness* and *understandability* cannot be assessed automatically because only humans can judge whether the data format matches the data or whether they understand the data. *Verifiability* depends very much on the actual use-case and is hard to generalise. Thus, it has been categorised as not automatically assessable. The *amount of data* can be used for versioning as well as for intra- and inter-marketplace comparison. The *representational consistency* can also be used for both types of comparison. While it is technically possible to change the representation (access API or data format), it seems inappropriate to do so just to lower the quality of the product. Nevertheless, different versions could be different guarantee levels that the representation does not change over certain time intervals. Thus, it has been categorised as applicable to versioning. Moreover, it is suitable for automated assessment but requires multiple measurements, similar to *availability*. The *representational conciseness* cannot be automatically assessed and is, thus, not suitable for either comparison mode. Nevertheless, it can be used for versioning, in that there could be a low quality version that simply stores all data in one *binary large object* and a high-quality version in which the data is organised in an appropriate relational structure.

An overview of all criteria, their applicability for versioning, their automated computability, and their applicability for intra- and inter-marketplace evaluation is presented in Table 6.1, where criteria that are applicable without restrictions are annotated with ✔, those that have a limited applicability are annotated with ✓/✖, and those not applicable at all are annotated with ✖.

Having described the most relevant data quality criteria as well as evaluated them with regards to versioning, automated evaluation, and usability for comparison between data provider on the same or on different data marketplaces, they shall now be grouped into different sets for easy future reference.

Table 6.1.: Overview of Quality Criteria.

Category	IQ Criterion	Versioning	Automated	Intra DM	Inter DM
Content-related	Accuracy	✓	✓/✗	✓/✗	✓/✗
	Completeness	✓	✓	✓	✓
	Customer Support	✓	✓/✗	✓/✗	✓/✗
	Documentation	✓	✓/✗	✓/✗	✓/✗
	Interpretability	✗	✗	✗	✗
	Relevancy	✗	✗	✗	✗
Technical	Availability	✓	✓	✓/✗	✓
	Latency	✓	✓	✓/✗	✓
	Response Time	✓	✓	✓/✗	✓
	Security	✓	✓	✓/✗	✓
	Timeliness	✓	✓	✓/✗	✓
Intellectual	Believability	✗	✗	✗	✗
	Objectivity	✗	✓/✗	✓/✗	✓/✗
	Reputation	✗	✗	✓/✗	✗
Instantiation-related	Amount of Data	✓	✓	✓	✓
	Representational Conciseness	✓	✗	✗	✗
	Representational Consistency	✓	✓	✓	✓
	Understandability	✗	✗	✗	✗
	Verifiability	✗	✗	✗	✗

Firstly, quality criteria can be trivially categorised by whether they can be assessed automatically A, manually M, or whether they are hybrids H. Resulting in the following sets:

$$\begin{aligned}
\mathsf{A} &= \{\textit{Amount of Data, Availability, Completeness, Latency, Representational}\\
&\qquad \textit{Consistency, Response Time, Security, Timeliness}\}\\
\mathsf{M} &= \{\textit{Believability, Interpretability, Relevancy, Representational Conciseness,}\\
&\qquad \textit{Reputation, Understandability, Verifiability}\}\\
\mathsf{H} &= \{\textit{Accuracy, Customer Support, Documentation, Objectivity}\}
\end{aligned}$$

In order to be able to address all quality criteria, the set $\mathsf{Q_a}$ is defined as union of the previous:

$$\mathsf{Q_a} = \mathsf{A} \cup \mathsf{M} \cup \mathsf{H} \tag{6.1}$$

Regarding pricing and versioning, all criteria that allow for the dynamic creation of a large number of versions build the set V. As a coincidence, most of them can be computed automatically and allow for inter-marketplace comparison as well as for intra-marketplace comparison, supposing the data marketplace operator provides the infrastructure for these technical criteria. Given that this thesis models an ideal marketplace it is reasonable to make the assumption that this infrastructure is provided. There is one exemption to this rule, namely *accuracy*. While allowing the creation of numerous versions, it cannot be (fully) automatically computed and, thus, is also limited in its applicability for comparison. The overall set V comprises the following attributes:

$$\begin{aligned}\mathsf{V} = \{&\textit{Accuracy, Amount of Data, Availability, Completeness, Latency}\\ &\textit{Response Time, Timeliness}\}\end{aligned}$$

Furthermore there are criteria that generally allow for versioning but where the number of versions is strongly limited. For instance, it is not sensible to create a large number of *customer support* tiers. Sticking with the Goldilocks principle, it seems reasonable to provide three categories for all attributes in this set, which will be referred to as G. All of the criteria in this set but *representational consistency* and *security* are limited in their applicability as comparison criterion.

$$\begin{aligned}\mathsf{G} = \{&\textit{Customer Support, Documentation, Representational Conciseness,}\\ &\textit{Representational Consistency, Security}\}\end{aligned}$$

As these are the only relevant quality criteria regarding versioning, they are combined in $\mathsf{Q_v}$ for which $\mathsf{Q_v} \subset \mathsf{Q_a}$ holds:

$$\mathsf{Q_v} = \mathsf{V} \cup \mathsf{G} \tag{6.2}$$

The last set X consists of quality criteria that require so much manual input or user knowledge that they neither allow for versioning, automated assessment nor for comparison of different offers. Thus, they are not really applicable in this context. Furthermore, this set contains two criteria that cannot be used for versioning but do allow for a (limited) comparison of offerings by different providers or on different data marketplaces. These are *objectivity* and *reputation*. While it is hard to measure objectivity in a way that allows for it to be used in pricing, reputation can – as argued above – be calculated on a data marketplace and, thus, in principal be used in pricing.

$$\begin{aligned}\mathsf{X} = \{&\textit{Believability, Interpretability, Relevancy, Reputation, Objectivity,}\\ &\textit{Understandability, Verifiability}\}\end{aligned}$$

6.4.2. A Data Quality Score

Having introduced and categorised different quality criteria in the previous section, this section presents an overall quality score and applies it to relative pricing. Inspired by Baetge et al. [BKW13], who suggested using a scoring model to evaluate professional football players – like data it is difficult to find an objective value for professional sportsmen and women –, here, it is proposed to approach data pricing utilising a scoring model for data quality. In the context of quality-driven query planning a similar approach was proposed by Naumann [Nau02].

Given the three sets of criteria: A, computationally assessable; M, manually assessable; and H, hybrid criteria that can partially be computed but can significantly be enhanced by an additional manual assessment, let a_i, m_i, and h_i denote the according scoring values for the respective criteria. Moreover, let a_i, m_i, and h_i be in the interval $[0, 1]$ to allow for comparison. Then, a simple overall score can be defined as:

$$QS = \frac{1}{|\mathsf{Q_a}|} \sum_{a_i \in \mathsf{A}} a_i + \sum_{m_j \in \mathsf{M}} m_j + \sum_{h_k \in \mathsf{H}} h_k \tag{6.3}$$

As such, this model is very generic and not very expressive as different criteria are of different importance. Therefore, it is very likely that users have different preferences for different attributes; consequently for each set of measures, a weighting vector $W = (w_1, \ldots, w_{n_q}), w_i \in \mathbb{R}_0^+$, is introduced, where $n_q = |\mathsf{Q_a}|$ denotes the number of elements in $\mathsf{Q_a}$. As in [Nau02], it is requested that:

$$\forall q_i \in \mathsf{Q_a} \exists! w_i \in W \text{ and } \sum_{w_i \in W} w_i = 1 \tag{6.4}$$

Using the formalisms of Equation 6.1 and Equation 6.3:

$$\sum_{w_i \in W_\mathsf{A}} w_i + \sum_{w_j \in W_\mathsf{M}} w_j + \sum_{w_k \in W_\mathsf{H}} w_k = 1 \tag{6.5}$$

Several methods exist to ask users for their preferences. Naumann [Nau02] mentions, direct specification, pair wise comparison and an eigenvector model. Here, a direct specification is suggested, by means of a slider-based GUI. As an example four criteria have been depicted in Figure 6.2. However, in order to get meaningful results, all criteria have to be used. If this method is implemented correctly, all sliders should start in the middle suggesting equal weights. Once one slider is moved all others react accordingly to ensure that the sum of all weight equals 1. In this way, no inconsistencies can occur and users receive direct visual feedback of their preferences.

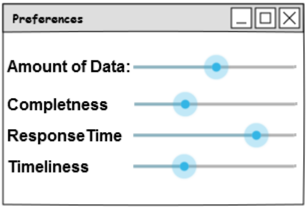


Figure 6.2.: Example GUI: Asking Users for Their Preferences.

Based on this, the overall score QS can be defined as:

$$QS = \sum_{a_i \in \mathsf{A}} w_{\mathsf{A}i} a_i + \sum_{m_j \in \mathsf{M}} w_{\mathsf{M}j} m_j + \sum_{h_k \in \mathsf{H}} w_{\mathsf{H}k} h_k \tag{6.6}$$

On a less abstract level, QS determines the overall quality of a data set. It is obvious that a data quality score is not a price. However, it can be seen as relative value when comparing offers of different vendors or when data is to be traded for data. Moreover, given a price P for a data set, the quality score QS can be used to give customers guidance regarding the *quality for money* (QM) they receive by calculating:

$$QM = \frac{QS}{P} \tag{6.7}$$

In summary the quality score can help customers to choose a provider and providers may benefit as it makes evident how their offer preforms compared to the competition's. As a result, they may adapt prices or quality to improve their competitiveness. Furthermore, with additional knowledge gathered on a data marketplace, it would be possible to use this score as price indication, i. e., learning from customers' choices which data with what quality is sold at what price. This can further improve over time.

6.4.3. Basic Data Quality Measures

This section introduces three quality criteria in A, namely *Amount of Data, Completeness,* and *Timeliness*, as illustrative examples of how data quality can be measured. These will then be applied to an example in the next subsection. Subsequently, Section 6.4.5 will introduce more advanced measures to complete the picture. As argued in Section 6.3, quality criteria will be defined using the universal relation u of each provider in order to allow for a better overview and comparability. For the same reasons, all quality criteria are scaled in the interval [0,1], with 0 being the worst and 1 being the highest score.

The *amount of data*, should measure the size of the data in bytes [Nau02]. While this definition has some expressiveness, it is difficult to apply the measure in versioning and even when comparing two offerings, as providers might store the very same data at different compression rates. Comparing two offerings of relational data, the *amount of data* can be measured by calculating the proportion of selected rows or columns compared to the maximum available between providers. As auxiliary means, let Y be a subset of X_u and let $A_i \Theta_i a_i$ be a selection with $A_i \in X_u$, $a_i \in dom(A_i)$, $\Theta_i \in \{<, \leq, >, \geq, =, \neq\}$. Then, a number of n_σ selection criteria, combined to the condition C, can be denoted as $C = \wedge_{A_i \Theta a_i}, 1 \leq i \leq n_\sigma$.

While most scores will consider each provider individually, others need more than one provider to be meaningfully calculated. For example, the *amount of data* for one provider is more meaningful if it is compared to the *amount of data* of other providers. Therefore, it is supposed that k_p providers are to be compared, the index l is used to refer to the individual offerings as u_l. However, such measures are rather the exemption than the rule; if only one provider is concerned, in the following, the index will be omitted to help readability.

Regarding *amount of data*, the offering that contains the largest response to the query is denoted u_l^*. To allow for comparability, it is further supposed that the views under consideration have the same schema. For the universal relation, this implies the same schema $U = (X_u, \cdot)$. Now, the score for *amount of data — AoD* — can be defined as:

$$AoD(u_l) := \frac{|\pi_Y(\sigma_C(u_l))|}{|\pi_Y(\sigma_C(u_l^*))|} \tag{6.8}$$

In the context of versioning, it is sensible to split this measure into two: one that measures the number of selected tuples (rows) — *AoR* — and one that measures the number of attributes (columns) — *AoC* —, each in relation to the total number of tuples and attributes respectively. This leads to the following, simpler, notations:

$$AoR(u_l) := \frac{|\sigma_C(u_l)|}{|u_l^*|} \tag{6.9}$$

$$AoC(u_l) := \frac{|Y|}{|X_{u_l}^*|} \tag{6.10}$$

Completeness measures the amount of content actually available in the relation to be sold, compared to the maximum amount of data possible. As previously argued, this is done under CWA. While this might seems to be a restriction, it is practically impossible to determine the actual completeness of most of the data traded on data marketplaces under OWA. In contrast, calculating an internal completeness makes data sets comparable. In fact in this work *completeness* is rather a *null-freeness* score, nevertheless, the term *completeness* is used for consistency. To measure *completeness*, an auxiliary function is defined, similar to [HK08], that determines if a value v is null:

$$nv(v) := \begin{cases} 0 \text{ if } v \neq \perp \\ 1 \text{ if } v = \perp \end{cases} \tag{6.11}$$

Based on this, the overall number of null values can be calculated as:

$$n(u) = \sum_{\mu \in u} \sum_{A_i \in X_u} nv(\mu[A_i]) \tag{6.12}$$

Now, *completeness* can be calculated:

$$c(u) = \frac{|u| \times |X_u| - n(u)}{|u| \times |X_u|} \tag{6.13}$$

Alternatively, this may be written as:

$$c(u) = \frac{|\{\mu[A], \mu \in u, A \in X_u | \mu[A] \neq \perp\}|}{|u| \times |X_u|} \tag{6.14}$$

Measures for *timeliness* have been described for instance in [BS06; BWPT98]. According to these sources, *timeliness*, i. e., the freshness of data, depends on a number of characteristics: a) delivery time, i. e., the time at which the datum is being delivered; b) input time, i. e., the time at which the datum was entered into the system; c) age, i. e., the age of the datum when entered into the system; and d) volatility, i. e., the typical time period a datum keeps its validity. Volatility can be interpreted as change frequency, analysed in [CG03]. Based on these fundamental indicators, Batini and Scannapieca [BS06] define the currency of data as $Age + (DeliveryTime - InputTime)$, which shows how up-to-date a datum is. In this thesis, it will be abstracted from age, as it is supposed that time-sensitive data is entered into the system immediately. Furthermore, in most cases it is only relevant when a datum was last updated and how long it remains typically valid. Since volatility and time have the same unit, the overall value is dimensionless. Adopting the definition of Batini and Scannapieca [BS06], the *timeliness* of a record t_μ is a function of delivery time, input time and volatility. Given that delivery time is a constant that does not depend on the query result, only the latter two are independent variables:

$$t_\mu(\mathit{InputTime, Volatility}) = max\left\{0, 1 - \frac{\mathit{DeliveryTime{-}InputTime}}{\mathit{Volatility}}\right\} \tag{6.15}$$

In order to make *timeliness* indeed measurable, it is supposed that a *LastUpdated* attribute exists for all m_r relations r_j that have been joined in u of a provider. This implies that $\mathit{LastUpdated} \in X_{r_j}, 1 \leq j \leq m_r$. Furthermore, it is supposed a volatility constant v_{r_j} exists for each relation r_j. For this measure, arguably only the timeliness of the most time critical relation r_j is relevant. In the case that more than one timeliness values from original sources are important multiple timeliness scores should be calculated. The smallest volatility value, i. e., the only relevant value, is denoted v^*. The according *LastUpdated* field is indicated as *LastUpdated**. Then, the overall timeliness score can be calculated as average timeliness for all tuples in u using Equation 6.15:

$$tim(u) = \frac{\sum\limits_{\mu \in u} t_\mu(\mu[LastUpdated^*], v^*)}{|u|} \tag{6.16}$$

At this point, it should be mentioned that the scores, even though standardised to be in the range between 0 and 1, are not really comparable because some scores might be biased towards one edge, e. g., *completeness* is likely to be closer to 1 than to 0, most of the time. To solve this issue, each measure can be standardised individually before calculating the overall score [Nau02] if it is supposed that two or more alternatives are to be compared. However, a quality score is rather pointless if it is not used to compare different offerings. Naumann [Nau02] states that there is no ideal method of standardisation as methods either have the disadvantage of unequal ranges, as just indicated, or they do not scale proportionally, i. e., if measure a was twice as high as measure b before, it is not necessarily after the standardisation. In the context of this work, it is more important to obtain comparable measures than to realise proportional scaling. Thus, a transformation will be applied, suggested in [Nau02], that enables sores to be exactly in the interval $[0,1]$ at the expense of not scaling proportionally. To this end, it is supposed that there are $1 \leq i \leq n_q$ quality criteria and $1 \leq j \leq k_p$ data providers to be compared. Then, d_{ij} denotes the untransformed quality score i for provider j. Furthermore, $d_i^{\min}$ denotes the minimal value for criterion i measured over all providers and $d_i^{\max}$ the maximum value. The normalised score v_{ij} can be calculated as follows for positive scores (more is better):

$$v_{ij} = \frac{d_{ij} - d_i^{\min}}{d_i^{\max} - d_i^{\min}}$$

and for negative scores:

$$v_{ij} = \frac{d_i^{\max} - d_{ij}}{d_i^{\max} - d_i^{\min}}$$

Now, the overall quality score for different criteria can be calculated based on v_{ij} rather than d_{ij}.

6.4.4. A Quality Score Example

Having introduced a data quality model in the previous section, it shall now be applied to a sample case. As an example, two providers of weather forecast data are considered. Weather data has a number of characteristics that makes it particularly interesting. For instance, it is relevant that weather forecast data is a consumable information good which allows for considering timeliness when pricing because most of the time weather data is only relevant for future dates. To create this data, the weather is constantly observed and different data are collected using a number of weather stations. These raw weather data are then used to forecast the weather for days to come. More precisely particular attributes of the weather, such as temperature, are forecast. Thus, it is supposed that weather data providers A and B both provide past, current, and forecast weather data, i. e., providers constantly fill their database with new data as well as update forecast data which becomes more precise the closer the forecast date comes.

Given that sensors and forecasting models are not perfect, it is further supposed that the data is not complete as some data are lost because of malfunctions. In the following provider A uses very reliable sensors but fewer, which results in more complete but less extensive data. Nevertheless, because of the better sensors, provider A can forecast three days, which provider B cannot. In contrast, provider B collects more data (more attributes) using less reliable sensors and has more weather stations. Moreover, both providers collect similar but not identical data. Provider A offers data for *AirPresure* in hPa, *Humidity* in percent, *Temperature* in degree centigrade, and *Precipitation* (rainfall) in mm. Provider B offers the same as provider A without *Precipitation* and additionally *WindSpeed* in km/h and *Cloudage* in percent. The data sets of both providers also include the date and station for which the weather is forecast (or has been recorded), as well as when the data were last updated.

In this sample scenario, a customer, suppose an airline, wants to buy weather forecast data at 5 pm (17:00) on 7th May 2017 for the next three days from three different airports (`FRA, LHR, AMS`). The assumption is made that the volatility of weather forecast data is 24 hours. The relevant data sets of providers A and B are depicted in Table 6.2 and Table 6.3, respectively. To provide a realistic example, the data has been randomly sampled from Weather Underground[3]. For reasons of clarity and comprehensibility only the relevant view on u is depicted. Nevertheless, the quality score could be applied to all of u. After the customer has submitted their query, the data market calculates the possible result

[3] http://www.wunderground.com/, accessed: 2015-05-31.

sets for providers A (Table 6.2) and B (Table 6.3) and applies the quality score calculation with user supplied weights. For simplicity reasons, only some measures in A will be demonstrated, namely *Amount of Data, Completeness,* and *Timeliness* have been chosen, as illustrative examples. Nevertheless, other measures can be applied in very much the same way.

Table 6.2.: Relation u_A for Provider A.

Station	AirPressure	Humidity	Temperature	Precipitation	Date	LastUpdated
FRA	⊥	52	17	0	2017-05-08	14:00
FRA	1020	43	19	0	2017-05-09	15:00
FRA	1005	40	15	41	2017-05-10	16:00
LHR	1025	69	16	17	2017-05-08	14:00
LHR	1008	⊥	14	85	2017-05-09	15:00
LHR	1003	70	12	70	2017-05-10	16:00

Table 6.3.: Relation u_B for Provider B.

Station	AirPressure	Humidity	Temperature	WindSpeed	Cloudage	Date	Last Update
FRA	1022	⊥	18	8	70	2017-05-08	14:00
FRA	⊥	50	20	⊥	25	2017-05-09	14:00
LHR	1015	⊥	⊥	23	⊥	2017-05-08	13:00
LHR	1004	79	13	⊥	93	2017-05-09	13:00
AMS	⊥	59	13	16	⊥	2017-05-08	12:00
AMS	1002	82	12	23	97	2017-05-09	12:00

Supposing that all weights are equal, i. e., $w_i = \frac{1}{3}$, the overall quality scores can be calculated as:

$$QS = \frac{1}{3}AoD + \frac{1}{3}c(u) + \frac{1}{3}tim(u)$$

Plugging in the respective formulas results in:

$$QS = \frac{1}{3}\frac{|\pi_Y(\sigma_C(u))|}{|\pi_Y(\sigma_C(u_j^*))|} + \frac{1}{3}\frac{|\{\mu[A], \mu \in u, A \in X_u | \mu[A] \neq \perp\}|}{|u| * |X_u|} + \frac{1}{3}\frac{\sum_{\mu \in u} t(\mu[LastUpdated^*], v^*)}{|u|}$$

The actual values for each score as well as the overall score for $w_i = \frac{1}{3}$ for all i and an alternative weighting ($w_1 = \frac{3}{20}, w_2 = \frac{1}{2}, w_3 = \frac{7}{20}$) are presented in Table 6.4.

Table 6.4.: Quality Score Results for Providers A and B.

	Provider A		**Provider** B	
	Raw Score	Normalised Score	Raw Score	Normalised Score
Amount of Data	0.875	0	1	1
Completeness	$0.9\bar{4}$	1	0.7857	0
Timeliness	$0.91\bar{6}$	1	$0.8\bar{3}$	0
Weighted Overall Score ($w_i = \frac{1}{3}$)	0.9120	$\frac{2}{3}$	0.8730	$\frac{1}{3}$
Weighted Overall Score ($w_1 = \frac{1}{2}, w_2 = \frac{3}{20}, w_3 = \frac{7}{20}$)	0.9	0.5	0.9095	0.5

From the calculations it is evident that, using equal weights, provider A has the higher quality. However, if a customer has a strong interest in the amount of data, they may be better off buying from provider B. Furthermore, it becomes evident that normalising the scores does not necessarily lead to a clearer result. However, this is also owing to the fact that only two providers have been compared in this example. Transferring the quality score to pricing, suppose provider A offers their data for £ 1,200.00 and provider B for £ 1,100.00. Then, a customer with an equal appreciation for all quality criteria can calculate the respective quality for money ratio (using Equation 6.7) as quality score point per £ 1,000.00, from this it follows that provider B offers the better quality for money:

$$QM_A = \frac{0.9120}{1.2} = 0.7600$$
$$QM_B = \frac{0.8730}{1.1} = 0.7937$$

Another problem that becomes evident in this example is the fact that the amount of data is very hard to judge without further domain knowledge: considering only the number of records and attributes one might miss that provider A offers no data for the station at AMS or that provider B does only two days of forecasting. This is left to the customer, who has to adjust their queries to ensure they receive all the data they want.

6.4.5. Advanced Quality Measures

So far, the focus has been on establishing a model and demonstrating its applicability in principle. This section will show that a measure can indeed be developed for all (to some extend) automatically assessable quality criteria discussed in Section 6.4.1. Furthermore, some more advanced measures demanded by the interview partners of [MSLV12] will be introduced. In the interviews, it became clear that besides measures of data quality, it is also important to measure the usefulness of the data for a customer in other respects. In light of this, the previously defined measures *completeness* will be refined and *novelty* will be introduced as a new criterion. To start with, the automated measures not covered in Section 6.4.3 will be defined, namely *availability, latency, representational consistency, response time,* and *security.*

All of these but *representational consistency* can be expressed as quotients. *Availability* cannot be calculated for just one point in time but has to be calculated over a period of time for the past. Supposing that the past can be indicative for the future, here the probability that the service is available in the future is denoted as its past availability. To this end, the availability is checked n_v times in a given period and the availability score *ava* is defined as the quotient of successful connection attempts s and all attempts n_v:

$$ava(s,n_v) := \frac{s}{n_v} \tag{6.17}$$

In contrast, for *latency* a spot calculation is possible. However, it might make sense to average the spot values over a certain period of time. Nevertheless, only the spot calculation will be provided here. In order to assess the latency, which theoretically can be infinitely large – in case of unavailability –, an acceptable threshold t_l has to be chosen. Then, given an actual latency value l the latency score *lat* can be defined as the quotient of l and t_l. In this case, 1 one would refer to a bad score, to stay in line with all other scores, the quotient is subtracted from 1, supposing that $l \leq t_l$:

$$lat(l,t_l) := 1 - \frac{l}{t_l} \tag{6.18}$$

Very similar to this, the score for *response time* can be calculated as spot-score or as an average score. Here, the spot-score is defined in analogy to *latency* as 1 minus the quotient of a measured response time r and a threshold value t_r, supposing that $r \leq t_r$:

$$res(r,t_r) := 1 - \frac{r}{t_r} \tag{6.19}$$

For *security*, a set of security criteria S has to be defined that is generally relevant in the context of data marketplaces. For instance, this could contain encrypted transmission of different levels or encrypted on-site calculations. Together with the set of actually supported security features $F_{\text{sec}} \subseteq S$, a quotient for the security score can be calculated.

$$sec(F_{sec}) \quad := \frac{|F_{\text{sec}}|}{|S|} \tag{6.20}$$

In order to measure *representational consistency*, two or more representations of u of a provider have to be compared over time. Thus, a function is defined that determines the number of changes between $U = (X_U, \cdot)$ (now) and $U' = (X'_U, \cdot)$ (then). In order to be able to calculate a score, firstly, the auxiliary sets X^*_U and X'^*_U are defined as sets of attribute and attribute domain pairs:

$$X^*_U = \{(A_1, \text{dom}(A_1)\}, \dots, \{(A_n, \text{dom}(A_{n_a})\}; A_i \in X_U \; 1 \leq i \leq n_a \tag{6.21}$$

$$X'^*_U = \{(A'_1, \text{dom}(A'_1)\}, \dots, \{(A'_n, \text{dom}(A'_{n'_a})\}; A'_i \in X'_U 1 \leq i \leq n'_a \tag{6.22}$$

Counting the elements in X'^*_U that are also present in X^*_U returns the number of elements that have not changed, i. e., have neither been deleted, renamed, nor assigned a new domain. It should be noted that added columns have not been taken into account, as they do not have a negative effect on existing queries. Formally, the consistency score can be defined as:

$$con(X_U, X'^*_U) := \frac{|X'^*_U \cap X^*_U|}{|X'^*_U|} \tag{6.23}$$

Completeness as defined in Section 6.4.3, supposes that all attributes are of equal importance to a customer. Obviously this is not the case. Thus, it can be extended by allowing users to provide a subset ($Y_u \subseteq X_u$) which contains all attributes relevant for them. Then, altering Equation 6.14 the new completeness score may be calculated as :

$$c(u) \;=\frac{|\{\mu[A],\mu\in u,A\in Y_u|\mu[A]\neq\perp\}|}{|u|\times|Y_u|} \tag{6.24}$$

Next, the newly requested criterion *novelty* is introduced. Novelty is supposed to provide information regarding how much new information can be gained by buying data from a data vendor. In order to calculate this measure, users are required to upload a sample of their data; this sample is then matched with the data on offer to calculate how much new data the customer will receive. The reason behind this is that a data set is arguably of greater value to a customer if it extends their own data significantly. In the context of *accuracy* in Section 6.4.1, it has been discussed that matching of different data sets is not trivial. To define a simple integratebility score, let u be the universal relation of the data provider and let r_c be a relation of a customer, who seeks extension of their data by integrating them with the provider's data. To this end, it is supposed that both relations have a set of attributes $K_u \subseteq X_u$ and $K_{r_c} \subseteq X_{r_c}$, respectively, that distinctly identify a tuple, referred to as key attribute(s). Furthermore, the assumption is made that these key attributes must not contain a null value. Then, the two relations are integratable if $K_{r_c} \subseteq K_u$. Based on this, the percentage of tuples that can potentially be extended by new data can be calculated:

$$novt(r_c,u)=\frac{|\{\mu\in r_c|\mu[K_{r_c}]\in\pi_{K_{r_c}}(u)\}|}{|r_c|} \tag{6.25}$$

Alternatively, this may be written as:

$$novt(r_c,u)=\frac{|\pi_{K_u}(u)\cap\pi_{K_{r_c}}(r_c)|}{|r_c|}=\frac{|\pi_{K_u}(u\bowtie r_c)|}{|r_c|} \tag{6.26}$$

Considering that this score is the percentage of customer records that can be extended with data from the vendor, the transition to analysing the usefulness to the customer can be made. The usefulness is greater the more new attributes are added, i. e., the larger $|X_u \backslash X_{r_c}|$ is, the greater the number of total fields added. Building on Equation 6.26 this can be expressed as:

$$newf(r,u)=\frac{|\pi_{K_u}(u\bowtie r_c)|\times|X_u\backslash X_{r_c}|}{|r_c|} \tag{6.27}$$

Regarding manual assessment, i. e., the assessment of criteria in $\mathsf{M} = \{$*Believability, Interpretability, Relevancy, Representational Conciseness, Reputation, Understandability, and Verifiability*$\}$, it is proposed to provide users of the data marketplace with a user interface that allows them to review the quality of these criteria. For instance, they could be asked how well they understand the data to measure *understandability*. More concretely, it is proposed to use a Likert scale user interface (basically, the common 5 star ranking used throughout the Web) to ask a user's opinion in a non-intrusive manner. These scores have to be normalised to result in a value between 0 and 1 in order to harmonise with the other scores.

Finally, measures are presented for the hybrid criteria $\mathsf{H} = \{$*Accuracy, Customer Support, Documentation, Objectivity*$\}$. However, by definition they cannot be calculated automatically. Thus, it has to be determined in how far they can be calculated and which part has to be processed manually. For example, in case of customer support, the number of available support channels (e-mail, telephone, chat, etc.) can be assessed automatically, while the quality of each individual channel has to be evaluated manually, for instance, using the model proposed in the previous paragraph. In consequence, this results in a number of calculable measures as well as a number of manual measures. Reviewing all hybrid criteria in this way results in two ultimate sets of criteria A' and M', with $\mathsf{A}' \cup \mathsf{M}' = \mathsf{H}$. In the following A' will be described.

For all criteria, but *accuracy*, the quotient score model previously introduced can be applied. For *customer support* (sup) and *documentation* (doc), sets of required features can be defined ($R_{\text{sup}}, R_{\text{doc}}$) and a score calculated, similar to *security*, by dividing the actual number of implemented features ((F_{sup}, F_{doc})) by the respective ideal set.

$$sup(F_{\text{sup}}) := \frac{|F_{\text{sup}}|}{|R_{\text{sup}}|} \tag{6.28}$$

$$doc(F_{\text{doc}}) := \frac{|F_{\text{doc}}|}{|R_{\text{doc}}|} \tag{6.29}$$

For *objectivity*, the number of sources used to verify could be requested as a threshold t_s. Then, the actually provided sources s_p can be set in relation to this, the argument being that if more sources back the data it is more objective.

$$obj(s_p) \quad := \frac{|s_p|}{|t_s|} \tag{6.30}$$

Regarding *accuracy*, besides the universal relation of the data provider u, a correct corresponding reference relation r_r with the same schemas $U(X_u, \cdot) = R(X_{r_r}, \cdot)$ is required. Based on these two, the *accuracy* of u may be determined. *Accuracy* can be categorised in syntactic accuracy and semantic accuracy. While the first is concerned with a value being of its domain, the latter actually checks for the correctness of a value [BS06]. The latter is inherently more complex for it also requires the matching of tuples in different relations which is all but a trivial task [BS06; BGM+09; MJC+07]. It has been opted to abstract from this problem here. This is reasonable as it is the main aim of this section to demonstrate that many quality criteria measures can be sensibly used for the comparison of offers and ultimately for pricing. Based on this, a measure introduced by Batini and Scannapieca [BS06] shall be used as representative for many possible measures. To this end, a Boolean auxiliary function $c(a)$ is needed that returns 0 if the value of a is syntactically correct and, 1 if it is not. Even though semantic accuracy is excluded, it is necessary to match tuples in u to tuples of r_r, which can be difficult because incorrect semantics might prevent a meaningful matching. Thus a second, Boolean function $m(\mu,r_r)$ is introduced that returns 0 if a tuple $\mu \in u$ can be matched to a tuple in r_r, and 1 if not. Additionally, a third Boolean function $\beta(cond)$ is required that returns 1 if the condition *cond* is met, and 0 if not. Based on all this, the *accuracy score*, i. e., the percentage of semantically correct and correctly matched tuples, can be defined as in Equation 6.31:

$$acc(u,r_r) = \frac{1}{|u|} \sum_{\mu \in u} \beta \left(\left(\sum_{A \in X_u} c(\mu[A]) = 0 \right) \wedge (m(\mu, r_r) = 0) \right) \tag{6.31}$$

6.5. Quality-Based Pricing

The last section was mainly concerned with a comparison of data offerings and, therefore, has taken rather a customer's point of view. Now, a focus shift is conducted and a provider's view is taken. More precisely a pricing mechanism will be developed that helps providers to discriminate their customers based on their willingness to pay and their preferences. In return, customers receive a relational data product that matches their needs.

Given the fact that data providers can decrease the quality of their product, this can be used for discounts as suggested in [Tan14; TASB14; TSBV13]. Building on the data quality measures established in Section 6.4.3, this section proposes an approach of data pricing in which not only one quality dimension of a relational data product is adjusted according to a user's willingness to pay but all dimensions are, taking user preferences into account. To this end, providers advertise a price. If this price exceeds a user's willingness to pay, they may suggest a price and reveal their preferences for certain quality criteria. Subsequently, a data product tailored to their needs and willingness to pay is created and delivered.

A physical good with a similar procedure of a quality reduction for different use cases is ethanol. If it is sold for drinking, it is commonly expensive and highly taxed. However, if it is used as a fuel or as a solvent, it is comparatively cheap. In order to prevent abuse, i. e., drinking the cheaper alcohol, additives are used that make it bitter or toxic. Transferring this idea to data, it can be offered cheaper if it is of lesser quality as it provides less utility to consumers.

As a prerequisite, it is supposed that users know their preferences and can express them in the following form:

$$q_i \succsim q_j \text{ f. a. } q_i, q_j \in \mathsf{Q}_\mathsf{v}$$

Regarding the attributes of preferences, at this point two assumptions mentioned in [PR13] are relevant: a) completeness, i. e., preferences exist for all combinations q_i and q_j; b) transitivity, i. e., preferences are totally ordered, formally: if $q_1 \succsim q_2 \wedge q_2 \succsim q_3 \Rightarrow q_1 \succsim q_3$. In order to express these preferences users are asked to provide their appreciation of each quality measure in the way they were asked in the comparison model, discussed in Section 6.4.2 and presented in Figure 6.2.

The basic approach of the overall pricing model is the following: A data provider has relational data with a given quality level, which they are willing to sell at price P. If customers want to pay less, they are given the chance to propose an alternative price $W < P$. Then the data products will be tailored to their needs.

In Section 6.4.3 seven quality criteria were identified that allow for *continuous* versioning (tailoring). This means that for these criteria an arbitrarily large number of versions can be created. They were assembled in the set V = {*Accuracy, Amount of Data, Availability, Completeness, Latency, Response Time, Timeliness*}. Furthermore, five criteria have been established for which a limited number of versions can be created, i. e., which allow for discrete versioning, collocated in G = {*Customer Support, Documentation, Security, Representational Conciseness, Representational Consistency*}. To handle all alike, all criteria will be treated in a way to create discrete versions. In the following the differentiation is subordinated and it will be referred to all quality criteria as: $Q = \mathsf{V} \cup \mathsf{G}$. However, the order will be of importance, hence, from now on, a list of quality criteria q will be used: $q = (q_1, \ldots, q_{n_q})$ with $n_q = |Q|$ elements.

The remainder of this section first introduces a notation of utility and a means to derive versions for each quality criterion based on utility. Secondly, it will be demonstrated how prices can be derived for these versions in Section 6.5.2. Based on this, the pricing problem will be shown to be representable as a Multiple-Choice Knapsack Problem (MCKP) in Section 6.5.3. Subsequently, approaches to solve the MCKP will be discussed. Section 6.5.5 shows that versions can indeed be calculated based on the derived scores. This section is concluded by an example illustrating how knapsack-based pricing can be applied in Section 6.5.6.

6.5.1. Introducing Utility

In Section 5.1 it has been established that goods provide utility. Commonly, micro economists investigate utility functions for a set of n_g goods, which will be referred to as benefit function $b = f(x_1, \ldots, x_{n_g})$ [SMS11] to not confuse utility and the universal relation. In the context of this work, sets of goods will not be looked at, rather this thesis will focus on one relational data good and its quality attributes, the utility of which may be formalised as $b = f(q_1, \ldots, q_{n_q})$, where q_i represents the quality scores for quality criterion q_i.

Here, it will be supposed that quality criteria are independent, i. e., that the *consumption* of one quality criterion does not have an effect on the utility of other quality criteria. While this is not the case for extremes, e. g., an incomplete data set is less likely to be accurate than a complete one, this is a necessary simplification to handle all dimensions in the following model. Furthermore, it can be argued that when looking at one item only, and keeping the others constant (ceteris paribus), it is a valid assumption. Economist say, the quality criteria are *perfect substitutes* with a constant *marginal rate of substitution*, i. e., the willingness to change one quality criterion for another is constant but not necessarily 1, for instance in Figure 6.3 one unit q_1 is as good as two units q_2.

Regarding utility functions there are commonly three basic assumptions, presented here based on [SMS11]:

1. Non-satiation, i. e., more of a good is always better, formally:
   $b'(x) > 0 \text{ f. a. } x$

2. Decreasing marginal utility, i. e., every additional unit increases the utility less than its predecessor, formally: $b''(x) < 0$ f. a. x

3. Decreasing marginal rate of substitution, which means that well mixed combinations of goods are preferred over extremes. However, having stated that, in this work, the attributes are regarded as perfect substitutes, a constant marginal rate of substitution is given and this formalism will not be investigated further.

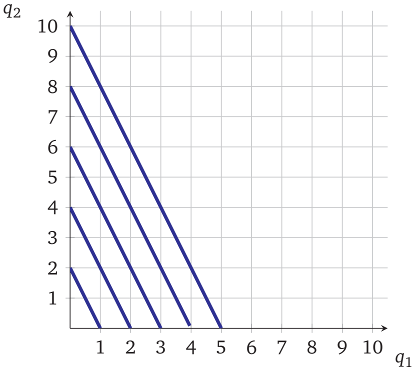


Figure 6.3.: Perfect Substitutes, q_1 and q_2 with a Constant Rate of Substitution: -2.

Two well-known function types satisfy these criteria, logarithm functions – for which the natural logarithm has been chosen as representative – as well as any root function $\sqrt[a]{x}, a \in \mathbb{N}_{\geq 2}$. Given that the quantity of a good cannot be negative, the relevant domain for both functions is $\mathbb{R}_0^+$. Figure 6.4 presents the square root function as well as the natural logarithm, shifted by one so that it passes the origin, as possible utility functions. The illustration shows that while generally increasing – meeting condition 1 –, the slope is decreasing – meeting condition 2.

Formally, Equation 6.33 and Equation 6.34 show that the aforementioned conditions are met by the natural logarithm.

$$b(x) = \ln(x) \tag{6.32}$$

$$b'(x) = \frac{1}{x} \qquad > 0 \text{ f. a. } x \in \mathbb{R}_0^+ \tag{6.33}$$

$$b''(x) = -\frac{1}{x^2} \qquad < 0 \text{ f. a. } x \in \mathbb{R}_0^+ \tag{6.34}$$

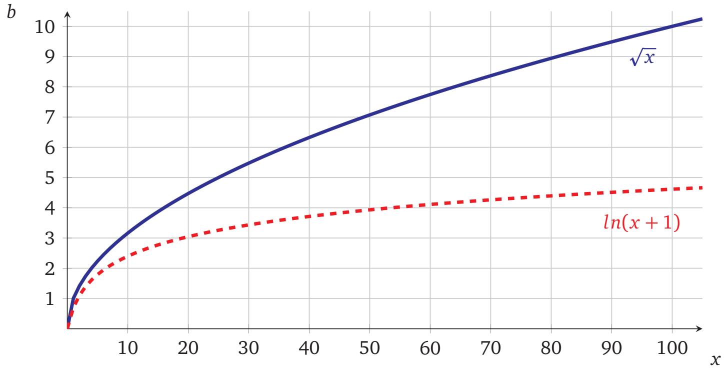


Figure 6.4.: Natural Logarithm and Square Root Function as Possible Utility Functions.

The fact that any root function satisfies the utility function conditions, is presented in Equation 6.36 and Equation 6.37.

$$b(x) = \sqrt[a]{x} \qquad = x^{\frac{1}{a}} \qquad a \in \mathbb{N}_{\geq 2} \tag{6.35}$$

$$b'(x) = \frac{1}{a}x^{\frac{1-a}{a}} \qquad = \frac{1}{a}x^{-\frac{a-1}{a}} = \frac{1}{a\sqrt[a]{x^{a-1}}} \qquad > 0 \text{ f. a. } x \in \mathbb{R}_0^+ \tag{6.36}$$

$$b''(x) = -\frac{a-1}{a} \times \frac{1}{a}x^{\frac{1-2a}{a}} \quad = -\frac{a-1}{a^2}x^{-\frac{2a-1}{a}} = -\frac{a-1}{a^2\sqrt[a]{x^{2a-1}}} \quad < 0 \text{ f. a. } x \in \mathbb{R}_0^+ \tag{6.37}$$

In this thesis, it is proposed to create versions based on the expected utility. Thus, the utility function is used to create m_l utility-based versions or levels so that $b_j - b_{j-1} = const.$ f. a. $j, 1 \leq j \leq m_l$. To this end, the quality scores which have been standardised to fit the domain [0,1] will be scaled to match a sector of the utility function's domain $[x_{\min}, x_{\max}]$, e. g., [0,100] for the square root. It is worth a thought that data with some quality scores beneath a certain threshold t_q are useless. To address this, it is also possible to transform only the interval $[t_q, 1], 0 \leq t_q \leq 1$ from the original score to the representative sector of the utility function, i. e., at quality score t_q the utility level of that quality score is 0.

To arrive at the necessary minimum quality score for each utility level, the inverted utility function is used, e. g., x^2 for $\sqrt{x}$ and $e^x - 1$ for $\ln(x+1)$. Graphically, this is shown in Figure 6.5 for the natural logarithm and in Figure 6.6 for the square root function.

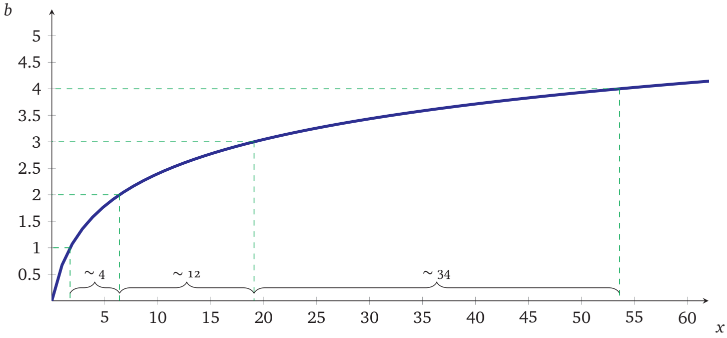


Figure 6.5.: Groups of Same Utility for $\ln(x+1)$.

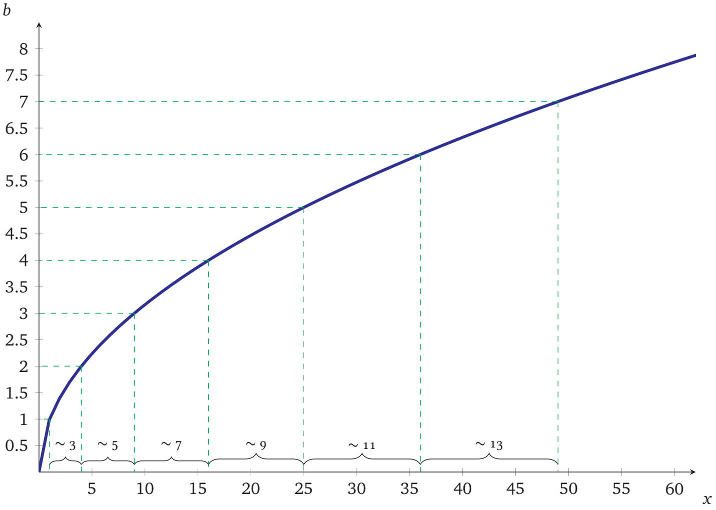


Figure 6.6.: Groups of Same Utility for $\sqrt{x}$.

As these exemplary figures show, the number of utility levels that can be arrived at, given a domain (in the example [0,60]), vary between the two. While for the square root model the utility-based levels increase linearly with x because the difference of two levels can be calculated as $x^2-(x-1)^2=2x-1$, for the natural logarithm they increase exponentially in x because the difference of two levels is $\mathrm{e}^x-1-(\mathrm{e}^{x-1}-1)=\mathrm{e}^x-\frac{\mathrm{e}^x}{\mathrm{e}}=\mathrm{e}^x(1-\frac{1}{\mathrm{e}})$. Thus, in the following, the square root function will be used as it produces more reasonable utility levels. The case can be made that other root functions $\sqrt[a]{x}$, which scale polynomially with the degree $a-1$, could be used as well. However, this is a matter of implementation as the model is – as will be seen – independent from the function. A positive side-effect of using the square root with, for instance, a domain of [1,100] and $m_l=10$ utility levels, as hereby proposed for this work, is that the examples are more illustrative.

The utility-based quality level vector l contains the concrete values of the utility level l_j in order. In the example manifestation presented here, it is supposed that $l_j=j$ f. a. $j, 1\leq j\leq m_l$. While this applies for those quality criteria that allow for continuous versioning (i. e., $q\in\mathsf{V}$), for those that only allow for discrete versioning (i. e., $q\in\mathsf{G}$) a smaller number has to be chosen, here three utility levels $l_1=3, l_2=6, l_3=9$ are chosen form the utility function for $q\in\mathsf{G}$ – according to Goldilocks principle, discussed in Section 5.4.3. To differentiate between the utility level vectors of both sets, they have an according superscript, resulting in the two vectors l^{V} and l^{G}. Since quality levels in l^{G} do not correspond to concrete quality scores, determining a value for them is meaningless. It is rather advisable to manually determine the amount of service for each level. Exemplary figures for both variants are presented in Table 6.5, where levels for the second type have been marked with an X. The used utility function and according versions are depicted in Figure 6.7.

Table 6.5.: Used Utility Levels Mapped to Versions; Showing Required Quality Score (QS).

Utility Level (l_j)	0	1	2	3	4	5	6	7	8	9	10
QS Required to Reach Version ($q\in\mathsf{V}$)	0	1	4	9	16	25	36	49	64	81	100
QS Required to Reach Version ($q\in\mathsf{G}$)	0	⊥	⊥	X	⊥	⊥	X	⊥	⊥	X	⊥

While in reality the utility does differ between customers, the general trend is the same, and will here be approximated by the same function. Furthermore, it is acknowledged that not all quality criteria have the same importance for customers. For example, completeness may be more important for a customer than timeliness because they want to do some time-independent analysis, while for another customer timeliness might be more important because they base time-critical decisions on the data. To represent this in the model, the utility gained from each q_i's quality score is weighted with a user provided ω_i that represents the importance of all quality criteria relative to each other.

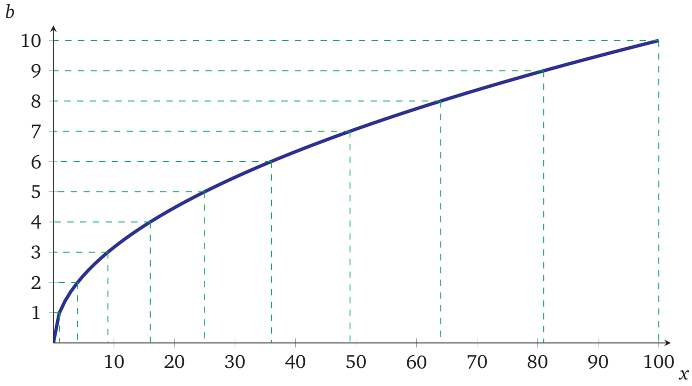


Figure 6.7.: Exemplary Used Utility Functions with Ten Utility Levels.

To receive ω_i, users are asked to express their preferences using the slider GUI proposed in the context of quality comparison and depicted in Figure 6.2. The effects this has on the utility function is illustrated in Figure 6.8. Moreover, this results in a weight vector ω such that:

$$\forall q_i \exists ! \omega_i, 1 \leq i \leq n_q \text{ and } \sum_{i=1}^{n_q} \omega_i = 1$$

Based on all this, a weight matrix b can be calculated for each user. This matrix shows for which quality criterion q_i with an according weight ω_i what actual utility b_{ij} can be reached for the different utility levels l_j^{V} and l_j^{G}. It is calculated as follows:

$$b_{ij} = \begin{cases} \omega_i \times l_j^{\mathsf{V}} \text{ f. a. } q_i \in \mathsf{V} \\ \omega_i \times l_j^{\mathsf{G}} \text{ f. a. } q_i \in \mathsf{G} \end{cases}$$

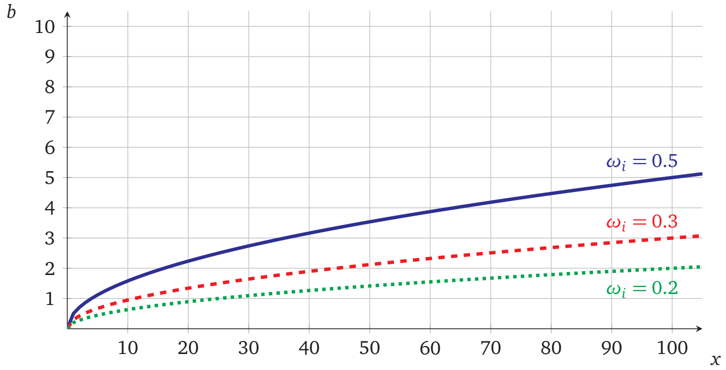


Figure 6.8.: Example Utility Functions $\omega\sqrt{x}$ for $\omega_1 = 0.5, \omega_2 = 0.3,$ and $\omega_3 = 0.2$.

6.5.2. Price Attribution

Having extensively discussed for each quality criterion how a utility level is arrived at, this subsection elaborates on how prices can be attached to the different levels. As proposed elsewhere [TSBV13; TASB14], this work builds on the idea that providers offer data for an ask price P and customers may suggest an alternative (lower) bid price W. If $W < P$ the quality of the data is lowered to meet the price W suggested by the customer.

Besides the overall ask price P providers want to achieve, they have to specify the importance of different quality criteria from their point of view. This may either be done based on the cost the different quality criteria caused when being created or based on the perceived utility of the different criteria. As argued before, the utility-based approach is preferable; however, the cost-based approach can serve as point of reference if no further information is available. Additionally, it is also an option to attribute an equal weight to all quality criteria. Similar to the user weighting vector ω, providers define a weight vector κ such that:

$$\forall q_i \exists! \kappa_i, 1 \leq i \leq n_q \text{ and } \sum_{i=1}^{n_q} \kappa_i = 1$$

For the actual distribution of the overall ask price P to the different quality levels and quality criteria two fundamentally different approaches can be implemented. In any case the overall price would be distributed to the different quality criteria using κ. Then, prices can be attributed to the different quality levels using the utility levels or using the relative satisfaction of each quality criterion. The first will lead to linear prices corresponding to the benefit, which is arguably a fair way of pricing a data product. In this case, the price w_{ij} for each quality criterion q_i at each quality level l_j is calculated using a formula of the form $w_{ij}(b_{ij})$, in detail:

$$w_{ij} = P \times \kappa_i \times \frac{b_{ij}}{b_{i,n_q}}$$

The alternative is to model prices linear to the actual quality scores required to reach this level. This will result in increasing prices for the utility levels. However, looking at it from the discount perspective, this means that the biggest discount is granted for the sacrifice of the first utility level and then decrease. The calculation of w_{ij} in this case is conducted based on the inverted utility function $w_{ij}(x) = P \times \kappa_i \times b^{-1}(x)$, in this case $b^{-1}(x) = x^2$ and the overall utility levels in l:

$$w_{ij} = \begin{cases} P \times \kappa_i \times \frac{b^{-1}(l_j^{\mathsf{V}})}{b^{-1}(l_{m_l}^{\mathsf{V}})} \text{ f. a. } q_i \in \mathsf{V}, 1 \leq j \leq l_{m_l}^{\mathsf{V}} \\ \\ P \times \kappa_i \times \frac{b^{-1}(l_j^{\mathsf{G}})}{b^{-1}(l_{m_l}^{\mathsf{G}})} \text{ f. a. } q_i \in \mathsf{G}, 1 \leq j \leq l_{m_l}^{\mathsf{G}} \end{cases}$$

It cannot be decided per se which of the two alternatives is the better one. There are some quality scores, such as the amount of data, for which it is sensible to grant a good discount if less data is to be delivered. In other cases, such as accuracy it might make more sense to scale prices according to the utility levels. That being said, what model to choose is a business decision that has to be made for each individual criterion depending on the attributes of the criterion as well as on the intended fairness of the pricing model. Given the stronger decrease when using the inverted utility function, the average price across all levels is smaller than in the linear case; this speaks in favour of the latter model from a customer's perspective. After all, it is not important what product is actually delivered as the cost of creating it is marginal. What is more important is that customers get a fair discount for their scarifies of quality. This is achieved by either of the two.

6.5.3. Fair Knapsack Pricing

The last two sections developed a framework to attribute utility as well as a price to different quality criteria for relational data products. This section demonstrates, given these prerequisites, that the pricing problem can be shown to be a Multiple-Choice Knapsack Problem (MCKP). The knapsack problem, which will be described according to Kellerer et al. [KPP04], is comparatively old; according to the authors, it was already studied in 1897. A common illustration for the knapsack problem is that of a mountaineer, Simon for demonstrative purposes, who is packing his backpack (i. e., knapsack) for a climbing tour. Simon has a number of items available that he considers useful for his climbing trip. Each of these items (numbered from 1 to m_l) provides him with a benefit b_i and has a weight w_i. Since Simon, despite being a strong chap, can only carry a limited amount of weight, he wants to optimise his knapsack in a way that he maximises his utility given a maximum weight W. More formally, the auxiliary vector a stores for each item available if it has been put in the knapsack as $a_i \in \{0,1\}$. Given these presupposition, the knapsack problem can be formalised as follows:

$$\text{maximise} \sum_{i=1}^{m_l} b_i a_i \tag{6.38}$$

$$\text{subject to} \sum_{i=1}^{m_l} w_i a_i \leq W \tag{6.39}$$

Solving this will lead to an optimal solution vector $a^* = (a_1^*, \ldots, a_n^*)$ specifying which objects to choose. The optimal benefit is referred to as z^*.

This standard knapsack problem has been extended in several ways. One of the most flexible knapsack models is the herein-applied MCKP [KPP04]. Other areas of application include capital budgeting, menu planning, and transforming non-linear knapsack problems to MCKPs [Pis95; KPP04]. In a MCKP, items are chosen from n_q sets of available items rather than from just one set of available items, an additional restriction being that from each set exactly one item has to be chosen. Using the variables from the previous sections and extending the vector a to a matrix, pricing can be formalised using the MCKP as presented in [KPP04].

In the following, Equation 6.40 and Equation 6.41 extended the original knapsack problem, presented in Equation 6.38 and Equation 6.39, respectively, to multiple sets to choose from. Equation 6.42 restricts the choice to one item per set and Equation 6.43 determines that items are indivisible.

$$\text{maximise} \sum_{i=1}^{n_q} \sum_{j=1}^{m_l} b_{ij} a_{ij} \tag{6.40}$$

$$\text{subject to} \sum_{i=1}^{n_q} \sum_{j=1}^{m_l} w_{ij} a_{ij} \leq W \tag{6.41}$$

$$\text{and} \sum_{j=1}^{m_l} a_{ij} = 1, i = 1, \ldots, n_q \tag{6.42}$$

$$\text{and } a_{ij} \in \{0; 1\}, i = 1, \ldots, n_q, j = 1, \ldots, m_l \tag{6.43}$$

6.5.4. Solving the Multiple-Choice Knapsack Pricing Problem

In order to create a custom-tailored relational data product, the Multiple-Choice Knapsack Pricing Problem (MCKPP) has to be solved. This is not a trivial task. Even the basic knapsack problem belongs to the class of $\mathcal{NP}$-complete problems, a proof of which can be found in [GJ79]. In very simple terms and according to the current state of research, $\mathcal{NP}$-complete problems are in theory solvable (by exponential-time algorithms) but not in practice (by polynomial-time algorithms), and if one of them were, they all would be [VW11]. The MCKP is also $\mathcal{NP}$-complete [IHTI78], as it can be reduced from the ordinary knapsack problem [KPP04]. Consequently, for a very large input, an exact solution cannot be expected within reasonable time, so that approximations are necessary.

Despite the fact that MCKP is $\mathcal{NP}$-complete, it can be solved in pseudo-polynomial time using, for instance, dynamic programming; several algorithms have been presented to achieve this [Pis95]. Most algorithms start by solving the linear MCKP to obtain an upper bound. For the linear MCKP the restriction $a_{ij} \in \{0; 1\}$ has been relaxed to $a_{ij} \in [0,1]$, which means it allows the choosing of a fraction of an item [Pis95].

In order to solve the linear MCKP, the concept of dominance is introduced which filters elements that will never be chosen in an optimal solution, here

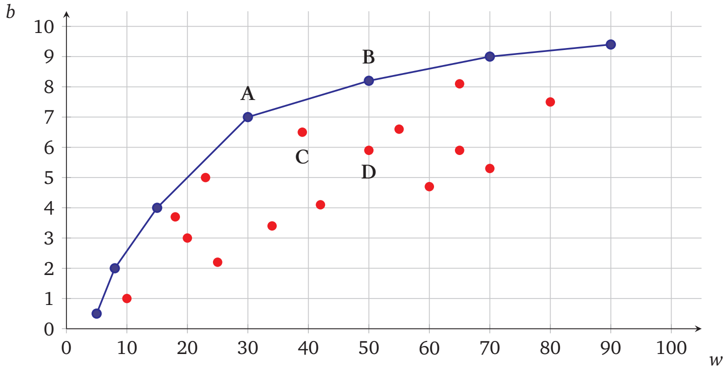


Figure 6.9.: Illustration of Dominance and LP-Dominance.

presented as shown in [SZ79]. For two items j,k in the same set i, item j dominates item k if:

$$w_{ij} < w_{ik} \text{ and } b_{ij} > b_{ik}$$

Sinha and Zoltners [SZ79] proved that under these conditions a_{ik} is never part of an optimal solution. Furthermore, they proved that a_{ik} is never part of an optimal solution if it is LP-dominated by two items. The concept of LP-dominance, which also allows for linear relaxation, is formalised for three items h,j,k in the same set i with the following rankings $w_{ij} < w_{ik} < w_{ih}$ and $u_{ij} < u_{ik} < u_{ih}$ as follows:

$$\frac{b_{ih} - b_{ik}}{w_{ih} - w_{ik}} > \frac{b_{ik} - b_{ij}}{w_{ik} - w_{ij}}$$

The concept is illustrated in Figure 6.9; it can be seen that element B dominates element D and that elements A and B LP-dominate element C. That being said, this common reduction to the so-called set of LP-extreme items is not necessary for MCKPP because of the way in which the respective utility and weights are calculated. This is easily verifiable by comparing Figure 6.7 to Figure 6.9. However, items of an equal ratio (which can appear in MCKPP when prices scale linearly with utility) must not be eliminated from the sets as this would lead to a single random item because all have the same ratio of benefit to weight.

Algorithm 6.1 presents a general greedy algorithm to solve the MCKPP. It has been adapted from the greedy algorithm outlined in [KPP04]. The main difference is that the original algorithm contained a preparation step to derive the LP-extremes of each set, which is not necessary for the MCKPP. The algorithm eventually results in a matrix a indicating which items to choose, a value $W - \bar{c}$, which represents the total cost of these items, and a score z, indicating the total utility achieved.

Algorithm 6.1 Greedy Algorithm to Solve MCKPP, Adapted from [KPP04]

```
# In the following, $i$ is the index for quality scores and $n$ denotes the number of quality scores.
#Furthermore, $j$ is the utility level index and $m$ denotes the total number of levels.
#Initialise:
for $i = 1$ to $n$ do
    $\bar{c} = W - w_{i1}$                                  ▷ The residual weight
    $z = u_{i1}$                                            ▷ The achieved utility
    for $j = 2; j < m; j++$ do
        $\tilde{b_{ij}} = b_{ij} - b_{i,j-1}$                ▷ The incremental benefit matrix
        $\tilde{w_{ij}} = w_{ij} - w_{i,j-1}$                ▷ The incremental weight matrix
        $\tilde{e}_{ij} = \frac{\tilde{b}_{ij}}{\tilde{w}_{ij}}$   ▷ The incremental efficiency matrix
    end for
end for
#Sort:
L := sort($\tilde{e}_{ij}$)          ▷ One dimensional list of $\tilde{e}_{ij}$; original indices are maintained
#Solve:
for all $\tilde{e}_{ij}$ in $L$ do
    if $\bar{c} - \tilde{w_{ij}} > 0$ then            ▷ Fill knapsack as long as there is space left
        $z += \tilde{p_{ij}}$
        $\bar{c} -= \tilde{w_{ij}}$
        $a_{ij} = 1$
        $a_{i,j-1} = 0$
    else        ▷ The so-called split item $a_{st}$ has been found, i. e., criterion $s$ at level $t$
        $a_{ts} = \frac{\bar{c}}{\tilde{w_{ts}}}$
        $a_{t,s-1} = 1 - a_{ts}$
        $z += \tilde{p_{st}}$
        break loop
    end if
end for
```

The greedy algorithm, presented in a pricing-tailored form in Algorithm 6.1, has a runtime of $\mathcal{O}(n \log n)$ owing to the sorting in Line 14. This form of a greedy-type algorithm is often used as a starting point for further procedures such as branch and bound [KPP04]. Furthermore, the split solution is generally a good heuristic solution; however, it has to be pointed out that as a solution algorithm – despite being used here for demonstrative purposes as it is easy to comprehend – it is unsuited. The reason for this is that its performance is arbitrarily bad, i. e., while performing quickly, the solution is not guaranteed to be the optimal solution [KPP04].

Further approximation algorithms exist that do have certain performance guarantees. The performance of algorithms is measured using the optimal solution value z^* and the achieved solution value z'. An algorithm is referred to as ϵ-approximation ($\epsilon \in [0,1]$, 0 being a perfect solution) if the following equation holds for all problem instances I [KPP04; MT90]:

$$\frac{z'(I)}{z^*(I)} \geq 1-\epsilon \Leftrightarrow \frac{z^*(I)-z'(I)}{z^*(I)} \leq \epsilon \tag{6.44}$$

Gens and Levner [GL98] have presented a binary search approximation algorithm running in time $\mathcal{O}(n_t \log n_q)$, where n_q is the number of quality criteria and n_t is the total number of items over all quality criteria $n_t = \sum_{i=1}^{n_q} m_{l\,i}$. At this point, it should be mentioned that $m_{l\,i}$ is used here to indicate that depending on whether $q_i \in \mathsf{V}$ or $q_i \in \mathsf{G}$, ${m_l}^{\mathsf{G}}$ or ${m_l}^{\mathsf{V}}$ has to be substituted. However, the guarantee is $\epsilon = 0.8$, which is still a considerably bad result even though the authors argue that the actual performance may be much better than that.

Using dynamic programming, a fully polynomial time approximation scheme can be developed [KPP04]. Lawler [Law77] presents an ϵ-approximation that runs in $\mathcal{O}(n_t \log n_t + \frac{n_t n_q}{\epsilon})$, the first term being due to sorting which might be omitted here. A similar approach is also presented in [KPP04].

Approaches to solve the MCKP to optimality include branch and bound [DKW84], dynamic programming [DW87] (which is often used to solve dynamic pricing challenges [NRRS05]), hybrid algorithms of the former [DRW95], and expanding core algorithms [Pis95]. Pisinger [Pis95] presents a minimal expanding core algorithm solving the MCKP to optimality. It is based on the idea that the problem is first solved for a core set of classes, i. e., quality criteria, $C \subseteq \mathsf{Q}_\mathsf{v}$ based on the split item. Then, gradually more classes are added. This results in a runtime of $\mathcal{O}(n_t + W \sum_{q_i \in C} m_{l\,i})$, where W denotes the weight limit. This results in a linear solution time for a minimal core and pseudo-polynomial time for larger cores [Pis95]. Pseudo-polynomial time refers to the fact that the algorithm will show exponential behaviour when confronted with exponentially-large input numbers [GJ79].

Based on this elaboration it can be stated that the MCKPP can be solved to optimality in time $\mathcal{O}(n_t + W \sum_{q_i \in C} m_{l\,i})$ and ϵ-approximated in time $\mathcal{O}(n_t \log n_t + \frac{n_t n_q}{\epsilon})$. This is summarised in Theorem 6.1 and Corollary 6.2.

Theorem 6.1. Given a list of quality criteria q that can be measured on the interval [0,1], a customer-provided vector of preferences for these criteria ω, a bid price of the customer W, an ask price of the provider P, a utility function b, a cost distribution vector κ, and a weighting function $w(x)$ or $w(b)$, the MCKPP can be solved to optimality in time $\mathcal{O}(n + W\Sigma_{q_i \in C} m_{l\,i})$.

Sketch of Proof.

1. Translate inputs into MCKPP as demonstrated in Sections 6.5.1 to 6.5.3.
2. Use Pisinger's [Pis95] algorithm to solve the according MCKP.

Corollary 6.2. Given the above, the MCKPP can be ϵ-approximated in time $\mathcal{O}(n_t \log n_t + \frac{n_t n_q}{\epsilon})$

Sketch of Proof.

1. Translate inputs into MCKPP as demonstrated in Sections 6.5.1 to 6.5.3.
2. Use any ϵ-approximation algorithm to solve the according MCKP.

Notwithstanding these sketched proofs, the MCKP can commonly be solved quickly in practise [DRW95]. Given that in the MCKPP the weights correlate with the benefits per definition, this results in strongly correlated data instances, which are particularly hard for knapsack algorithms, as no dominated items exist [Pis95; KPP04].

In 1995, Pisinger [Pis95] presented computational experiments on commodity hardware for his algorithm. Nearly a decade later, in 2004 results for the same algorithm on more recent commodity hardware were presented in [KPP04]. Table 6.6 shows the results relevant for strongly correlated data instances in the problem scope of MCKPP. Two realistic cases will be presented for each year *Case 1* considers 100 quality dimensions and 10 quality levels and *Case 2* considers 100 quality dimensions and 100 quality levels. Furthermore two different maximum bid prices are considered *Max Bid 1* is set to 1,000 and *Max Bid 2* to 10,000. The time is reported in seconds.

Table 6.6.: Experimental Results for MCKP Calculations Using Pisinger's [Pis95] Algorithm.

	1995 [Pis95]		**2004 [KPP04]**	
	Max Bid 1	Max Bid 2	Max Bid 1	Max Bid 2
Case 1	0.37	5.16	0.061	0.561
Case 2	0.33	6.93	0.52	0.828

Given that another decade has passed since, it is a reasonable assumption that the MCKPP can practically be solved in less than a second on commodity hardware. However, computational experiments will have to prove this.

6.5.5. Modifying Quality

To briefly recapitulate, Tang et al. have demonstrated that it is feasible to decrease data quality in exchange for a discount. One the one hand, they have analysed the possible of randomly choosing an XML sub-tree from an XML document so that the price of the sub-tree is equivalent to the offered price [TASB14]. On the other hand, they have shown how data accuracy can be reduced by determining values for relational data from a probability distribution, the distance of which to the original distribution depends on the height of the discount [TSBV13]. Both methods modify data products according to the discount and could, thus, be used in the pricing model suggested here. However, XML data has not been the primary focus of this work as only quality measures for relational data have been investigated so far. Thus, the completeness ideas presented by Tang et al. [TASB14] are not applicable to this work directly. Nevertheless, while this work focused on relational data, the presented framework is applicable to any type of data for which appropriate quality measures can be found.

At this point, it is argued that for any of the quality measures presented previously an algorithm can be found that creates a quality decreased relational data product according to a proposed discount. For accuracy, this has extensively been described in [TSBV13]. Largely, modifications to the quality can be group into three categories:

1. The modification of accompanying services, e. g., delivery conditions and comprehensiveness of *support*

2. The modification of the data itself, e. g., decreasing the *accuracy*
3. The modification of the view on the data, e. g., a limited *amount of data*

As examples this section will present algorithms to modify the *completeness* as representation of the second as well as *timeliness* and *amount of data* as representation of the third. No representation for the first will be given as this is a rather trivial contractual task and does not affect the data as such. Nevertheless, an example will be provided in the next subsection.

Obviously, the order in which the quality is decreased plays an important role. For instance, if null values are inserted first and then the accuracy is reduced, the accuracy reduction might build on a wrong distribution. Here, it is suggested to apply criteria first that reduce the size, i. e., criteria of the third type and also completeness, before the rest of the quality is lowered.

The first quality measure to be looked at in more detail is *completeness*. In Section 6.4.3, Equation 6.14 defined completeness as the number of non-null value cells divided by the overall number of cells:

$$c(u) = \frac{|\{\mu[A], \mu \in u, A \in X_u | \mu[A] \neq \perp\}|}{|u| \times |X_u|} \tag{6.45}$$

Alternatively, this may be written as:

$$c(u) = 1 - \frac{|\{\mu[A], \mu \in u, A \in X_u | \mu[A] = \perp\}|}{|u| \times |X_u|} \tag{6.46}$$

For simplification purposes, n_v shall refer to the number of null values:

$$n_v = |\{\mu[A], \mu \in u, A \in X_u | \mu[A] = \perp\}| \tag{6.47}$$

Thus:

$$c(u) = 1 - \frac{n_v}{|u| \times |X_u|} \tag{6.48}$$

This implies that in order to reduce the completeness further, null values have to be inserted at random. In the following u^* is the universal relation to be sold before any modification and u afterwards. The same applies to all other relevant variables, n_v^* is the number of null values before and $n_{v\,t}$ after the quality modification. The suffix t indicates a target value. Furthermore, $x_{\max}$ denotes the maximum of the domain of the utility function and x the utility score at the chosen level. To lower the completeness the actual value for completeness has to be determined and the target value for completeness has to be calculated based on the selected quality level.

$$c_t = \frac{x}{x_{\max}} \times c(u^*) \tag{6.49}$$

Based on this the target number of null values $n_{v\,t}$ can be calculated:

$$\frac{x}{x_{\max}} \times c(u^*) \overset{!}{=} 1 - \frac{n_{v\,t}}{|u| \times |X_u|} \tag{6.50}$$

Resulting in:

$$n_{v\,t} = \left\lceil |u| \times |X_u| \times \left(1 - \frac{x}{x_{\max}} c(u^*)\right)\right\rceil \tag{6.51}$$

Note that the ceiling function has to be used in Equation 6.51 to ensure n_t is an integer as no half null values exist. Alternatively, the floor function could be used, this is at the providers discretion but would result in a slightly better quality. Based on this target value for null values n_t Algorithm 6.2 presents an exemplary method to achieve the modified data set u. At this point, it should be noted that this algorithm does not check whether a null value already is in place. This may result in a scenario where little to no null values are added. While for relatively high completeness scores this is unlikely, bad scores might not be effectively lowered. However, it allows for efficient computing and the quality can be effectively lowered for relevant data products, i. e., for those that have a high completeness beforehand.

Algorithm 6.2 Adapting the Relation to the Completeness Score

```
#In the following $u$ is the relation to be sold and $X_u$ denotes the according attributes.
#$\mu_i$ indicates the $i^{\text{th}}$ tuple in $u$
#$a = n_t - n^*$; by definition: $n_t > n^*$
function IncreaseNulls($a, u, X_u$)
    while $a-- > 0$ do
        $i := rand(0, |u|)$                ▷ Choose a random integer between 0 and $|u|$
        $j := rand(0, |X_u|)$              ▷ Choose a random integer between 0 and $|X_u|$
        $\mu_i[A_j] := \perp$
    end while
    return $u$
end function
```

The *amount of data*, has been defined as comparison measure in Equation 6.8. In this context a closer look at the amount of attributes or columns — *AoC* — will be taken. In Equation 6.10 it has been defined as:

$$AoC(u) := \frac{|Y|}{|X_u|} \tag{6.52}$$

Given $AoC_t = \frac{x}{x_{\max}}$, $y_t = |Y|$ can be calculated. Similar to *completeness*, the ceiling function is used to ensure that y_t is an integer.

$$y_t = |Y| = \left\lceil AoC_t \times |X_u| \right\rceil \tag{6.53}$$

Furthermore, it is supposed that the customer provides a list of attributes $Z \subseteq X_{u^*}$ in decreasing order of their liking, such that the first attribute is the most important and the last is the least important. If Z is not provided, it is at the provider's discretion which attributes actually to deliver. While this could also be the standard, giving customers the choice seems more fair. Algorithm 6.3 presents an exemplary method to achieve $Y \subseteq X_{u^*}$ provided y_t,u^*, Z.

Algorithm 6.3 Adapting the Relation to the Amount of Columns Score

```
#In the following $u$ is the relation to be sold and $Y$ the according set of attributes
function AdaptColumns($y_t, u^*, Z$)
    $n := 0$
    $Y := []$                          ▷ An empty list, to be filled with attributes.
    while $n{+}{+} < y_t$ do
        $Y += shift(Z)$                ▷ Remove the first element in Z and add it to Y.
    end while
    $u := \pi_Y(u^*)$
    return $u, Y$
end function
```

Timeliness does not require an algorithm as it is concerned with delayed delivery. However, it requires some calculus, presented in the following. It was defined in Equation 6.16 as

$$tim(u) = \frac{\sum\limits_{\mu \in u} t(\mu[LastUpdated^*], v^*)}{|u|} \tag{6.54}$$

In order to further analyse it regarding the quality score, Equation 6.15 hast to be plugged in to result in:

$$tim(u) = \frac{\sum\limits_{\mu \in u} max \left\{0, 1 - \frac{DeliveryTime - \mu[LastUpdated^*]}{v^*}\right\}}{|u|} \tag{6.55}$$

For better readability, DeliveryTime will be denoted as DT and LU shall represent $\mu[LastUpdated^*]$. Furthermore, the max function can be omitted supposing that the target score $t_t = \frac{x}{x_{\max}}$ is positive. Additionally $|u|$ will be represented by n. Thus:

$$tim(u) = \frac{\sum\limits_{\mu \in u} 1 - \frac{DT - LU}{v^*}}{n} \tag{6.56}$$

Plugging in a target value t_t:

$$t_t \overset{!}{\geq} \frac{\sum\limits_{\mu \in u} 1 - \frac{DT - LU}{v^*}}{n} \tag{6.57}$$

$$t_t \times n \geq \sum_{\mu \in u} 1 - \sum_{\mu \in u} \frac{DT - LU}{v^*} \tag{6.58}$$

$$t_t \times n \times v^* \geq n \times v^* - \sum_{\mu \in u} DT - LU \tag{6.59}$$

Given that only LU is variable:

$$t_t \times n \times v^* \geq n \times v^* - \left(n \times DT - \sum_{\mu \in u} LU \right) \tag{6.60}$$

$$t_t \times n \times v^* - n \times v^* \geq -n \times DT + \sum_{\mu \in u} LU \tag{6.61}$$

$$\sum_{\mu \in u} LU \leq n \times v^* \times (t_t - 1) + n \times DT \tag{6.62}$$

$$\frac{1}{n} \times \sum_{\mu \in u} LU \leq v^* \times (t_t - 1) + DT \tag{6.63}$$

Equation 6.63 shows what the average timeliness depending on the target value t_t should be and could also be denoted as:

$$AvgLU(t) \leq v^* \times (t_t - 1) + DT \tag{6.64}$$

Alternatively:

$$LU_t \leq v^* \times (t_t - 1) + DT \tag{6.65}$$

The delivery time will always be the current time. Thus, it will be represented by the variable *now*, which will be replaced by the current timestamp upon query time. This allows for further modification resulting in:

$$LU_t \leq now - v^* \times (1 - t_t) \tag{6.66}$$

Introducing a delay function:

$$d(v^*, t_t) := v^* \times (1 - t_t) \tag{6.67}$$

Results in:

$$LU_t \leq now - d(v^*, t_t) \tag{6.68}$$

On first sight one might require each data set to have an average timeliness not greater than LU_t. However, using the overall average of a data set is slightly problematic, as this allows the selection of data that is very old together with very fresh data and then only use the fresh data. To avoid this, the timeliness of any record is required to be not greater than LU_t. In this way it is ensured that records with a timeliness score worse than or equal to what has been paid for are delivered. In practical terms customers do query a view u on u^* that is defined as:

$$u = \sigma_{\mu[LastUpdated^*] \leq now - d(v, t_t)}(u^*) \tag{6.69}$$

In this model it is important that when records are updated, the original record is kept so that customers can still access the older record rather than receiving an empty result set.

6.5.6. Fair Knapsack Pricing Example

This section illustrates the custom-tailoring of a data product using the previously described fair knapsack pricing model, supposing that a customer has opted to buy data from provider *A* from Section 6.4.4. The data presented in Table 6.2, represent the data before it is modified, i. e., u^*. The advertised price P remains £ 1,200.00 but the customer is only willing to pay $W = 1{,}000.00$. Again, for reasons of clarity and comprehensibility this section investigates only three quality attributes, namely:

$$\begin{aligned}\mathsf{V} &= \{\mathit{Timeliness}(q_1), \mathit{Amount\ of\ Data\ (Columns)}(q_2)\}\\ \mathsf{G} &= \{\mathit{Customer\ Service}(q_3)\}\end{aligned}$$

For *Customer Service*, the provider offers three service levels:

1. E-mail support with a 48 hour response guarantee
2. Telephone support 9 to 5 and 24 hours response time e-mail support
3. Telephone and e-mail support 24 / 7

Furthermore, at quality level 0 no support is provided. The customer-provided preferences for the different quality criteria are: $\omega = (0.35, 0.5, 0.15)$ and the provider specifies $\kappa = (0.5{,}0.3{,}0.2)$. In the following, the index i (rows) refers to quality criteria and the index j (columns) refers to utility levels. Based on this the utility matrix b, the weight matrix w as well as the incremental utility matrix $\tilde{b}$ and the incremental weight matrix $\tilde{w}$ can be calculated. Eventually, this can be used to arrive at the incremental efficiency $\tilde{e}$:

$$b = \begin{pmatrix} 0.35 & 0.7 & 1.05 & 1.4 & 1.75 & 2.1 & 2.45 & 2.8 & 3.15 & 3.5 \\ 0.5 & 1 & 1.5 & 2 & 2.5 & 3 & 3.5 & 4 & 4.5 & 5 \\ & & 0.45 & & & 0.9 & & & 1.35 & \end{pmatrix}$$

$$w = \begin{pmatrix} 6 & 24 & 54 & 96 & 150 & 216 & 294 & 384 & 486 & 600 \\ 3.6 & 14.4 & 32.4 & 57.6 & 90 & 129.6 & 176.4 & 230.4 & 291.6 & 360 \\ & & 26.\bar{6} & & & 106.\bar{6} & & & 240 & \end{pmatrix}$$

$$\tilde{b} = \begin{pmatrix} 0.35 & 0.35 & 0.35 & 0.35 & 0.35 & 0.35 & 0.35 & 0.35 & 0.35 & 0.35 \\ 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ & & 0.45 & & & 0.45 & & & 0.45 & \end{pmatrix}$$

$$\tilde{w} = \begin{pmatrix} 6 & 18 & 30 & 42 & 54 & 66 & 78 & 90 & 102 & 114 \\ 3.6 & 10.8 & 18 & 25.2 & 32.4 & 39.6 & 46.8 & 54 & 61.2 & 68.4 \\ & & 26.\bar{6} & & & 80 & & 133.\bar{3} & & \end{pmatrix}$$

$$\tilde{e} = \begin{pmatrix} 0.0729 & 0.0194 & 0.0117 & 0.0083 & 0.0065 & 0.0053 & 0.0045 & 0.0039 & 0.0034 & 0.0031 \\ 0.1389 & 0.0463 & 0.0278 & 0.0198 & 0.0154 & 0.0126 & 0.0107 & 0.0093 & 0.0082 & 0.0073 \\ & & 0.0169 & & & 0.0056 & & & 0.0033 & \end{pmatrix}$$

This results in the following ordered list of $\tilde{e}_{ij}$. Nota bene, $\tilde{e}_{1,1}$,$\tilde{e}_{21}$, and $\tilde{e}_{31}$ are missing because they are used to initialise the knapsack.

$$\begin{aligned} &\{\tilde{e}_{2,2} = 0.0463, \quad \tilde{e}_{2,3} = 0.0278, \quad \tilde{e}_{2,4} = 0.0198, \quad \tilde{e}_{1,2} = 0.0194, \quad \tilde{e}_{2,5} = 0.0154, \quad \tilde{e}_{2,6} = 0.0126, \\ &\tilde{e}_{1,3} = 0.0117, \quad \tilde{e}_{2,7} = 0.0107, \quad \tilde{e}_{2,8} = 0.0093, \quad \tilde{e}_{1,4} = 0.0083, \quad \tilde{e}_{2,9} = 0.0082, \quad \tilde{e}_{2,10} = 0.0073, \\ &\tilde{e}_{1,5} = 0.0065, \quad \tilde{e}_{3,2} = 0.0056, \quad \tilde{e}_{1,6} = 0.0053, \quad \tilde{e}_{1,7} = 0.0045, \quad \tilde{e}_{1,8} = 0.0039, \quad \tilde{e}_{1,9} = 0.0034, \\ &\tilde{e}_{3,3} = 0.0033, \quad \tilde{e}_{1,10} = 0.0031\} \end{aligned}$$

The knapsack is initialised as follows:

$$\begin{aligned} x_{i1} &:= 1 \text{ f. a. } i, 1 \leq i \leq n_q \\ z &:= \sum_{i=1}^{n} b_{i1} = 1.3 \\ \bar{c} &:= W - \sum_{i=1}^{n} w_{i1} = 963.7\bar{3} \end{aligned}$$

Then, the MCKPP is solved using Algorithm 6.1. The sequence of $\bar{c}$, and the selected $\tilde{e}_{ij}$ in each step is shown in Table 6.7. Again, it has to be pointed out that this is just for illustrative purposes as better but less easy to comprehend algorithms exist to solve this.

Table 6.7.: Development of $\bar{c}$ After Each Processing Step.

Iteration	1	2	3	4	5	6	7	8	9	10
Selected $\tilde{e}_{ij}$	$\tilde{e}_{2,2}$	$\tilde{e}_{2,3}$	$\tilde{e}_{2,4}$	$\tilde{e}_{1,2}$	$\tilde{e}_{2,5}$	$\tilde{e}_{2,6}$	$\tilde{e}_{1,3}$	$\tilde{e}_{2,7}$	$\tilde{e}_{2,8}$	$\tilde{e}_{1,4}$
$\bar{c}$	952,93	934,93	909,73	891,73	859,33	819,73	789,73	742,93	688,93	646,93
Iteration	11	12	13	14	15	16	17	18	19	20
Selected $\tilde{e}_{ij}$	$\tilde{e}_{2,9}$	$\tilde{e}_{2,10}$	$\tilde{e}_{1,5}$	$\tilde{e}_{3,2}$	$\tilde{e}_{1,6}$	$\tilde{e}_{1,7}$	$\tilde{e}_{1,8}$	$\tilde{e}_{1,9}$	$\tilde{e}_{3,3}$	$\tilde{e}_{1,10}$
$\bar{c}$	585,73	517,33	463,33	383,33	317,33	239,33	149,33	47,33	-86,00	-200,00

As can be seen in Table 6.7, the items that do not fit in the knapsack are $\tilde{e}_{3,3}$ and $\tilde{e}_{1,10}$[4]. By implication, this means that items $x_{1,9}, x_{2,10}$ and $x_{3,2}$ are chosen, i. e., *Timeliness* at level 9, *Amount of Data* at full level and *Customer Support* at level 2. Thus, *Customer Support* is served as 9 to 5 telephone support and 24 hours response time e-mail support. Nevertheless, this is a contractual category that does not influence the data. Since *Amount of Data* is served at full level it also does not influence the data. Therefore, $u = u^*$ because no modifications to the data have to be made. However, given that time restrictions apply, in that the data is not queryable as soon as it is entered into the system, a view has to be derived at. The volatility of 24 hours is known from Section 6.4.4, and the target quality score of $t_t = \frac{81}{100} = 0.81$ is a result of the optimisation process, based on the quality level 9. Now the delay may be calculated using Equation 6.67:

$$d(v, t_t) = v^* \times (1 - t_t)$$
$$d = 24 * 0{,}19 = 4.45$$

Having identified a delay of 4.45 hours and supposing the constant *now* is expressed in hours, too, the final delivery view can be expressed as:

$$\sigma_{\mu[LastUpdated^*] \leq now - 4.56}(u)$$

6.6. Data Marketplaces Conclusions and Future Work

This chapter has demonstrated how data quality on data marketplaces can be measured and how these measures can be applied when pricing relational data goods. Firstly, a customer's point of view has been taken and a quality scoring model has been developed that can support customers when choosing a data provider to buy from, indifferent of whether providers operate on the same or on different data marketplaces. Furthermore, this score can be used by data providers to learn about their customers' preferences as well as their standing compared to the competition. Implementing a learning algorithm, this can provide valuable insight and could potentially be a source of revenue for the data marketplace operator as well as an advantage over their competitors. Consequently, implementing this framework and extending it to gain further insights can be seen as the main tasks for future research regarding the quality scoring model.

The second part of this chapter then presented a model that allows providers to apply a NYOP scheme for data. This enables them to tap the willingness to

[4] Note that the split has not been considered to keep the example simple.

pay of customers that would otherwise not buy their relational data product. By adjusting the quality it can be ensured that a customer gets exactly what they pay for. In fact, using this model providers do not have to specify a price publicly at all. They also could use an internal price P and still apply the same pricing model. While this would require users to bid exactly the price they are willing to pay it lacks transparency. An alternative would be advertising a price P^p greater than P publicly. This would result in additional profits from customers paying a price W for which $P \leq W \leq P^p$ holds. This work has excluded the issue of potential cannibalisation, i. e., that customers who would have bought expensive products switch to a cheaper version if it becomes available. This is an organisational aspect subject to future research. Furthermore, it should be evaluated whether this pricing model is perceived as fair as this is an important issue when pricing [Rei02; NRRS05]. To this end, an alternative pricing model could be experimented with, in which not all prices are calculated automatically but users are provided with feedback regarding the actual quality levels while entering their prices and preferences. In this case they would know what quality level they receive and can experiment with input variables. This might also increase the perceived fairness. Furthermore, using statistical analyses on the bid prices, data providers can learn what value customers attribute to their offerings.

In this context, truth revelation might be an issue [NRRS05]; the question remains if customers can actually cheat the system by not mentioning their true preference. At this point, no formal proof can be provided but the argument is made that if the used algorithm indeed delivers optimal results, then, customers cannot cheat the system as it delivers a custom-tailored product for exactly the suggested price. Depending on the number and size of the quality-based utility levels, there might be a little room to minimise the residual capacity $\bar{c}$ which might still occur; however, this is ineluctable.

Regarding the used weighting function it can be summarised that using the inverted utility function, the average price across all levels is smaller than in the linear case, which speaks in favour of the latter model from a customer's perspective. After all, for the provider it is not that important what product is actually delivered as the cost of creating it are marginal and every sold unit adds to their profits. What is more important is that customers get a fair discount for their scarifies of quality. This is achieved by either of the two.

Developing a quality-based pricing model, it has been shown that pricing on a data marketplace can be expressed as a *Multiple-Choice Knapsack Problem.* An implementation is an important future work in order to evaluate the algorithm presented in Section 6.5.4 in the context of pricing. In this regard, it is particularly interesting from which number of quality levels and quality scores the algorithms becomes inefficient – if at all. Conducting such experiments, it can be verified if the assumption that using the proposed solution to the MCKPP the quality level can be determined in fractions of seconds can be held. Furthermore, some work has to be invested into the question of how to actually create the required relational data products on the spot as this might also take a considerable amount of time. For the methods presented in Section 6.5.5 it can be said that run time is negligible as every record has to be processed at most once, yielding $\mathcal{O}(n)$, where n denotes the number of requested tuples. However, other adaptations might be more difficult.

Thus far it has not been explicitly stated but MCKPP can be applied to multiple vendors as well. In this case not the scores of one provider have to be mapped to the quality levels but only the best scores of all providers. This results in a problem scenario, where some providers might not be able to deliver all quality criteria. Solving the MCKPP for all providers given a customer's query and preference, it can also be determined which provider offers the best product for a customer.

To summarise, Figure 6.10 shows all components that influence the MCKPP, namely *Quality Criteria*, *Customer Info* comprising the preference vector ω and a bid price W, *Provider Info* comprising a weighting vector κ and an ask price P, a *versioning function* b, a *weighting function* $w(x)$ or $w(b)$, and a *Quality Adaptation Algorithm* for each *Quality Criterion.* It is a distinct feature of this model that all components can be adjusted to match the needs of data marketplace providers as well as the needs of data providers. Formally, this chapter is summarised in Theorem 6.3.

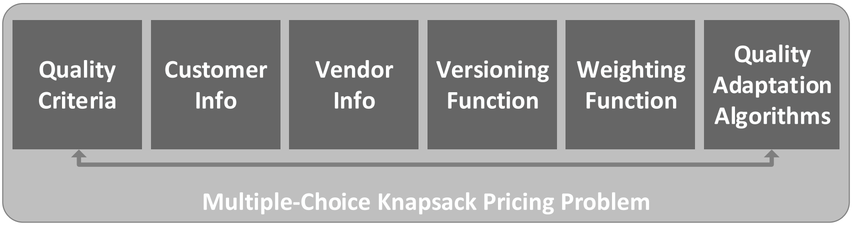


Figure 6.10.: All Components of the Multiple-Choice Knapsack Pricing Problem.

Theorem 6.3. Given all of the features presented in Figure 6.10 and supposing the *Quality Adaptation Algorithms* have a complexity less than that of the MCKPP, a custom-tailored data product can be created in pseudo-polynomial time.

Sketch of Proof.

1. Translate inputs into MCKPP as demonstrated in Sections 6.5.1 to 6.5.3.
2. Use PISINGER's [Pis95] algorithm to solve the according MCKP.
3. Apply MCKP result to create custom-tailored data products using *Quality Adaptation Algorithms*.

7. Conclusion

In this day and age, a world without digital information processing is inconceivable. Living in an information-driven economy and society, data and the information that is drawn from it are used for leisure tasks, such as planning an elaborate holiday trip as well as for steering businesses, with data-driven decision making being on the rise. In this environment, the quality of information, manifested in availability, correctness, accessibility and other facets, is more important than ever before.

As the amount of data available online is growing at a tremendous speed, leading to a point at which information is omnipresent, its full potential cannot yet be harnessed. This is owing to the fact that the sheer amount of data available makes it distinctly more difficult to locate and retrieve those pieces of information most relevant to a specific use case. While mainstream information needs are well-satisfied by current algorithmic search engines, high-quality information needs in niche domains are difficult to satisfy.

Given the recognition that information can be business-critical, it is hardly surprising that data, which enables information, has been discovered as a valuable good. Consequently, data is now being traded on data marketplaces which act as an intermediary between data providers and users of data. Despite the emergence of data marketplaces and the recognition that data indeed has an inherent value, pricing of data has remained difficult as there has been little understanding of how value can be attributed to data. In this context, the main problem has been the fact that various customers attribute different value to data goods depending, among other things, on the quality of the data products.

This thesis has, therefore, addressed the two aspects high-quality information provisioning and quality-based data pricing, to advance the understanding of the important resources that are data and information. In the course of this conclusion, the main contributions of this thesis will be recapitulated, before this work is concluded by giving an outlook on possible future developments of Web information provisioning and data pricing.

7.1. Main Contributions

The main contributions of this work in the adjacent fields of high-quality information provisioning and quality-based data pricing will be outlined in the following two subsections.

7.1.1. High-Quality Web Information Provisioning

In the first part of this thesis, the topic of high-quality information provisioning on the Web has been addressed by introducing a complex process that helps to deliver comprehensive information in areas where using state of the art Web search engines can be tedious and time-consuming. For such scenarios, the WiPo process offers a unique and innovative approach to information provisioning compared to traditional Web search engines, including the additional feature of offline availability. This process, which has been prototypically implemented, enables a topic-centric information service, tailored precisely to a user's needs and budget.

To achieve this, information is sourced from the Internet and undergoes comprehensive data mining as well as curation in order to result in *curated documents*, stored in a *curated database*. Querying the WiPo service, users provide keywords, supply a list of relevant links and have the option to upload files through an easy-to-use GUI. Similar to traditional search engines, a user's query is run against an existing document index based on the *curated database*. In contrast to standard search indices, the new approach gives a WiPo operator full control over the contained content, ensuring quality, and dynamically reacts to the users' precise information needs. In addition, WiPo allows for sophisticated personalisation, enabling users to configure their own explicit search profile which helps with finding relevant documents and improving ranking. Moreover, the WiPo service can even be configured so that searches do not only include (semi-) public data but also private documents of users or proprietary information of a WiPo operator. Additionally, users are informed whenever new information is available for them. The final results are presented in an integrated manner that makes it easy to see the original sources as well relations between results. The service is available on all kinds of devices including mobile devices, enabling users to carry their information repository with them and keep relevant parts offline – hence the name: Web in your Pocket (WiPo). The WiPo approach is characterised by three unique features, elaborated on next.

Private Sources WiPo is unique in that it does not rely solely on public sources but allows for the inclusion of private sources, very much in the spirit of the *Memex* [Bus45]. In that sense, WiPo can also be seen as a portable *Memex* for the Internet age.

Data Curation High-quality information is achieved by exploiting curation, which, in the context of this thesis, means the long-term selection, cleansing, enrichment, preservation, and assembly of data and information and their respective sources. Curation typically requires a varying degree of manual labour – which may be provided either by a number of selected individuals or an anonymous crowd – and should be supported by algorithms in order to decrease the manual part of a curators' workload.

Offline Availability and Push-Paradigm Despite allowing for persisting content on a mobile device, the biggest difference between the WiPo approach and established search engines is a paradigm shift from *pull* (i. e., users need to request information) to *push* (i. e., information is automatically delivered to users).

The description of the WiPo process and its implementation has been complemented by an extensive discussion of four use cases (LandSAR, tourism, healthcare, agricultural business). This has made it evident that while use cases for WiPo can be inherently different, they can be gauged by the same sets of attributes. Moreover, it stands to reasons that the developed comparison framework is generally applicable to WiPo use cases and can serve as reference to new use cases being implemented.

In summary, it has been shown that WiPo, as an information service and single point of truth, can be particularly beneficial if a number of sources for a well-specified domain have to be integrated, kept up-to-date, and possibly persisted offline on a mobile device. Thus, it can be stated that the aim formulated in the Introduction of creating a software artefact that answers a user's domain-specific informational queries levering curation to satisfy high-quality information needs has been reached.

7.1.2. Quality-Based Relational Data Pricing

In the context of quality-based data pricing, a novel scoring framework for measuring data quality on data marketplaces has been developed. This scoring model comprises – in its current form – 19 quality categories, the majority of which (namely 12) can be assessed computationally. The most prominent examples are *accuracy, completeness,* and *timeliness.* Based on quality scores for each category, an overall quality score can be calculated for relational data offerings. This quality scoring model has five distinct features:

1. It is customisable to the specific preference of a buyer for certain quality criteria.
2. It can support customers with an intra-marketplace comparison when choosing a data provider to purchase their data from.
3. It allows for the calculation of a quality-for-money-score.
4. It can be supplied by data marketplace providers to enable intra-marketplace comparisons of different data offerings if marketplace providers use the same scores and interchange scoring data.
5. It can be used by data providers to learn about their customer's preferences as well as compare their products to competitors' products.

This quality scoring model – which is to the best of the author's knowledge the first specifically targeted at (price) comparisons of relational data products – alone can already provide guidance on the relative value of a relational data product regarding different quality criteria. In extension to this, however, an advanced pricing model was presented – based on the preliminary work of data quality criteria – that allows providers to apply a Name Your Own Price (NYOP) scheme for relational data. It enables data vendors to tap into their customers willingness to pay while, at the same time, allowing them to provide custom-tailored relational data products to their clients by adjusting the quality. In this way, it can be ensured that customers receive exactly the product they pay for.

As a means for dynamic pricing, the overall pricing problem has been transformed into a Multiple-Choice Knapsack Problem (MCKP), in the context of data pricing referred to as Multiple-Choice Knapsack Pricing Problem (MCKPP), consisting of the following seven components:

1. *Quality scores*, which allow for measuring the quality of the relational data product
2. *Quality modification algorithms and methods*, which ensure that the quality of a data product can be adjusted to create different versions
3. *Versioning function*, which translates different quality scores into product versions

4. *Customer information*, i. e., the bid price and a preference vector indicating how the bid price is to be distributed across different quality versions using the customised versioning function
5. *Weighting function*, which assigns weights to different versions
6. *Vendor information*, i. e., the ask price and a cost vector indicating how the ask price is to be distributed across different quality versions using the cost function
7. A suitable *knapsack algorithm*, which allows the solving of the pricing problem in a reasonable amount of time

Given all these factors, it could be shown that solving the MCKPP is possible within a runtime of $\mathcal{O}(n_t + W \sum_{q_i \in C} m_{l\,i})$, where q_i denotes a quality score, $m_{l\,i}$ denotes the according number of quality levels, n_t denotes the total number of elements (i. e., the sum of all versions over all quality levels), W denotes the weight limit (i. e., the bid price), and C denotes a core set of quality criteria for which the algorithm is solved first and then gradually extended. This results in a linear solution time for a minimal core and pseudo-polynomial time for larger cores, i. e., exponential behaviour for exponentially-large inputs. Computational experiments suggest that the MCKPP – for expected problem sizes of up to 100 quality scores and levels – will be solvable to optimality in less than a second. Notwithstanding this, there is also the possibility to ϵ-approximate a solution in $\mathcal{O}(n_t \log n_t + \frac{n_t n_q}{\epsilon})$, where n_q denotes the number of quality criteria. However, both complexity classes do not account for the time needed to calculate the data product but only the according levels. While this is generally an issue of future research, this work presented initial adaptation algorithms for basic quality scores which have a complexity smaller than that of the MCKPP and are thus negligible.

From this it can be concluded that the aim of the second part of this thesis – i. e., to develop a method which supports trading of data by allowing customers to acquire a custom-tailored relational data product at a fair price – has been reached. The presented pricing framework can be seen as a building block to enable trading of quality data based on a novel pricing mechanism which had not existed so far. The pricing model is generic and very flexible in nature; it consists of many components and can be easily extended to comprise additional (or different) quality scores, attribute cost and prices, or solve the pricing problem to different degrees of optimality. Therefore, it can be used by various data marketplace operators with different requirements for their quality-based pricing scheme.

7.2. Outlook

As the terms *information society* and *information economy* suggest, information is a determining factor in modern economy and society, respectively. As such, it is most likely that access to high-quality information will become of even higher importance for businesses and individuals. This thesis has advanced two interrelated fields in this context, namely that of high-quality Web information provisioning and that of data pricing.

For WiPo it would be most interesting to continue the started design science research process by developing the prototype further and applying it to one of the use cases described; the LandSAR case is the most promising as initial steps in this direction have already been taken. Furthermore, exploring the trend of crowd curation in-depth is an interesting field of study. In this context the development of a readily usable mobile app is particularly interesting.

If the current trend continues, it is foreseeable that the demand for high-quality information will last and probably increase. Consequently, the Web in your Pocket (WiPo) will continue to be a useful application for niche domains. While it is likely that ubiquitous broadband availability will be reached over time, which implies that offline functionality may become less important in the coming years, the curation component will continue to provide a benefit to WiPo users. However, the success of a WiPo installation will largely depend on two main factors: 1) a very good usability of the system that motivates users to adopt the WiPo approach instead of spending time with a well-known and easy-to-use general-purpose search engine that nevertheless only provides mediocre result quality and 2) the employed experts, as it is their responsibility to ensure that high-quality standards are fulfilled. If these requirements are met, WiPo can be a valuable tool for Web information provisioning.

As the economy becomes increasingly data-driven, it is very likely that trading of data will become more important over time. In this regard, this thesis has laid the basis for facilitating quality-based data trading in the near future. Implementing this approach on a real data marketplace is an important next step. This implementation may then serve as the basis for practical experiments to address such questions as what queries to allow, what quality preferences are common, how is the proposed model perceived by customers and data providers, how does it perform when comparing providers and can the assumption be held that the proposed solution can be solved efficiently in practise.

Nevertheless, data is only valuable if it can be used for a purpose, which requires companies to hire expert analysts, lately referred to as data scientists, who are capable of interpreting and making use of the data available. To date, areas for which expert knowledge exist and, consequently, data is being traded already, include financial data, weather data, and marketing data. However, many more data types exist, for most of which no business case has been identified yet. As soon as companies start to analyse new data types to gain competitive advantages, a market for such data will emerge and the trading of new data types will start. When this happens, the quality-based pricing model developed in this work is at hand.

All things considered, the models presented in this thesis can provide a significant benefit to their users and may serve as a basis for future developments in their respective domains.

References

Literature References

[Aal98] W. van der Aalst. 'The application of Petri nets to workflow management'. In: *Journal of circuits, systems, and computers* 8.01 (1998), pp. 21–66.

[ABD06] E. Agichtein, E. Brill, and S. Dumais. 'Improving web search ranking by incorporating user behavior information'. In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Ed. by S. Dumais. New York, NY, USA: Association for Computing Machinery, 2006, pp. 19–26.

[Ada95] D. Adams. *The Hitchhiker's Guide to the Galaxy*. Ballantine Books. Del Rey/Ballantine Books, 1995.

[Agu67] F. Aguilar. *Scanning the business environment*. Studies of the modern corporation. Macmillan, 1967.

[AH04] W. van der Aalst and K. van Hee. *Workflow Management: Models, Methods, and Systems*. Cooperative Information Systems. MIT Press, 2004.

[AHV95] S. Abiteboul, R. Hull, and V. Vianu. *Foundations Of Databases*. New York, U. S.: Addison Wesley, 1995.

[And07] C. Anderson. *The Long Tail: How Endless Choice is Creating Unlimited Demand*. Random House Business Books. London, UK: Random House Business, 2007.

[Arr62] K. J. Arrow. 'Economic welfare and the allocation of resources for invention'. In: *The Rate and Direction of Inventive Activity: Economic and Social Factors*. Ed. by H. M. Groves. Princeton University Press, 1962, pp. 609–626.

[AS11] W. van der Aalst and C. Stahl. *Modeling Business Processes: A Petri Net-Oriented Approach*. Cooperative information systems. MIT Press, 2011.

[BAB12] A. Beloglazov, J. Abawajy, and R. Buyya. 'Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing'. In: *Future Generation Computer Systems* 28.5 (2012), pp. 755–768.

[Bac00] J. Bachmann. *Der Information-Broker: Informationen suchen, sichten, präsentieren*. Die Integrata-Qualifizierung. Addison-Wesley, 2000.

[BAHT11] J. Baliga, R. Ayre, K. Hinton, and R. Tucker. 'Green Cloud Computing: Balancing Energy in Processing, Storage, and Transport'. In: *Proceedings of the IEEE* 99.1 (Jan. 2011), pp. 149–167.

[Bak97] J. Y. Bakos. 'Reducing Buyer Search Costs: Implications for Electronic Marketplaces'. In: *Management Science* 43.12 (1997), pp. 1676–1692.

[Bak98] Y. Bakos. 'The Emerging Role of Electronic Marketplaces on the Internet'. In: *Communications of the ACM* 41.8 (1998), pp. 35–42.

[Bal12a] R. Baldegger. *Management in a Dynamic Environment: Concepts, Methods and Tools*. Gabler Verlag, 2012.

[Bal12b] W.-T. Balke. 'Introduction to Information Extraction: Basic Notions and Current Trends'. In: *Datenbank-Spektrum* 12.2 (2012), pp. 81–88.

[Bar00] R. Barnes. 'Cloistered Bookworms in the Chicken-Coop of the Muses: The Ancient Library of Alexandria'. In: R. MacLeod. *The Library of Alexandria: Centre of Learning in the Ancient World*. London, UK: I. B. Tauris, 2000, pp. 61–77.

[Bas90] R. Basch. 'Measuring the quality of the data: Report on the fourth annual SCOUG retreat'. In: *Database Searcher* 6.8 (1990), pp. 18–24.

[Bat90] B. J. Bates. 'Information as an economic good: A re-evaluation of theoretical approaches'. In: *Mediation, information, and communication – Information and behavior*. Ed. by B. D. Ruben and L. A. Lievrouw. Transaction Publishers, 1990, pp. 379–394.

[BB07] A. Bryman and E. Bell. *Business Research Methods*. Oxford University Press, 2007.

[BBB+11] R. Baeza-Yates, P. Boldi, B. M. Bozz, S. Ceri, and G. Pasi. 'Trends in Search Interaction'. In: *Search computing*. Ed. by S. Ceri and M. Brambilla. Berlin: Springer, 2011, pp. 26–32.

[BBC+10] A. Bozzon, M. Brambilla, S. Ceri, F. Corcoglioniti, and N. Gatti. 'Chapter 14: Building Search Computing Applications'. In: *Search Computing*. Ed. by S. Ceri and M. Brambilla. Vol. 5950. Lecture Notes in Computer Science. Berlin and Heidelberg: Springer-Verlag, 2010, pp. 268–290.

[BBC+11] A. Bozzon, M. Brambilla, S. Ceri, P. Fraternali, and S. Vadacca. 'Exploratory search in multi-domain information spaces with liquid query'. In: *Proceedings of the 20th international conference companion on World wide web*. WWW '11. Hyderabad, India: ACM, 2011, pp. 189–192.

[BBCF10] A. Bozzon, M. Brambilla, S. Ceri, and Fraternali P. 'Chapter 13 Liquid query: Multi-domain exploratory search on the web'. In: *Search Computing*. Ed. by S. Ceri and M. Brambilla. Vol. 5950. Lecture Notes in Computer Science. Berlin and Heidelberg: Springer Berlin Heidelberg and Springer-Verlag Berlin Heidelberg, 2010.

[BC01a] H. K. Bhargava and V. Choudhary. ‘Information goods and vertical differentiation’. In: *Journal of Management Information Systems* 18.2 (2001), pp. 89–106.

[BC01b] H. K. Bhargava and V. Choudhary. ‘Second-degree price discrimination for information goods under nonlinear utility functions’. In: *System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on*. IEEE. 2001, pp. 1–6.

[BC08] H. K. Bhargava and V. Choudhary. ‘Research Note - When Is Versioning Optimal for Information Goods?’ In: *Management Science* 54.5 (2008), pp. 1029–1035.

[BCCV06] P. Buneman, A. Chapman, J. Cheney, and S. Vansummeren. ‘A provenance model for manually curated data’. In: *Proceedings of the 2006 international conference on Provenance and Annotation of Data*. IPAW’06. Chicago, IL: Springer-Verlag, 2006, pp. 162–170.

[BCDP11] T. Buganza, M. Corubolo, E. Della Valle, and E. Pellizzoni. ‘Analysis of Business Models for Search Computing’. In: *Search computing*. Ed. by S. Ceri and M. Brambilla. Berlin: Springer, 2011, pp. 256–271.

[BCGH10] R. Baumgartner, A. Campi, G. Gottlob, and M. Herzog. ‘Chapter 6: Web Data Extraction for Service Creation’. In: *Search Computing*. Ed. by S. Ceri and M. Brambilla. Vol. 5950. Lecture Notes in Computer Science. Berlin and Heidelberg: Springer-Verlag, 2010, pp. 94–113.

[BEP+08] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. ‘Freebase: a collaboratively created graph database for structuring human knowledge’. In: *Proc. 2008 ACM SIGMOD Int. Conf. on Management of Data*. Vancouver, Canada, 2008, pp. 1247–1250.

[Ber01] M. K. Bergman. ‘White Paper: The Deep Web: Surfacing Hidden Value’. In: *Journal of Electronic Publishing* 7.1 (2001).

[BESL13] D. Bunker, C. Ehnis, P. Seltsikas, and L. Levine. ‘Crisis Management and Social Media: Assuring Effective Information Governance for Long Term Social Sustainability’. In: *IEEE International Conference on Technologies for Homeland Security (HST)*. Nov. 2013, pp. 246–251.

[BF12] B. Bensoussan and C. Fleisher. *Analysis Without Paralysis: 12 Tools to Make Better Strategic Decisions*. Pearson Education, 2012.

[BFS03] P. Baldi, P. Frasconi, and P. Smyth. *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*. Wiley Series in Probability and Statistics. Wiley, 2003.

[BGM+09] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom. ‘Swoosh: A Generic Approach to Entity Resolution’. In: *The VLDB Journal* 18.1 (2009), pp. 255–276.

[BH12] J. Barney and W. Hesterly. *Strategic Management and Competitive Advantage: Concepts and Cases.* Pearson, 2012.

[BH99] F. Bieberbach and M. Hermann. 'Die Substitution von Dienstleistungen durch Informationsprodukte auf elektronischen Märkten'. In: *Electronic Business Engineering.* Ed. by M. Nüttgens and A.-W. Scheer. Springer, 1999, pp. 67–81.

[BHK+13] M. Balazinska, B. Howe, P. Koutris, D. Suciu, and P. Upadhyaya. 'A Discussion on Pricing Relational Data'. In: *In Search of Elegance in the Theory and Practice of Computation.* Ed. by V. Tannen, L. Wong, L. Libkin, W. Fan, W.-C. Tan, and M. Fourman. Vol. 8000. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 167–173.

[BHK11] E. Brynjolfsson, L. M. Hitt, and H. H. Kim. *Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?* Tech. rep. Social Science Research Network, 2011.

[BHL01] T. Berners-Lee, J. Hendler, and O. Lassila. 'The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities'. In: *Scientific American* (2001).

[BHS11] M. Balazinska, B. Howe, and D. Suciu. 'Data Markets in the Cloud: An Opportunity for the Database Community'. In: *PVLDB* 4.12 (2011), pp. 1482–1485.

[BHW11] M. Burghardt, M. Heckner, and Wolff Christian. 'Social Search'. In: *Handbuch Internet-Suchmaschinen 2.* Ed. by D. Lewandowski. Heidelberg: AKA, Akad. Verl.-Ges., 2011, pp. 3–28.

[BKM+00] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. 'Graph Structure in the Web'. In: *Proceedings of the 9th International World Wide Web Conference on Computer Networks : The International Journal of Computer and Telecommunications Netowrking.* Amsterdam, The Netherlands: North-Holland Publishing Co., 2000, pp. 309–320.

[BKO+11] B. Bishop, A. Kiryakov, D. Ognyanov, I. Peikov, Z. Tashev, and R. Velkov. 'FactForge: a fast track to the web of data'. In: *Semantic Web* 2.2 (2011), pp. 157–166.

[BKW13] J. Baetge, H. Klönne, and C. Weber. 'Möglichkeiten und Grenzen einer objektivierten Spielerbewertung im Profifußball'. In: *KoR* 12.06 (2013), pp. 310–319.

[BL13] I. Blohm and J. M. Leimeister. 'Gamification: Design of IT-Based Enhancing Services for Motivational Support and Behavioral Change'. In: *Business Information System & Engineering (BISE)* 4.5 (2013), pp. 275–278.

[BMPW98] S. Brin, R. Motwani, L. Page, and T. Winograd. 'What can you do with a Web in your Pocket?' In: *Data Engineering Bulletin* 21 (1998), pp. 37–47.

[BN08] J. Bleiholder and F. Naumann. 'Data Fusion'. In: *ACM Comput. Surv.* 41.1 (2008).

[BP98] S. Brin and L. Page. 'The Anatomy of a Large-Scale Hypertextual Web Search Engine'. In: *Computer Networks and ISDN Systems* 30.1-7 (1998), pp. 107–117.

[BR10] R. Baeza-Yates and P. Raghavan. 'Chapter 2: Next Generation Web Search'. In: *Search Computing*. Ed. by S. Ceri and M. Brambilla. Vol. 5950. Lecture Notes in Computer Science. Berlin and Heidelberg: Springer-Verlag, 2010, pp. 11–23.

[BR11] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval: The Concepts and Technology Behind Search*. Addison Wesley Professional, 2011.

[Bra08] D. Brabham. 'Moving the crowd at iStockphoto: The composition of the crowd and motivations for participation in a crowdsourcing application'. In: *First Monday* 13.6 (2008).

[BS03] H. K. Bhargava and S. Sundaresan. 'Contingency Pricing for Information Goods and Services Under Industrywide Performance Standard'. In: *J. Manage. Inf. Syst.* 20.2 (2003).

[BS06] C. Batini and M. Scannapieca. *Data Quality: Concepts, Methodologies and Techniques*. Data-Centric Systems and Applications. Springer, 2006.

[BSS05] M. Bernhardt, M. Spann, and B. Skiera. 'Reverse pricing'. In: *Die Betriebswirtschaft* 65.1 (2005), pp. 104–107.

[BTF11] P. Bodenbenner, C. Tempich, and L. Feuerstein. *Detecon Opinion Paper: Turning Data into Profit - Success Factors in Data-Centric Business Models*. Tech. rep. Detecon Consulting, 2011.

[Bun10] D. Bunker. 'Information systems management (ISM): Repertoires of collaboration for community warning (CW) and emergency incident response (EIR)'. In: *IEEE International Conference on Technologies for Homeland Security (HST)*. Boston, United States, Nov. 2010, pp. 216–221.

[Bus45] V. Bush. 'As we may think'. In: *Atlantic Monthly* 176 (1945), pp. 101–108.

[BWPT98] D. Ballou, R. Wang, H. Pazer, and G. K. Tayi. 'Modeling Information Manufacturing Systems to Determine Information Product Quality'. In: *Manage. Sci.* 44.4 (1998), pp. 462–484.

[Cas10] M. Castells. *The Rise of The Network Society: The Information Age: Economy, Society and Culture*. 2nd edition. Information Age Series Vol. 1. Wiley, 2010.

[CBD99] S. Chakrabarti, M. v. Berg, and B. Dom. 'Focused crawling: a new approach to topic-specific Web resource discovery'. In: *Computer Networks* 31.11–16 (1999), pp. 1623–1640.

[CCG+10] A. Campi, S. Ceri, G. Gottlob, A. Maesani, and S. Ronchi. 'Chapter 9: Service Marts'. In: *Search Computing*. Ed. by S. Ceri and M. Brambilla. Berlin and Heidelberg: Springer-Verlag, 2010, pp. 163–187.

[Cer10] S. Ceri. 'Chapter 1: Serach Computing'. In: *Search Computing*. Ed. by S. Ceri and M. Brambilla. Vol. 5950. Lecture Notes in Computer Science. Berlin and Heidelberg: Springer Berlin Heidelberg and Springer-Verlag Berlin Heidelberg, 2010, pp. 3–10.

[CG03] J. Cho and H. Garcia-Molina. 'Estimating frequency of change'. In: *ACM Trans. Internet Technol.* 3.3 (2003), pp. 256–290.

[CH93] S. Christensen and N. Hansen. 'Coloured Petri nets extended with place capacities, test arcs and inhibitor arcs'. In: *Application and Theory of Petri Nets 1993*. Ed. by M. Ajmone Marsan. Vol. 691. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1993, pp. 186–205.

[Cha33] E. H. Chamberlin. *The Theory of Monopolistic Competition*. Cambridge, MA: Harvard University Press, 1933.

[CHL+04] K. C.-C. Chang, B. He, C. Li, M. Patel, and Z. Zhang. 'Structured databases on the web'. In: *ACM SIGMOD Record* 33.3 (2004), p. 61.

[CHM11] M. J. Cafarella, A. Halevy, and J. Madhavan. 'Structured data on the web'. In: *Communications of the ACM* 54.2 (2011), pp. 72–79.

[Cho08] G. S. Choudhury. 'Case Study in Data Curation at Johns Hopkins University'. In: *Library Trends* 57.2 (2008), pp. 211–220.

[Cho10] V. Choudhary. 'Use of Pricing Schemes for Differentiating Information Goods'. In: *Information Systems Research* 21.1 (2010), pp. 78–92.

[Chr97] R. B. Chris Steyaert. 'Group Methods in Organizational Analysis'. In: *Qualitative methods in organizational research. A practical guide*. Ed. by G. S. Catherine Cassell. SAGE Publications Ltd, 1997, pp. 123–146.

[Cod70] E. F. Codd. 'A Relational Model of Data for Large Shared Data Banks'. In: *Commun. ACM* 13.6 (1970), pp. 377–387.

[Cod90] E. F. Codd. *The Relational Model For Database Management: Version 2*. New York, U. S.: Addison Wesley, 1990.

[Com95] D. Comer. *The Internet book: Everything you need to know about computer networking and how the Internet works*. Englewood Cliffs, N.J.: Prentice-Hall International, 1995.

[COP12] M. J. Carey, N. Onose, and M. Petropoulos. 'Data services'. In: *Commun. ACM* 55.6 (2012), pp. 86–97.

[CPT10] J. Cadle, D. Paul, and P. Turner. *Business Analysis Techniques: 72 Essential Tools for Success*. British Comp Society Series. British Computer Society, 2010.

[Cro98] M. Crotty. *The Foundations of Social Research: Meaning and Perspective in the Research Process*. SAGE Publications, 1998.

[CWC14] C.-K. Chau, Q. Wang, and D.-M. Chiu. 'Economic viability of Paris Metro Pricing for digital services'. In: *ACM Transactions on Internet Technology (TOIT)* 14.2-3 (2014), p. 12.

[CY07] W.-L. Chang and S.-T. Yuan. 'An overview of information goods pricing'. In: *International Journal of Electronic Business* 5.3 (2007), pp. 294–314.

[CZW98] Y. Chen, Q. Zhu, and N. Wang. 'Query processing with quality control in the World Wide Web'. In: *World Wide Web* 1.4 (1998), pp. 241–255.

[Dav06] T. H. Davenport. 'Aus Daten Geld machen'. In: *Harvard Business Manager* 4 (2006), pp. 73–83.

[DKA09] D. Dash, V. Kantere, and A. Ailamaki. 'An Economic Model for Self-Tuned Cloud Caching'. In: *ICDE*. Ed. by Y. E. Ioannidis, D. L. Lee, and R. T. Ng. IEEE, 2009.

[DKW84] M. Dyer, N. Kayal, and J. Walker. 'A branch and bound algorithm for solving the multiple-choice knapsack problem'. In: *Journal of Computational and Applied Mathematics* 11.2 (1984), pp. 231–249.

[Dop09] P. Dopichaj. 'Ranking-Verfahren für Web-Suchmaschinen'. In: *Handbuch Internet-Suchmaschinen*. Ed. by D. Lewandowski. Heidelberg: AKA, Akad. Verl.-Ges., 2009, pp. 101–115.

[DP00] T. Davenport and L. Prusak. *Working Knowledge: How Organizations Manage What They Know*. Harvard Business School Press, 2000.

[DR98] J. Desel and W. Reisig. 'Place/transition Petri Nets'. In: ed. by W. Reisig and G. Rozenberg. Vol. 1491. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1998, pp. 122–173.

[DRSV] S. Dillon, K. Rastrick, F. Stahl, and G. Vossen. 'WiPo for SAR: Taking the Web in your Pocket when doing Search and Rescue'. In preparation.

[DRSV14] S. Dillon, K. Rastrick, F. Stahl, and G. Vossen. 'Cases for the Web in the Pocket (WiPo): Surviving Offline with Online Data'. In: *International Journal of Information Technology and Web Engineering* 9.3 (2014).

[Dru69] P. F. Drucker. *Age Of Discontinuity*. Butterworth-Heinemann, 1969, p. 384.

[DRV06] A. Doan, R. Ramakrishnan, and S. Vaithyanathan. 'Managing information extraction: state of the art and research directions'. In: *SIGMOD Conference*. Ed. by S. Chaudhuri, V. Hristidis, and N. Polyzotis. ACM, 2006.

[DRW95] M. Dyer, W. Riha, and J. Walker. 'A hybrid dynamic programming/branch-and-bound algorithm for the multiple-choice knapsack problem'. In: *Journal of Computational and Applied Mathematics* 58.1 (1995), pp. 43–54.

[DSV12] S. Dillon, F. Stahl, and G. Vossen. 'Towards The Web in Your Pocket: Curated Data as a Service'. In: *Advanced Methods for Computational Intelligence.* Ed. by N. T. N. et al. Vol. 457. Studies in Computational Intelligence. Berlin: Springer-Verlag, 2012, pp. 25–34.

[DSVR13] S. Dillon, F. Stahl, G. Vossen, and K. Rastrick. 'A Contemporary Approach to Coping with Modern Information Overload'. In: *Communications of the ICISA: An International Journal* 14.1 (2013), pp. 1–24.

[DT07] P. Doorn and H. Tjalsma. 'Introduction: archiving research data'. In: *Archival Science* 7 (1 2007), pp. 1–20.

[DW87] K. Dudziński and S. Walukiewicz. 'Exact methods for the knapsack problem and its generalizations'. In: *European Journal of Operational Research* 28.1 (1987), pp. 3–21.

[EB13] C. Ehnis and D. Bunker. 'The Impact of Disaster Typology on Social Media Use by Emergency Services Agencies: The Case of the Boston Marathon Bombing'. In: *Proccedings of the 24th Australasian Conference on Information Systems.* Melbourne, Australia, Dec. 2013, pp. 1–11.

[Edm99] H. Edmunds. *The Focus Group Research Handbook.* NTC Business Books, 1999.

[FBL+15] C. Forster, P. Bruland, J. Lechtenbörger, B. Breil, and G. Vossen. 'Connecting Clinical Care and Research: Single-Source with x4T — Process Design, Architecture, and Use Cases'. In: *Information Technology* 57.1 (2015), pp. 63–72.

[FBR+12] F. Fritz, S. Balhorn, M. Riek, B. Breil, and M. Dugas. 'Qualitative and quantitative evaluation of EHR-integrated mobile patient questionnaires regarding usability and cost-efficiency'. In: *International Journal of Medical Informatics* 81.5 (2012), pp. 303–313.

[FE74] S. G. Faibisoff and D. P. Ely. 'Information and Information Needs'. In: *Information Reports and Bibliographies* 5 (5 1974).

[FEFP95] R. Fikes, R. Engelmore, A. Farquhar, and W. Pratt. *Network-based Information Brokers.* Tech. rep. KSL-95-13. Knowledge Systems Laboratory, Stanford Univeristy, 1995.

[Fer03] R. Ferber. *Information Retrieval: Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web.* Dpunkt.Verlag GmbH, 2003.

[FHS+14] A. Flizikowski, W. Hołubowicz, A. Stachowicz, L. Hokkanen, T. Kurki, N. Päivinen, and T. Delavallade. 'Social Media in Crisis Management–the iSAR+ Project Survey'. In: *Proceedings of the 11th International ISCRAM Conferenc*. University Park, Pennsylvania, USA, 2014, pp. 707–711.

[FKM+05] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. 'Evaluating implicit measures to improve web search'. In: *ACM Transactions on Information Systems* 23.2 (2005), pp. 147–168.

[FLM+06] A. Farahat, T. LoFaro, J. C. Miller, G. Rae, and L. A. Ward. 'Authority Rankings from HITS, PageRank, and SALSA: Existence, Uniqueness, and Effect of Initialization'. In: *SIAM Journal on Scientific Computing* 27.4 (2006), p. 1181.

[For14] C. Forster. 'Webanwendungsentwicklung mit XML-Techniken'. Dissertation. WWU Münster, 2014.

[FOS97] P. C. Fishburn, A. M. Odlyzko, and R. C. Siders. 'Fixed Fee Versus Unit Pricing for Information Goods: Competition, Equilibria, and Price Wars'. In: *Internet Publishing and Beyond: The Economics of Digital Information and Intellectual Property*. MIT Press, 1997, pp. 167–189.

[FOTB14] K. C. Feeney, D. O'Sullivan, W. Tai, and R. Brennan. 'Improving Curated Web-Data Quality with Structured Harvesting and Assessment'. In: *Int. J. Semant. Web Inf. Syst.* 10.2 (2014), pp. 35–62.

[FS01] S. Feldman and C. Sherman. 'The High Cost of Not Finding Information: An IDC White Paper'. In: (2001).

[FV12] C. Forster and G. Vossen. 'Exploiting XML Technologies in Medical Information Systems'. In: *BLED 2012 Proceedings*. Bled, Slovenia, 2012, pp. 70–83.

[GBR09] J. Griesbaum, B. Bekavac, and M. Ritterberg. 'Typologie der Suchdienste im Internet'. In: *Handbuch Internet-Suchmaschinen*. Ed. by D. Lewandowski. Heidelberg: AKA, Akad. Verl.-Ges., 2009.

[GGRN12] A. Gneezy, U. Gneezy, G. Riener, and L. D. Nelson. 'Pay-what-you-want, identity, and self-signaling in markets'. In: *Proceedings of the National Academy of Sciences* 109.19 (2012), pp. 7236–7240.

[GHH13] J. Gottschlich, I. Heimbach, and O. Hinz. 'The Value Of Users' Facebook Profile Data - Generating Product Recommendations For Online Social Shopping Sites'. In: *ECIS 2013*. 2013, p. 117.

[GHW01] A. V. Goldberg, J. D. Hartline, and A. Wright. 'Competitive Auctions and Digital Goods'. In: *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '01. Washington, D.C., USA: Society for Industrial and Applied Mathematics, 2001, pp. 735–744.

[Gil09] J. Giles. '"Knowledge engine" is unveiled / Got a burning question? Ask Alpha'. In: *The New Scientist* 202.2707 (2009), pp. 18–19.

[Gil11] R. Gill. *Theory and Practice of Leadership*. SAGE Publications, 2011.

[Gil93] P. Gilster. *The Internet navigator*. New York: Wiley, 1993.

[GJ79] M. R. Garey and D. S. Johnson. *Computers And Intractability: A Guide to the Theory of NP-Completeness*. New York, U. S.: W. H. Freeman and Company, 1979.

[GL79] H. Genrich and K. Lautenbach. 'The analysis of distributed systems by means of predicate/transition-nets'. In: *Semantics of Concurrent Computation*. Ed. by G. Kahn. Vol. 70. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1979, pp. 123–146.

[GL81] H. Genrich and K. Lautenbach. 'System Modelling with High-Level Petri Nets'. In: *Theoretical Computer Science* 13.1 (1981). Special Issue Semantics of Concurrent Computation, pp. 109–135.

[GL98] G. Gens and E. Levner. 'An approximate binary search algorithm for the multiple-choice knapsack problem'. In: *Information Processing Letters* 67.5 (1998), pp. 261–265.

[Gou12] R. Gould. *Creating the Strategy: Winning and Keeping Customers in B2B Markets*. Kogan Page Series. Kogan Page, 2012.

[GRC05] W. Ge, M. Rothenberger, and E. Chen. 'A Model for an Electronic Information Marketplace'. In: *Australasian Journal of Information Systems* 13.1 (2005).

[Gri03] M. Grieger. 'Electronic marketplaces: A literature review and a call for supply chain management research'. In: *European Journal of Operational Research* 144.2 (2003), pp. 280–294.

[GST+02] J. Gray, A. S. Szalay, A. R. Thakar, C. Stoughton, and J. v. Berg. *Online Scientific Data Curation, Publication, and Archiving*. Ed. by Microsoft Research. 2002.

[Gut84] E. Gutenberg. *Grundlagen Der Betriebswirtschaftslehre: Zweiter Band: Der Absatz*. 17th edition. Berlin, Heidelberg: Springer, 1984.

[Ham12] L. V. Hammon. *Crowdsourcing: eine Analyse der Antriebskräfte innerhalb der Crowd*. Schriftenreihe Innovative Betriebswirtschaftliche Forschung und Praxis. Kovač, 2012.

[Has12] T. Haselmann. *Cloud-Services in kleinen und mittleren Unternehmen: Nutzen, Vorgehen, Kosten*. Wissenschaftliche Schriften der WWU Münster, Reihe IV, Band 6. Monsenstein und Vannerdat, 2012.

[HC02] K. L. Hui and P. Y. Chau. 'Classifying Digital Products'. In: *Communications of the ACM* 45.6 (2002), pp. 73–79.

[Hei11] P. B. Heidorn. ‘The Emerging Role of Libraries in Data Curation and E-science’. In: *Journal of Library Administration* 51.7-8 (2011), pp. 662–672.

[HHS06] A. Herrmann, M. Heitmann, and F. Stahl. ‘Digitale Produkte richtig verkaufen’. In: *Harvard Business Manager* 8 (2006), pp. 8–12.

[HHS11] O. Hinz, I.-H. Hann, and M. Spann. ‘Price discrimination in e-commerce? An examination of dynamic pricing in name-your-own price markets’. In: *MIS quarterly* 35.1 (2011), pp. 81–98.

[HK01] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Data Management Systems Series. Morgan Kaufmann Publishers, 2001.

[HK08] B. Heinrich and M. Klier. In: *Daten- und Informationsqualität: Auf dem Weg zur Information Excellence*. Wiesbaden, Germany: Vieweg+Teubner Verlag, 2008.

[HMW14] J. Hoffart, D. Milchevski, and G. Weikum. ‘STICS: Searching with Strings, Things, and Cats’. In: *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR ’14. Gold Coast, Queensland, Australia: ACM, 2014, pp. 1247–1248.

[HN05] H. Hansen and G. Neumann. *Wirtschaftsinformatik 1: Grundlagen und Anwendungen*. Grundwissen der Ökonomik. Lucius & Lucius, 2005.

[Hoe14] T. Hoeren. In: *European Intellectual Property Review* (12 2014), pp. 751–754.

[Hoe15] T. Hoeren. ‘Datenschutz in der Cloud: Probleme der Werbewirtschaft bei der Auslagerung von Daten auf amerikanische Cloud-Anbieter’. In: *Annual Multimedia 2015*. Regensburg/Berlin: Walhalla u. Praetoria Verlag GmbH & Co. KG, 2015, pp. 24–26.

[How06] J. Howe. ‘The rise of crowdsourcing’. In: *Wired magazine* 14.6 (2006), pp. 174–178.

[HSBW13] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. ‘YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia’. In: *Artificial Intelligence* 194 (2013), pp. 28–61.

[HT03] T. Hey and A. Trefethen. ‘The Data Deluge: An e-Science Perspective’. In: *Grid Computing - Making the Global Infrastructure a Reality*. Ed. by F. Berman, G. C. Fox, and A. J. Hey. Wiley, 2003, pp. 809–824.

[HV10] T. Haselmann and G. Vossen. *Database-as-a-Service für kleine und mittlere Unternehmen*. Tech. rep. Münster: Institut für angewandte Informatik, 2010.

[IAJG06] P. G. Ipeirotis, E. Agichtein, P. Jain, and L. Gravano. 'To search or to crawl?: towards a query optimizer for text-centric tasks'. In: *SIGMOD Conference*. Ed. by S. Chaudhuri, V. Hristidis, and N. Polyzotis. ACM, 2006, pp. 265–276.

[IHTI78] T. Ibaraki, T. Hasegawa, K. Teranaka, and J. Iwase. 'The multiple choice knapsack problem'. In: *J. Oper. Res. Soc. Japan* 21 (1978), pp. 59–94.

[J L10] J. L. Hong. 'Deep web data extraction'. In: *2010 IEEE International Conference on Systems, Man, and Cybernetics*. New York: Ieee Press Books, 2010, pp. 3420–3427.

[Jen89] K. Jensen. 'Coloured Petri Nets: A High Level Language for System Design and Analysis'. In: *Advances in Petri Nets 1990*. Ed. by G. Rozenberg. Vol. 483. Lecture Notes in Computer Science. Springer, 1989, pp. 342–416.

[Jen98] K. Jensen. 'An Introduction to the Practical Use of Coloured Petri Nets'. In: *Lectures on Petri Nets II: Applications, Advances in Petri Nets, the Volumes Are Based on the Advanced Course on Petri Nets*. Heidelberg, Germany: Springer-Verlag, 1998, pp. 237–292.

[JS82] G. Jaeschke and H. J. Schek. 'Remarks on the Algebra of Non First Normal Form Relations'. In: *Proceedings of the 1st ACM SIGACT-SIGMOD Symposium on Principles of Database Systems*. PODS '82. Los Angeles, California: ACM, 1982, pp. 124–138.

[Jür97] Jürgen Bode. 'Der Informationsbegriff in der Betriebswirtschaftslehre'. In: *Zeitschrift für betriebswirtschaftliche Forschung* 05 (1997), pp. 449–468.

[JV97] M. Jarke and Y. Vassiliou. 'Data warehouse Quality Design: A Review of the DWQ project'. In: *Proceedings of the International Conference on Information Quality (IQ)*. 1997.

[KAA+04] T. Kulikova, P. Aldebert, N. Althorpe, et al. 'The EMBL Nucleotide Sequence Database'. In: *Nucleic Acids Research* 32.Database-Issue (2004), pp. 27–30.

[Kar07] J. Karla. 'Implementierung von Regelkreisen in Geschäftsmodellen für Web 2.0-Publikumsdienste'. In: *HMD Praxis der Wirtschaftsinformatik* 44.3 (2007), pp. 17–26.

[Kar08] J. Karla. 'Differenzierung von Erlösmodellen für Weblog-Betreiber — eine anwendungsfallbezogene Analyse'. In: *HMD Praxis der Wirtschaftsinformatik* 45.3 (2008), pp. 80–90.

[KB08] S. Kvale and S. Brinkmann. *InterViews: Learning the Craft of Qualitative Research Interviewing*. Sage Publications, 2008.

[KDF+11] V. Kantere, D. Dash, G. Francois, S. Kyriakopoulou, and A. Ailamaki. 'Optimal Service Pricing for a Cloud Cache'. In: *Knowledge and Data Engineering, IEEE Transactions on* 23.9 (2011), pp. 1345–1358.

[KDGA11] V. Kantere, D. Dash, G. Gratsias, and A. Ailamaki. ‘Predicting Cost Amortization for Query Services’. In: *SIGMOD Conference*. Ed. by T. K. Sellis, R. J. Miller, A. Kementsietsidis, and Y. Velegrakis. ACM, 2011.

[Keu14] B. Keuter. ‘Bidirektionale Abbildung zwischen Geschaeftsprozessmodellen und IT-Kommunikationssystemen:’ PhD thesis. 2014.

[KK10] A. Kittur and R. E. Kraut. ‘Beyond Wikipedia: Coordination and Conflict in Online Production Groups’. In: *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*. CSCW ’10. Savannah, Georgia, USA: ACM, 2010, pp. 215–224.

[KL02] S. J. Kim and S. H. Lee. ‘An Improved Computation of the PageRank Algorithm’. In: *Lecture Notes in Computer Science* 2291 (2002), pp. 73–85.

[Kle99] J. M. Kleinberg. ‘Authoritative sources in a hyperlinked environment’. In: *Journal of the ACM* 46.5 (1999), pp. 604–632.

[KNS09] J.-Y. Kim, M. Natter, and M. Spann. ‘Pay What You Want: A New Participative Pricing Mechanism’. In: *Journal of Marketing* 73.1 (2009), pp. 44–58.

[KPP04] H. Kellerer, U. Pferschy, and D. Pisinger. *Knapsack Problems*. Berlin, Germany: Springer, 2004.

[KR12] J. Kurose and K. Ross. *Computernetzwerke: der Top-Down-Ansatz*. Always learning. Pearson, 2012.

[Kri09] V. Krishna. *Auction Theory*. Elsevier Science, 2009.

[KUB+12a] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu. ‘Query-based data pricing’. In: *PODS*. 2012, pp. 167–178.

[KUB+12b] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu. ‘Query-Market Demonstration: Pricing for Online Data Markets’. In: *PVLDB* 5.12 (2012), pp. 1962–1965.

[KUB+13] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu. ‘Toward practical query pricing with QueryMarket’. In: *SIGMOD Conference*. 2013, pp. 613–624.

[Kun12] M. Kunter. ‘Pay What You Want-Pricing : Erfolgsfaktoren, Gestaltungsvariablen, Anwendungsbeispiele’. In: *Wirtschaftswissenschaftliches Studium : WiSt ; Zeitschrift für Studium und Forschung*. Wirtschaftswissenschaftliches Studium : WiSt ; Zeitschrift für Studium und Forschung. - Beck, ISSN 0340-1650, ZDB-ID 1202856. - Vol. 41.2012, 6, p. 302-307 41.6, (6) (2012), pp. 302–307.

[LAF12] A. Löser, S. Arnold, and T. Fiehn. ‘The GoOLAP Fact Retrieval Framework’. In: *Business Intelligence*. Vol. 96. Lecture Notes in Business Information Processing. 2012, pp. 84–97.

[Lan12] L. Langman. 'Commoditization'. In: *The Wiley-Blackwell Encyclopedia of Globalization*. Ed. by G. Ritzer. Wiley-Blackwell, 2012.

[Law77] E. L. Lawler. 'Fast Approximation Algorithms for Knapsack Problems'. In: *Foundations of Computer Science, 1977., 18th Annual Symposium on*. 1977, pp. 206–213.

[LC11] D. Lacquiere and S. Courtman. 'Use of the iPad in paediatric anaesthesia'. In: *Anaesthesia* 66.7 (2011), pp. 629–630.

[Lee11] D. Lee. 'JXON: an Architecture for Schema and Annotation Driven JSON/ XML Bidirectional Transformations'. In: *Proceedings of Balisage: The Markup Conference*. Vol. 7. Balisage Series on Markup Technologies. Montréal, Canada, 2011.

[Len03] K. Lenz. 'Modellierung und Ausführung von E-Business-Prozessen mit XML-Netzen'. PhD thesis. 2003, p. 252.

[Lev10] M. Levene. *An Introduction to Search Engines and Web Navigation*. 2nd edition. Hoboken: Wiley, 2010.

[Lew09a] D. Lewandowski, ed. *Handbuch Internet-Suchmaschinen: Nutzerorientierung in Wissenschaft und Praxis*. Heidelberg: AKA, Akad. Verl.-Ges., 2009.

[Lew09b] D. Lewandowski. 'Spezialsuchmaschinen'. In: *Handbuch Internet-Suchmaschinen*. Ed. by D. Lewandowski. Heidelberg: AKA, Akad. Verl.-Ges., 2009, pp. 53–69.

[Lin09] F. Linde. 'Ökonomische Besonderheiten von Informationsgütern'. In: *Wissens- und Informationsmanagement*. Ed. by F. Keuper and F. Neumann. Springer, 2009, pp. 291–320.

[Liu07] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Data-centric systems and applications. Springer, 2007.

[LK14] B.-R. Lin and D. Kifer. 'On Arbitrage-free Pricing for General Data Queries'. In: *Proc. VLDB Endow.* 7.9 (2014), pp. 757–768.

[LL99] M. Levene and G. Loizou. *A Guided Tour Of Relational Databases And Beyond*. London, UK: Springer, 1999.

[LLD11] G. Liu, K. Liu, and Y.-y. Dang. 'Research on discovering Deep Web entries based ontopic crawling and ontology'. In: *Conference on Electrical and Control Engineering (ICECE), 2011 International*. 2011, pp. 2488–2490.

[LLMS13] C. Li, D. Y. Li, G. Miklau, and D. Suciu. 'A Theory of Pricing Private Data'. In: *Proceedings of the 16th International Conference on Database Theory*. ICDT '13. Genoa, Italy: ACM, 2013, pp. 33–44.

[LLS10] K. Laudon, J. Laudon, and D. Schoder. *Wirtschaftsinformatik: eine Einführung*. Pearson Studium - IT. Pearson Deutschland, 2010.

[LM00] R. Lempel and S. Moran. 'The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect'. In: *ACM TRANSACTIONS ON INFORMATION SYSTEMS* 19 (2000), pp. 387–401.

[LM06] A. N. Langville and C. D. Meyer. *Google's PageRank and Beyond: The science of search engine rankings.* Princeton [u.a.]: Princeton Univ. Press, 2006.

[LMHS09] C. A. Lee, R. Marciano, C.-y. Hou, and C. Shah. 'From harvesting to cultivating: transformation of a web collecting system into a robust curation environment'. In: *Proceedings of the 9th ACM/IEEE-CS joint Conference on Digital Libraries.* Ed. by F. Heath and M. L. Rice-Lively. JCDL '09. Austin, TX, USA: ACM, 2009, pp. 423–424.

[LMLG04] P. Lord, A. Macdonald, L. Lyon, and Giaretta David. 'From Data Deluge to Data Curation'. In: *Proceedings of the UK e-Science All Hands Meeting.* Nottingham, 2004, pp. 371–375.

[LO01] K. Lenz and A. Oberweis. 'Modeling Interorganizational Workflows with XML Nets'. In: *HICSS.* IEEE Computer Society, 2001.

[LO03] K. Lenz and A. Oberweis. 'Inter-organizational Business Process Management with XML Nets'. In: *Petri Net Technology for Communication-Based Systems.* Ed. by H. Ehrig, W. Reisig, G. Rozenberg, and H. Weber. Vol. 2472. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2003, pp. 243–263.

[LRU14] J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining Of Massive Datasets.* 2nd edition. Cambridge, UK: Cambridge University Press, 2014.

[LS11] F. Linde and W. Stock. *Informationsmarkt: Informationen im I-Commerce anbieten und nachfragen.* Einführung in die Informationswissenschaft. Oldenbourg Wissenschaftsverlag, 2011.

[LSDJ06] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. 'Content-based multimedia information retrieval: State of the art and challenges'. In: *ACM Trans. Multimedia Comput. Commun. Appl* 2 (2006), pp. 1–19.

[LT06] K. Laudon and C. Traver. *E-commerce: business, technology, society.* Pearson/Prentice Hall, 2006.

[Luo12] J. Luomakoski. 'Why did electronic B2B marketplaces fail? Case study of an agricultural commodity exchange'. Dissertation. University of Jyväskylä, 2012.

[Mac00] R. MacLeod. 'Introduction: Alexandria in History and Myth'. In: *The Library of Alexandria: Centre of Learning in the Ancient World.* London, UK: I. B. Tauris, 2000, pp. 1–15.

[Mac14] R. MacManus. *Trackers: How technology is helping us monitor and improve our health.* Kindle Edition. David Bateman Ltd, 2014.

[Mac62] F. Machlup. *The Production and Distribution of Knowledge in the United States.* Princeton paperbacks. Princeton University Press, 1962.

[Mar06] G. Marchionini. 'Exploratory search: from finding to understanding'. In: *Commun. ACM* 49.4 (2006).

[MD12] K. Möller and L. Dodds. 'The Kasabi Information Marketplace'. In: *21st World Wide Web Conference, Lyon, France.* 2012.

[MF10] A. Marques and J. Figueiredo. 'An Approach to Decentralizing Search, Using Stigmergic Hyperlinks'. In: *ENTERprise Information Systems.* Ed. by J. Quintela Varajão, M. Cruz-Cunha, G. Putnik, and A. Trigo. Vol. 109. Communications in Computer and Information Science. Springer Berlin Heidelberg, 2010, pp. 289–298.

[MJC+07] J. Madhavan, S. Jeffery, S. Cohen, X. Dong, D. Ko, C. Yu, and A. Halevy. 'Web-scale data integration: You can only afford to pay as you go'. In: CIDR. 2007.

[MKK+08] J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy. 'Google's Deep Web crawl'. In: *Proc. VLDB Endow* 1 (2008), pp. 1241–1252.

[MLB+11] S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang, and A. Ghalsasi. 'Cloud computing — The business perspective'. In: *Decision Support Systems* 51.1 (2011), pp. 176–189.

[MO06] J. Musser and T. O'Reilly. *Web 2.0: Principles and Best Practices.* Sebastopol and CA: O'Reilly Media, 2006.

[MS95] S. T. March and G. F. Smith. 'Design and natural science research on information technology'. In: *Decision Support Systems* 15.4 (1995), pp. 251–266.

[MSF+05] E. S. de Moura, C. F. dos Santos, D. R. Fernandes, A. S. Silva, P. Calado, and M. A. Nascimento. 'Improving Web Search Efficiency via a Locality Based Static Pruning Method'. In: *Proceedings of the 14th International Conference on World Wide Web.* WWW '05. Chiba, Japan: ACM, 2005, pp. 235–244.

[MSHP09] C. Maaß, A. Skusa, A. Heß, and G. Pietsch. 'Der Markt für Internet-Suchmaschinen'. In: *Handbuch Internet-Suchmaschinen.* Ed. by D. Lewandowski. Heidelberg: AKA, Akad. Verl.-Ges., 2009.

[MSLV12] A. Muschalle, F. Stahl, A. Löser, and G. Vossen. 'Pricing Approaches for Data Markets'. In: *Proceedings of the Workshop Business Intelligence for the Real Time Enterprise.* Istanbul, Turkey, 2012.

[MT12] N. Mankiw and M. Taylor. *Grundzüge der Volkswirtschaftslehre*. Schäffer-Poeschel, 2012.

[MT90] S. Martello and P. Toth. *Knapsack problems: algorithms and computer implementations*. Wiley-Interscience series in discrete mathematics and optimization. J. Wiley & Sons, 1990.

[Mur09] S. Murray. *The Library: An Illustrated History*. ALA editions. Skyhorse Pub., 2009.

[MUV84] D. Maier, J. D. Ullman, and M. Y. Vardi. 'On the foundations of the universal relation model'. In: *ACM Transactions on Database Systems (TODS)* 9.2 (1984), pp. 283–308.

[MV00] U. Masermann and G. Vossen. 'SISQL: schema-independent database querying (on and off the Web)'. In: *Database Engineering and Applications Symposium, 2000 International*. 2000, pp. 55–64.

[MWLZ09] S. E. Madnick, R. Y. Wang, Y. W. Lee, and H. Zhu. 'Overview and Framework for Data and Information Quality Research'. In: *J. Data and Information Quality* 1.1 (2009), 2:1–2:22.

[MYB87] T. W. Malone, J. Yates, and R. I. Benjamin. 'Electronic markets and electronic hierarchies'. In: *Communications of the ACM* 30.6 (1987), pp. 484–497.

[Nau02] F. Naumann. *Quality-Driven Query Answering for Integrated Information Systems*. Vol. 2261. Lecture Notes in Computer Science. Springer, 2002.

[NDH02] R. Nieschlag, E. Dichtl, and H. Hörschgen. *Einführung in die Lehre von der Absatzwirtschaft*. 19th edition. Duncker & Humblot, 2002.

[Neh65] J. Nehnevajsa. 'Information Needs of Society: Future Patterns'. In: *Proceedings of the 1965 Congress International Federation for Documentation (IFD1965)*. Washington DC, USA, 1965.

[Nel65] T. H. Nelson. 'Complex Information Processing: A File Structure for the Complex, the Changing and the Indeterminate'. In: *Proceedings of the 1965 20th National Conference*. ACM '65. Cleveland, Ohio, USA: ACM, 1965, pp. 84–100.

[Nel99b] T. H. Nelson. 'The Unfinished Revolution and Xanadu'. In: *ACM Comput. Surv.* 31.4es (1999).

[NL05] E. Newcomer and G. Lomow. *Understanding SOA with Web Services*. Independent technology guides. Addison-Wesley, 2005.

[Nor11] K. North. *Wissensorientierte Unternehmensführung*. 5th edition. Gabler, 2011.

[NR00] F. Naumann and C. Rolker. *Assessment methods for information quality criteria.* Tech. rep. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Informatik, 2000.

[NRRS05] Y. Narahari, C. Raju, K. Ravikumar, and S. Shah. 'Dynamic pricing models for electronic business'. In: *Sadhana (Academy Proceedings in Engineering Sciences).* Vol. 30. 2 & 3. Indian Academy of Sciences. 2005, pp. 231–256.

[Obe96] A. Oberweis. *Modellierung und Ausführung von Workflows mit Petri-Netzen.* Teubner Reihe Wirtschaftsinformatik. Vieweg+Teubner Verlag, 1996.

[OLC11] B. Otto, Y. W. Lee, and I. Caballero. 'Information and Data Quality in Networked Business: a key concept for enterprises in its early stages of development'. In: *Electronic Markets* 21.2 (2011), pp. 83–97.

[PAM11] C. L. Palmer, S. Allard, and M. Marlino. 'Data curation education in research centers'. In: *Proceedings of the 2011 iConference.* iConference '11. New York, NY, and USA: ACM, 2011, pp. 738–740.

[Pet62] C. A. Petri. 'Kommunikation mit Automaten'. PhD thesis. Universität Hamburg, 1962.

[Pig20] A. C. Pigou. *The Economics of Welfare.* London: MacMillian and Co., 1920.

[Pis95] D. Pisinger. 'A minimal algorithm for the multiple-choice knapsack problem'. In: *European Journal of Operational Research* 83.2 (1995), pp. 394–410.

[PLW02] L. L. Pipino, Y. W. Lee, and R. Y. Wang. 'Data Quality Assessment'. In: *Commun. ACM* 45.4 (2002), pp. 211–218.

[PN07] G. C. Peng and M. B. Nunes. 'Using PEST Analysis as a Tool for Refining and Focusing Contexts for Information Systems Research'. In: *6th European Conference on Research Methodology for Business and Management Studies.* Lisbon, Portugal, 2007, pp. 229–236.

[PN09] T. Püschel and D. Neumann. 'Management of Cloud Infastructures: Policy-Based Revenue Optimization'. In: *ICIS.* Ed. by J. F. Nunamaker and W. L. Currie. Association for Information Systems, 2009, p. 178.

[Por08] M. E. Porter. 'The Five Competitive Forces That Shape Strategy'. In: *Harvard Business Review* 86 (1 2008), pp. 78–93.

[Por77] M. U. Porat. *The information economy: Definition and Measurement.* Vol. 77-12(1). OT special publication. Washington: The Office, 1977.

[Por85] M. Porter. *Competitive Advantage: Creating and Sustaining Superior Performance.* Free Press, 1985.

[Por97] M. E. Porter. 'How Competitive Forces Shape Strategy.' In: *Harvard Business Review* 57 (2 1997), pp. 137–145.

[Pot00] D. T. Potts. 'Before Alexandria: Libraries in the Ancient Near East'. In: R. MacLeod. *The Library of Alexandria: Centre of Learning in the Ancient World*. London, UK: I. B. Tauris, 2000, pp. 19–33.

[PPG+13] D. Potoglou, S. Patil, C. Gijón, J. F. Palacios, and C. Feijóo. 'The Value Of Personal Information Online: Results From Three Stated Preference Discrete Choice Experiments In The UK'. In: *ECIS*. 2013, p. 189.

[PR13] R. S. Pindyck and D. L. Rubinfeld. *Mikroökonomie*. 8. überarbeitete Auflage. München: Pearson Deutschland GmbH, 2013.

[Pre01] M. Prensky. 'Digital Natives, Digital Immigrants Part 1'. In: *On the Horizon* 9.5 (2001), pp. 1–6.

[PT10] A. Pirkola and T. Talvensaari. 'Addressing the Limited Scope Problem of Focused Crawling Using a Result Merging Approach'. In: *Proceedings of the 2010 ACM Symposium on Applied Computing*. SAC '10. Sierre, Switzerland: ACM, 2010, pp. 1735–1740.

[PTRC07] K. Peffers, T. Tuunanen, M. Rothenberger, and S. Chatterjee. 'A Design Science Research Methodology for Information Systems Research'. In: *Journal of Management Information Systems* 24.3 (2007), pp. 45–77.

[Pun05] K. F. Punch. *Introduction to Social Research: Quantitative and Qualitative Approaches*. 2nd edition. SAGE Publications, 2005.

[Qui09] S. Quirmbach. 'Universal Serach: Kontextuelle Einbindung von Ergebnissen unterschiedlicher Quellen und Auswirkungen auf das User Interface'. In: *Handbuch Internet-Suchmaschinen*. Ed. by D. Lewandowski. Heidelberg: AKA, Akad. Verl.-Ges., 2009, pp. 220–248.

[Raj09] A. Rajaraman. 'Kosmix: Exploring the Deep Web using Taxonomies and Categorization'. In: *IEEE Data Engineering Bulletin*. 2009.

[Ram11] M. L. Ramírez. 'Whose Role Is It Anyway? A Library Practitioner's Appraisal of the Digital Data Deluge'. In: *Bulletin of the American Society for Information Science and Technology* 37.5 (2011), pp. 21–23.

[Ras14] C. Rasche. 'PESTEL-Compliance'. In: *Das Wirtschaftsstudium : wisu* (2014).

[RB09] K. Riemer and F. Brüggemann. 'Personalisierung der Internetsuche'. In: *Handbuch Internet-Suchmaschinen*. Ed. by D. Lewandowski. Heidelberg: AKA, Akad. Verl.-Ges., 2009, pp. 148–169.

[RDGG14] J. Radianti, J. Dugdale, J. J. Gonzalez, and O.-C. Granmo. 'Smartphone Sensing Platform for Emergency Management'. In: *Proceedings of the 11th International ISCRAM Conferenc*. University Park, Pennsylvania, USA, 2014, pp. 379–383.

[RE98] G. Rozenberg and J. Engelfriet. 'Elementary net systems'. In: ed. by W. Reisig and G. Rozenberg. Vol. 1491. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1998, pp. 12–121.

[Red96] T. Redman. *Data Quality for the Information Age*. Artech House Telecommunications Library. Artech House, 1996.

[Rei02] W. Reinartz. 'Customizing Prices in Online Markets'. In: *Symphonya. Emerging Issues in Management* 1 Market-Space Management (2002).

[Rei10] W. Reisig. *Petrinetze: Modellierungstechnik, Analysemethoden, Fallstudien*. Leitfäden der Informatik. Vieweg+Teubner Verlag, 2010.

[RG97] S. Rugge and A. Glossbrenner. *The information broker's handbook*. McGraw-Hill, 1997.

[RK96] J. Rehäuser and H. Krcmar. 'Wissensmanagement im Unternehemen'. In: *MANAGEMENTFORSCHUNG 6*. Ed. by G. Schreyögg and P. Conrad. 5 edition. Gruyter, Walter de GmbH, 1996, pp. 1–41.

[RN02] K. Richter and H. Nohr. *Elektronische Marktplätze: Potenziale, Funktionen und Auswahlstrategien*. Aachen, Germany: Shaker, 2002.

[RPS+05] C. Rusbridge, B. P., R. S., B. P., G. D., L. L., and A. M. 'The Digital Curation Centre: A vision for digital curation'. In: *Proceedings of Local to Global Data Interoperability - Challenges and Technologies*. 2005, pp. 31–41.

[RR98a] W. Reisig and G. Rozenberg, eds. *Lectures on Petri Nets I: Basic Models*. Vol. 1491. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1998.

[RR98b] W. Reisig and G. Rozenberg, eds. *Lectures on Petri Nets II: Applications: Advances in Petri Nets*. Vol. 1492. Lecture Notes in Computer Science. Springer, 1998.

[RST10] M. Reimann, O. Schilke, and J. S. Thomas. 'Toward an understanding of industry commoditization: Its nature and role in evolving marketing competition'. In: *International Journal of Research in Marketing* 27.2 (2010), pp. 188–197.

[SBI+13] M. Stonebraker, D. Bruckner, I. F. Ilyas, G. Beskales, M. Cherniack, S. B. Zdonik, A. Pagan, and S. Xu. 'Data Curation at Scale: The Data Tamer System'. In: *CIDR*. 2013.

[SC11] C. J. Stoffle and C. Cuillier. 'Living the Future: Introduction'. In: *Journal of Library Administration* 51.7-8 (2011), pp. 595–598.

[Sch00] B. F. Schmid. 'Elektronische Märkte'. In: *Handbuch Electronic Business*. Ed. by R. Weiber. Gabler Verlag, 2000, pp. 213–239.

[SF08] H. Simon and M. Fassnacht. *Preismanagement: Strategie - Analyse - Entscheidung - Umsetzung*. Gabler Lehrbuch. Gabler Verlag, 2008.

[SGH+14a] F. Stahl, A. Godde, B. Hagedorn, B. Köpcke, M. Rehberger, and G. Vossen. 'Implementing the WiPo Architecture'. In: *Proceedings of the EC Web*. München, 2014.

[SGH+14b] F. Stahl, A. Godde, B. Hagedorn, B. Köpcke, M. Rehberger, and G. Vossen. *Implementing the WiPo Architecture*. Tech. rep. 20. Münster: ERCIS, 2014.

[SGH+15] F. Stahl, A. Godde, B. Hagedorn, B. Köpcke, M. Rehberger, and G. Vossen. 'High Quality Information Delivery: Demonstrating the Web in Your Pocket for Cineast Tourists'. In: *Proceedings of the BTW 2015*. Hamburg, Deutschland, 2015.

[Sha48] C. E. Shannon. 'A Mathematical Theory of Communication'. In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423.

[SHL06] R. Sanderson, J. Harrison, and C. Llewellyn. 'A curated harvesting approach to establishing a multi-protocol online subject portal'. In: *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*. Chapel Hill, NC, USA: ACM, 2006, p. 355.

[SJPB08] G. Skobeltsyn, F. Junqueira, V. Plachouras, and R. Baeza-Yates. 'ResIn: A Combination of Results Caching and Index Pruning for High-performance Web Search Engines'. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08. Singapore, Singapore: ACM, 2008, pp. 131–138.

[SL07] B. Skiera and A. Lambrecht. 'Erlösmodelle im Internet'. In: *Handbuch Produktmanagement*. 3rd edition. Gabler Verlag, 2007, pp. 871–886.

[SLV15] F. Stahl, A. Löser, and G. Vossen. 'Preismodelle für Datenmarktplätze'. In: *Informatik-Spektrum* 38.2 (2015), pp. 133–141.

[Smi08] P. L. Smith II. 'Where IR you?: Using "open access" to extend the reach and richness of faculty research within a university'. In: *OCLC Systems & Services* 24.3 (2008), pp. 174–184.

[SMS11] J. Schumann, U. Meyer, and W. Ströbele. *Grundzüge der mikroökonomischen Theorie*. Springer-Lehrbuch. Springer Berlin Heidelberg, 2011.

[SN10a] P. A. Samuelson and W. D. Nordhaus. *Volkswirtschaftslehre*. 4th edition. Finanzbuch Verlag, 2010.

[SN10b] P. A. Samuelson and W. D. Nordhaus. *Economics*. Nineteenth International Edition. New York, US: McGraw Hill, 2010.

[SPM11] F. Stahl, S. E. Parkin, and A. van Moorsel. *Cooperative Information Security Knowledge: Content Validation and incentives to contribute*. Tech. rep. 1241. Newcastle University School of Computing Science, 2011.

[SSV13] F. Schomm, F. Stahl, and G. Vossen. 'Marketplaces for Data: An Initial Survey'. In: *ACM SIGMOD Record* 42.1 (2013), pp. 15–26.

[SSV14a] F. Stahl, F. Schomm, and G. Vossen. 'Data Marketplaces: An Emerging Species'. In: *Databases and Information Systems VIII*. Ed. by H.-M. Haav, A. Kalja, and T. Robal. IOS Press, 2014, pp. 145–158.

[SSV14b] F. Stahl, F. Schomm, and G. Vossen. 'The Data Marketplace Survey Revisited'. In: *Proceedings of the 11th International Baltic Conference on Databases and Information Systems*. Ed. by H.-M. Haav, A. Kalja, and T. Robal. Tallinn, Estonia: Tallinn University of Technology Press, 2014, pp. 135–146.

[SSW05] B. Skiera, M. Spann, and U. Walz. 'Erlösquellen und Preismodelle für den Business-to-Consumer-Bereich im Internet'. In: *Wirtschaftsinformatik* 47.4 (2005), pp. 285–293.

[ST06] M. Spann and G. J. Tellis. 'Does the Internet promote better consumer decisions? The case of name-your-own-price auctions'. In: *Journal of Marketing* 70.1 (2006), pp. 65–78.

[Sta05] F. Stahl. *Paid Content: Strategien Zur Preisgestaltung Beim Elektronischen Handel Mit Digitalen Inhalten*. Gabler Edition Wissenschaft. Deutscher Universitäts-Verlag, 2005.

[Sta51] H. von Stackelberg. *Grundlagen der theoretischen Volkswirtschaftslehre*. 2nd edition. Tübingen: Mohr Siebeck, 1951.

[Sto03] W. G. Stock. 'Weltregionen des Internet: Digitale Informationen im WWW und via WWW.' In: *Password* 2 (2003), pp. 26–28.

[SV12] F. Stahl and G. Vossen. 'From Unreliable Web Search to Information Provisioning based on Curated Data'. In: *EMISA Forum* 32.2 (2012), pp. 6–20.

[SV13] F. Stahl and G. Vossen. 'High Quality Information Provisioning and Data Pricing'. In: *IEEE 29th International Conference on Data Engineering Workshops (PhD Symposium)*. Brisbane, Austrailia, 2013, pp. 290–293.

[SV99] C. Shapiro and H. Varian. *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business School Press, 1999.

[SVOK11] F. Schönthaler, G. Vossen, A. Oberweis, and T. Karle. *Geschäftsprozesse für Business Communities: Modellierungssprachen, Methoden, Werkzeuge*. Oldenbourg, 2011.

[SZ79] P. Sinha and A. A. Zoltners. 'The Multiple-Choice Knapsack Problem'. In: *Operations Research* 27.3 (1979), p. 503.

[Tan07] W. C. Tan. 'Provenance in Databases: Past, Current, and Future'. In: *IEEE Data Eng. Bull.* 30.4 (2007), pp. 3–12.

[Tan14] R. Tang. 'Quality and Price of Data'. PhD thesis. National University of Singapore, 2014.

[TASB14] R. Tang, A. Amarilli, P. Senellart, and S. Bressan. 'Get a Sample for a Discount'. In: *Database and Expert Systems Applications.* Ed. by H. Decker, L. Lhotská, S. Link, M. Spies, and R. R. Wagner. LNCS 8644. Springer International Publishing, 2014, pp. 20–34.

[Tay62] R. S. Taylor. 'The process of asking questions'. In: *American Documentation* 13.4 (1962), pp. 391–396.

[TKRB11] C. Tempich, F. Kröger, V. Rieger, and P. Bodenbenner. *Cash Cow "Information": Erfolgsfaktoren für informationszentrierte Unternehmen in 2032.* Tech. rep. Detecon Consulting, 2011.

[TMSW14] N. Tandon, G. de Melo, F. Suchanek, and G. Weikum. 'WebChild: Harvesting and Organizing Commonsense Knowledge from the Web'. In: *Proceedings of the 7th ACM International Conference on Web Search and Data Mining.* WSDM '14. New York, New York, USA: ACM, 2014, pp. 523–532.

[TPS05] S. Theysohn, A. Prokopowicz, and B. Skiera. 'Der Paid Content-Markt – Eine Bestandsaufnahme und Analyse von Preisstrategien'. In: *Medienwirtschaft* 2.4 (2005), pp. 170–180.

[TSBV13] R. Tang, D. Shao, S. Bressan, and P. Valduriez. 'What You Pay for Is What You Get'. In: *Database and Expert Systems Applications.* Ed. by H. Decker, L. Lhotská, S. Link, J. Basl, and A. Tjoa. Vol. 8056. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 395–409.

[TWB+13] R. Tang, H. Wu, Z. Bao, S. Bressan, and P. Valduriez. 'The Price Is Right'. In: *Database and Expert Systems Applications.* Ed. by H. Decker, L. Lhotská, S. Link, J. Basl, and A. Tjoa. Vol. 8056. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 380–394.

[TWRB14] J. Talburt, T. L. Williams, T. C. Redman, and D. Becker. 'Information Quality Research Challenge: Predicting and Quantifying the Impact of Social Issues on Information Quality Programs'. In: *J. Data and Information Quality* 5.1-2 (2014), 2:1–2:3.

[UBS12] P. Upadhyaya, M. Balazinska, and D. Suciu. 'How to price shared optimizations in the cloud'. In: *Proc. VLDB Endow.* 5.6 (2012).

[VCK14] A. G. Vural, B. B. Cambazoglu, and P. Karagoz. 'Sentiment-Focused Web Crawling'. In: *ACM Trans. Web* 8.4 (2014), 22:1–22:21.

[VH07] G. Vossen and S. Hagemann. *Unleashing Web 2.0: From concepts to creativity.* Burlington (Mass.): Elsevier and M. Kaufmann publ., 2007.

[Vom14] L. Vomfell. *Surveying Data Marketplaces: Trends and Developments.* Bachelor Thesis. 2014.

[Vos08] G. Vossen. *Datenmodelle, Datenbanksprachen und Datenbankmanagementsysteme.* Oldenbourg Verlag, 2008.

[Vos86] G. Vossen. 'Entwurf und Bearbeitung von Datenbanken im Universalrelationen-Datenmodell'. PhD thesis. RWTH Aachen, 1986.

[VSSV] L. Vomfell, F. Stahl, F. Schomm, and G. Vossen. 'Marketplaces for Digital Data: Quo Vadis?' In preparation.

[VSSV15] L. Vomfell, F. Stahl, F. Schomm, and G. Vossen. 'In the Wake of the Cloud: Defining Data Marketplaces'. In: (2015). In preparation.

[VW11] G. Vossen and K.-U. Witt. *Grundkurs Theoretische Informatik*. 5th edition. Wiesbaden: Vieweg & Sohn Verlagsgeselschaft mbH, 2011.

[WA07] S. Wang and N. P. Archer. 'Electronic marketplace definition and classification: literature review and clarifications'. In: *Enterprise Information Systems* 1.1 (2007), pp. 89–112.

[Wan11] G. Wang. 'Improving Data Transmission in Web Applications via the Translation between XML and JSON'. In: *Communications and Mobile Computing (CMC), 2011 Third International Conference on.* Apr. 2011, pp. 182–185.

[WB10] S. Wu and R. D. Banker. 'Best Pricing Strategy for Information Services'. In: *J. AIS* 11.6 (2010).

[Wei00] R. Weiber. 'Herausforderung Electronic Business'. In: *Handbuch Electronic Business*. Ed. by R. Weiber. Gabler Verlag, 2000, pp. 3–37.

[Wei10] G. Weikum. 'Chapter 3: Search for Knowledge'. In: *Search Computing*. Ed. by S. Ceri and M. Brambilla. Vol. 5950. Lecture Notes in Computer Science. Berlin and Heidelberg: Springer Berlin Heidelberg and Springer-Verlag Berlin Heidelberg, 2010, pp. 24–39.

[Wei99] G. Weikum. 'Towards Guaranteed Quality and Dependability of Information Services'. In: *Datenbanksysteme in Büro, Technik und Wissenschaft*. Springer Verlag, 1999, pp. 379–409.

[WFH11] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd edition. Burlington and MA: Morgan Kaufmann, 2011.

[WHCA08] S. Wu, L. M. Hitt, P. S. Chen, and G. Anandalingam. 'Customized Bundle Pricing for Information Goods: A Nonlinear Mixed-Integer Programming Approach'. In: *Management Science* 54.3 (2008), pp. 608–622.

[Wie09] T. Wiechert. 'Datenschutz bei Suchmaschinen'. In: *Handbuch Internet-Suchmaschinen*. Ed. by D. Lewandowski. Heidelberg: AKA, Akad. Verl.-Ges., 2009, pp. 285–300.

[Wir10] B. Wirtz. *Electronic Business*. 3rd edition. Gabler Verlag, 2010.

[WLPS98] R. Y. Wang, Y. W. Lee, L. L. Pipino, and D. M. Strong. ‘Manage your information as a product’. In: *Sloan Management Review* 39.4 (1998), pp. 95–105.

[WMA13] L. Weng, F. Menczer, and Y.-Y. Ahn. ‘Virality Prediction and Community Structure in Social Networks’. In: *Scientific Reports* 3 (2013).

[WO01] J. Wilson and A. Oyola-Yemaiel. ‘The evolution of emergency management and the advancement towards a profession in the United States and Florida’. In: *Safety Science* 39.1–2 (2001), pp. 117–131.

[WP07] C. Wagner and P. Prasarnphanich. ‘Innovating Collaborative Content Creation: The Role of Altruism and Wiki Technology’. In: *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on.* Jan. 2007, pp. 18–18.

[WS96] R. Y. Wang and D. M. Strong. ‘Beyond accuracy: what data quality means to data consumers’. In: *J. Manage. Inf. Syst.* 12.4 (1996), pp. 5–33.

[XCZ+10] X. Xian, Z. Cui, P. Zhao, Y. Yang, and G. Zhang. ‘Utility Maximization Model for Deep Web Source Selection and Integration’. In: *Journal of Computers* 5.7 (2010).

Web References

[Ber09] T. Berners-Lee. *Linked Data*. 2009. URL: http://www.w3.org/DesignIssues/LinkedData.html (visited on 2015-05-16).

[Ber89] T. Berners-Lee. *Information Management: A proposal*. 1989. URL: http://www.w3.org/History/1989/proposal.html (visited on 2015-05-31).

[Bla90] S. Blakeslee. *Lost on Earth: Wealth of Data Found in Space*. 1990. URL: http://www.nytimes.com/1990/03/20/science/lost-on-earth-wealth-of-data-found-in-space.html (visited on 2015-05-31).

[Bra12] P. Bradley. *Rollyo closes its doors*. 2012. URL: http://philbradley.typepad.com/phil_bradleys_weblog/2012/09/rollyo-closes-its-doors.html (visited on 2015-05-11).

[Car14] Carrie Munk. *The ALS Association Expresses Sincere Gratitude to Over Three Million Donors*. 2014. URL: http://www.alsa.org/news/media/press-releases/ice-bucket-challenge-082914.html (visited on 2015-03-14).

[CIP13] CIPD. *PESTLE analysis*. 2013. URL: http://www.cipd.co.uk/hr-resources/factsheets/pestle-analysis.aspx (visited on 2015-01-24).

[Cro06] D. Crockford. *JSON: The Fat-Free Alternative to XML*. 2006. URL: http://www.json.org/fatfree.html (visited on 2015-05-31).

[DAR14] DARPA. 2014. URL: http://www.darpa.mil/newsevents/releases/2014/02/09.aspx (visited on 2015-03-02).

[Dav07] David Bailey. *An Insider's View Of Google Universal Search.* 2007. URL: http://searchengineland.com/an-insiders-view-of-google-universal-search-12059 (visited on 2015-05-31).

[DDCed] DDC. not dated. URL: http://www.dcc.ac.uk/resources/curation-lifecycle-model (visited on 2015-05-15).

[Dod12] L. Dodds. *About|kasabi.* 2012. URL: http://blog.kasabi.com/about/ (visited on 2015-05-15).

[Dum12] E. Dumbill. *Data markets compared.* 2012. URL: http://strata.oreilly.com/2012/03/data-markets-survey.html (visited on 2015-05-15).

[ecm13] ecma. *Standard ECMA-404 The JSON Data Interchange Format.* 2013. URL: http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf (visited on 2015-05-31).

[Eco10] T. Economist. *The Data Deluge - Businesses, Governments and Society are only Starting to Tap its Vast Potential.* 2010. URL: http://www.economist.com/node/15579717 (visited on 2015-05-16).

[Efr12] A. Efrati. *Google Gives Search a Refresh.* 2012. URL: http://online.wsj.com/news/articles/SB10001424052702304459804577281842851136290 (visited on 2015-05-31).

[EU12] EU. *Consolidated version of the Treaty on the Functioning of the European Union.* 2012. URL: http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:12012E/TXT (visited on 2015-03-29).

[Fox11] V. Fox. *Schema.org: Google, Bing & Yahoo Unite To Make Search Listings Richer Through Structured Data.* 2011. URL: http://searchengineland.com/schema-org-google-bing-yahoo-unite-79554 (visited on 2015-05-31).

[fut14] futurezone. *Personensuchmaschine 123people stellt Dienst ein.* 2014. URL: http://futurezone.at/b2b/personensuchmaschine-123people-stellt-dienst-ein/62.084.444 (visited on 2015-05-31).

[Gis14] H. Gislason. *DataMarket joins Qlik.* 2014. URL: https://blog.datamarket.com/2014/10/31/datamarket-acquired/ (visited on 2015-05-31).

[Gooeda] Google. *Google Basics.* not dated. URL: http://support.google.com/webmasters/bin/answer.py?hl=en&answer=70897 (visited on 2015-05-31).

[Gooedb] Google. *Google Begins Move to Universal Search.* not dated. URL: http://www.google.com/intl/en/press/pressrel/universalsearch_20070516.html (visited on 2015-05-31).

[Haq08] U. Haque. *Data is a Commodity, or How Not to Revolutionize.* 2008. URL: http://www.bubblegeneration.com/2008/01/data-is-commodity-or-how-not-to.cfm (visited on 2015-04-13).

[Hod14] H. Hodson. *Google's fact-checking bots build vast knowledge bank.* 2014. URL: http://www.newscientist.com/article/mg22329832.700-googles-factchecking-bots-build-vast-knowledge-bank.html (visited on 2015-05-31).

[Hufed] S. Huffman. *Search quality highlights: new monthly series on algorithm changes.* not dated. URL: http://insidesearch.blogspot.com/2011/12/search-quality-highlights-new-monthly.html (visited on 2015-05-31).

[Infeda] D. G. für Informationswissenschaft und Informationspraxis e.V. not dated. URL: http://www.dgd.de/BerufInfobroker.aspx (visited on 2015-05-31).

[Infedb] D. G. für Informationswissenschaft und Informationspraxis e.V. not dated. URL: http://www.dgi-info.de/index.php/qualifizierung (visited on 2014-09-02).

[Infedc] D. G. für Informationswissenschaft und Informationspraxis e.V. not dated. URL: http://www.dgi-info.de/index.php/fachgruppen-und-arbeitskreise/fachgruppen (visited on 2014-09-02).

[Infedd] D. G. für Informationswissenschaft und Informationspraxis e.V. not dated. URL: http://www.dgd.de/aginfobroker.aspx (visited on 2015-05-31).

[Infede] D. G. für Informationswissenschaft und Informationspraxis e.V. not dated. URL: http://www.dgi-info.de/index.php/fachgruppen-und-arbeitskreise/regionale-arbeitskreise (visited on 2015-05-31).

[Infedf] D. G. für Informationswissenschaft und Informationspraxis e.V. not dated. URL: http://www.dgd.de/regionalearbeitskreise.aspx (visited on 2015-05-31).

[Jased] M. Jasra. *Interview: Yahoo's Larry Cornett on Universal Search.* not dated. URL: http://www.searchengineguide.com/manoj-jasra/interview-yahoos-larry-cornett-on-univer.php (visited on 2015-05-31).

[jsoed] json.org. *Introducing JSON.* not dated. URL: http://www.json.org/ (visited on 2015-05-31).

[Jul14] Julia Bast. *Ice Bucket Challenge: Ein Kübel Eiswasser für den guten Zweck.* 2014. URL: http://www.faz.net/aktuell/gesellschaft/prominente-machen-bei-ice-bucket-challenge-mit-13107643.html (visited on 2015-05-29).

[Jur13] O. Jurevicius. *PEST & PESTEL Analysis.* 2013. URL: http://www.strategicmanagement%20insight.com/tools/pest-pestel-analysis.html (visited on 2015-01-24).

[KMKed] A. Kushal, S. Moorthy, and V. Kumar. *Pricing for Data Markets.* not dated. URL: http://www.cs.washington.edu/education/courses/cse544/11wi/projects/kumar_kushal_moorthy.pdf (visited on 2015-05-31).

[Koled] J. Kolb. *Data as a Renewable Commodity and Profit Center.* not dated. URL: http://jasonkolb.com/data-as-a-renewable-commodity-and-profit-center/ (visited on 2015-04-13).

[Koo13] A. Kooser. 2013. URL: http://www.cnet.com/news/intel-reveals-what-happens-in-a-single-internet-minute/ (visited on 2015-05-31).

[Kos10] R. Kosara. *The Rise and Fall of Swivel.com*. Last accessed: 2015-05-4. 2010. URL: https://eagereyes.org/criticism/the-rise-and-fall-of-swivel.

[Kosed] M. Koster. *Historical Web Services*. not dated. URL: http://www.greenhills.co.uk/historical.html (visited on 2015-06-01).

[Laned] LandSAR. *About LandSAR*. not dated. URL: https://www.landsar.org.nz/about/ (visited on 2015-06-01).

[Lib14] T. W. V. Library. *The WWW Virtual Library: History of the Virtual Library*. 2014. URL: http://vlib.org/admin/history (visited on 2015-05-31).

[Luc13] Lucianne Poole. *"Wiring" the Ottawa Hospital for success*. 2013. URL: http://researchworks.carleton.ca/2013/02/wiring-ottawa-hospital-success/ (visited on 2015-06-01).

[Mared] T. O. Marketing. *SES San Jose: Universal and Blended Search*. not dated. URL: http://www.toprankblog.com/2008/08/ses-san-jose-universal-and-blended-search/ (visited on 2012-06-12).

[MIAed] T. MIA. *MIA – ein Marktplatz für Informationen und Analysen*. not dated. URL: http://mia-marktplatz.de/ (visited on 2015-04-27).

[Mic11] Microsoft. *Windows Azure Marketplace*. 2011. URL: http://go.microsoft.com/fwlink/?LinkID=201129&clcid=0x409 (visited on 2015-05-31).

[Mil12a] P. Miller. *Data Market Chat*. 2012. URL: http://cloudofdata.com/category/podcast/data-market-chat/ (visited on 2015-05-31).

[Mil12b] P. Miller. *Data markets: in search of new business models*. 2012. URL: http://research.gigaom.com/report/data-markets-in-search-of-new-business-models/ (visited on 2015-05-31).

[Mil12c] P. Miller. *Nurturing the market for Data Markets*. 2012. URL: http://cloudofdata.com/2012/01/nurturing-the-market-for-data-markets/ (visited on 2015-05-31).

[Miteda] J. Mitchell. *A Year of Tweaks to Google Search: Are You "Fed Up?"* not dated. URL: http://www.readwriteweb.com/archives/a_year_of_tweaks_to_google_search_are_you_fed_up.php (visited on 2015-05-31).

[Mitedb] J. Mitchell. *Google+ Is Going To Mess Up The Internet*. not dated. URL: http://www.readwriteweb.com/archives/google_is_going_to_mess_up_the_internet.php (visited on 2015-05-31).

[Mitedc] J. Mitchell. *How Google Search Really Works: (Interview with Ben Gomes)*. not dated. URL: http://www.readwriteweb.com/archives/interview_changing_engines_mid-flight_qa_with_goog.php (visited on 2015-05-31).

[Nel08] T. H. Nelson. *Ted Nelson demonstrates Xanadu Space*. 2008. URL: https://www.youtube.com/watch?v=En_2T7KH6RA (visited on 2015-05-31).

[Nel99a] T. H. Nelson. *Ted Nelson's Computer Paradigm, Expressed as One-Liners.* 1999. URL: http://hyperland.com/TedCompOneLiners (visited on 2015-05-31).

[NF13] F. Nogatz and T. Frühwirth. *From XML Schema to JSON Schema: Translation with CHR.* 2013. URL: http://www.informatik.uni-ulm.de/pm/fileadmin/pm/home/fruehwirth/Papers/Nogatz-Fruehwirth_XSD-to-JSON-Schema-with-CHR.pdf (visited on 2015-05-31).

[Nig14] S. Niggemeier. *Leistungsschutzrecht wirkt: Mehrere Suchmaschinen zeigen Verlagsseiten nicht mehr an.* 2014. URL: http://www.stefan-niggemeier.de/blog/19058/leistungsschutzrecht-wirkt-mehrere-suchmaschinen-zeigen-verlagsseiten-nicht-mehr-an/ (visited on 2015-06-01).

[Nor14] R. Nord. *Datenmarkt.* 2014. URL: http://rtlnord.de/nachrichten/datenmarkt.html (visited on 2015-03-17).

[Off14] I. P. Office. *Intellectual Property Office: Legislation.* 2014. URL: http://www.iponz.govt.nz/cms/copyright/legal (visited on 2015-03-30).

[Oxfeda] Oxford Dictionaries. *Oxford Dictionaries: Curation.* not dated. URL: http://www.oxforddictionaries.com/definition/english/curate#curate-2 (visited on 2015-05-23).

[Oxfedb] Oxford Dictionaries. *Oxford Dictionaries: Data.* not dated. URL: http://www.oxforddictionaries.com/definition/english/data (visited on 2015-05-23).

[Pfl14] R. Pflanzenforschung.de. *Die Versalzung der Böden entzieht der Landwirtschaft fruchtbare Böden.* 2014. URL: http://www.pflanzenforschung.de/de/journal/journalbeitrage/2000-hektar-am-tag-die-versalzung-der-boeden-entzieht-d-10337 (visited on 2015-04-30).

[Rap07] RapidBI. *PESTLE analysis tool.* 2007. URL: https://rapidbi.com/the-PESTLE-analysis-tool/ (visited on 2015-01-24).

[Ros14] P. Rostkowski. *Third-Party Data Suppliers Need to Give Us What We Pay For.* 2014. URL: http://adage.com/article/digitalnext/party-data-suppliers-give-pay/245849/ (visited on 2015-04-23).

[Rosed] J. Rossiter. *Progress Report: Continued Product Focus.* not dated. URL: http://yahoo.tumblr.com/post/98474044364/progress-report-continued-product-focus (visited on 2015-05-31).

[Sat14] G. Satell. *A Look Back At Why Blockbuster Really Failed And Why It Didn't Have To.* 2014. URL: http://www.forbes.com/sites/gregsatell/2014/09/05/a-look-back-at-why-blockbuster-really-failed-and-why-it-didnt-have-to/ (visited on 2015-05-06).

[Sch14] S. Schechner. *Google Starts Removing Search Results Under Europe's 'Right to be Forgotten'.* 2014. URL: http://www.wsj.com/articles/google-starts-removing-search-results-under-europes-right-to-be-forgotten-1403774023 (visited on 2015-05-31).

[Sjo14] P. Sjofors. *The Top 10 Pricing Mistakes Companies Are Making*. 2014. URL: http://techcrunch.com/2014/12/27/the-top-ten-pricing-mistakes/?ncid=rss&utm_source=feedburner&utm_medium=feed&utm_campaign=Feed:+Techcrunch+(TechCrunch)&utm_content=FaceBook (visited on 2015-04-27).

[Sma09] A. Smarty. *Let's Try to Find All 200 Parameters in Google Algorithm*. 2009. URL: http://www.searchenginejournal.com/200-parameters-in-google-algorithm/15457/ (visited on 2015-05-31).

[Sul04] D. Sullivan. *Eurekster Launches Personalized Social Search*. 2004. URL: http://searchenginewatch.com/sew/news/2048359/eurekster-launches-personalized-social-search (visited on 2015-03-11).

[Sul10] D. Sullivan. *Schmidt: Listing Google's 200 Ranking Factors Would Reveal Business Secrets*. 2010. URL: http://searchengineland.com/schmidt-listing-googles-200-ranking-factors-would-reveal-business-secrets-51065 (visited on 2015-05-31).

[TA14] A. Travis and C. Arthur. *EU court backs 'right to be forgotten': Google must amend results on request*. 2014. URL: http://www.theguardian.com/technology/2014/may/13/right-to-be-forgotten-eu-court-google-search-results (visited on 2015-03-12).

[tag13] tagesschau.de. *Breaking news: Alles bleibt beim Alten*. 2013. URL: http://www.tagesschau.de/inland/leistungsschutzrecht126.html (visited on 2015-06-01).

[tag14a] tagesschau.de. *0:1 im Streit Verlage gegen Google*. 2014. URL: http://www.tagesschau.de/wirtschaft/leistungsschutzrecht-101.html (visited on 2015-06-01).

[tag14b] tagesschau.de. *Springer und Burda nur noch einzeilig*. 2014. URL: http://www.tagesschau.de/wirtschaft/google-verlage-101.html (visited on 2015-06-01).

[tag14c] tagesschau.de. *Verlage geben Google nach - vorerst*. 2014. URL: http://www.tagesschau.de/wirtschaft/vgmedia-google-101.html (visited on 2015-06-01).

[Tho13] J. Thomson. *Why CFOs Are Drowning In Data But Starving For Information*. 2013. URL: http://www.forbes.com/sites/jeffthomson/2013/10/30/why-cfos-are-drowning-in-data-but-starving-for-information/ (visited on 2015-05-31).

[Urhed] UrhG. *Gesetz über Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz)*. not dated. URL: http://www.gesetze-im-internet.de/urhg/BJNR012730965.html#BJNR012730965BJNG004400140 (visited on 2015-06-01).

[W3C00] W3C. *A Little History of the World Wide Web*. 2000. URL: http://www.w3.org/History.html (visited on 2015-05-31).

[W3Ced] W3C. *Resource Description Framework (RDF)*. not dated. URL: http://www.w3.org/RDF/ (visited on 2015-05-31).

[Wiked] Wikipedia. *Wikipedia:Copyrights)*. not dated. URL: https://en.wikipedia.org/wiki/Wikipedia:Copyrights (visited on 2015-06-01).

[WIPed] WIPO. *Inside WIPO: What is WIPO*. not dated. URL: http://www.wipo.int/about-wipo/en/ (visited on 2015-06-01).

[WM15] R. Winkler and B. Mullins. *How Google Skewed Search Results*. 2015. URL: http://www.wsj.com/articles/how-google-skewed-search-results-1426793553 (visited on 2015-05-31).

[Woled] WolframAlpha. *About Wolfram|Alpha*. not dated. URL: http://www.wolframalpha.com/about.html (visited on 2015-05-31).

[Wor15] WorldWideWebSize. *The size of the World Wide Web (The Internet)*. 2015. URL: http://www.worldwidewebsize.com/ (visited on 2015-05-31).

[Wored] World Health Organization. *2014 West African Ebola outbreak: feature map*. not dated. URL: http://www.who.int/features/ebola/storymap/en/ (visited on 2015-03-14).

[Wried] A. Wright. *Exploring a 'Deep Web' That Google Can't Grasp*. not dated. URL: http://www.nytimes.com/2009/02/23/technology/internet/23search.html?_r=1&ref=business (visited on 2015-05-31).

[yahed] yahoo! *yahoo! Meilensteine*. retrieved form the Internet archive as it is no longer available online through the comapany Website. not dated. URL: http://web.archive.org/web/20070918225007/http://yhoo.client.shareholder.com/press/history.cfm (visited on 2015-05-16).

List of Abbreviations and Acronyms

ADI	Arbeitskreis Dresdner Informationsvermittler
AIT	Arbeitskreis für Informationsvermittler Thüringen
AKI	Arbeitskreis für Information
AKRIBIE	Arbeitskreis für Information Bielefeld, Ostwestfalen-Lippe
API	Application Programming Interface
ARPA	Advanced Research Projects Agency
BAK	Berliner Arbeitskreis Information
BI	Business-Intelligence
BRAGI	Brandenburgische Arbeitsgemeinschaft Information
CEO	Chief Executive Officer
CWA	Closed World Assumption
DRM	Digital Rights Management
ENV	Environment
ERP	Enterprise Resource Planning
EU	European Union
GDP	Gross Domestic Product
GUI	Graphical User Interface
HITS	Hypertext Induced Topic Search
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
IR	Information Retrieval
IT	Information Technology
MAID	Münchener Arbeitskreis für Information und Dokumentation
MCKP	Multiple-Choice Knapsack Problem
MCKPP	Multiple-Choice Knapsack Pricing Problem
NF^2	Non-First-Normal-Form
NGO	Non-Governmental Organisation
NoSQL	Not only SQL
NYOP	Name Your Own Price
OWA	Open World Assumption
PEST	Political, Economic, Social, Technological
PESTEL	Political, Economic, Social, Technological, Environmental, Legal
PMP	Paris Metro Pricing
PWYW	Pay What You Want
RDF	Resource Description Framework
REST	Representational State Transfer
SaaS	Software-as-a-Service
SALSA	Stochastic Approach for Link-Structure Analysis
SCC	Strongly Connected Component
SME	Small and Medium-sized Enterprise
SOA	service-oriented architecture
SPEET	Social, Political/Legal/Ethic, Economic/Business, Environmental, Technological
SQL	Structured Query Language
STEEP	Social, Technological, Economic, Ethical, Political
SWOT	Strengths, Weaknesses, Opportunities, Threats
UDF	User Defined Function
URL	Uniform Resource Locator
VERONICA	Very Easy Rodent-Oriented Net-wide Index to Computer Archives
W3C	World Wide Web Consortium
WAIS	Wide Area Information Servers
WIPO	World Intellectual Property Organization
WiPo	Web in your Pocket
WWW	World Wide Web
XML	Extensible Markup Language

List of Most Important Symbols

A	Set of Automatic Evaluation Quality Criteria
A	Attribute of a Relation
a	Knapsack Solution Matrix
b	Benefit
C	Set of Comparison-Only Quality Criteria
$D = (\boldsymbol{R},\Sigma_{\boldsymbol{R}})$	Database Schema over Relations $\boldsymbol{R}$ with Constraints $\Sigma_{\boldsymbol{R}}$
d	Database Instance
$\mathrm{dom}(A)$	Domain of Attribute(s) A
G	Set of Goldilocks Versioning Quality Criteria
H	Set of Hybrid Evaluation Quality Criteria
κ	Cost Attribution Vector
M	Set of Manual Evaluation Quality Criteria
$M = (\boldsymbol{u}, \boldsymbol{U}, \Sigma_M)$	Marketplace over Universal Relations $\boldsymbol{u}$ with Schemas $\boldsymbol{U}$ and Constraints Σ_M
m_l	Number of Quality Levels
μ	Tuple
$\mu[X]$	Tuple over Attributes X
n_q	Number of Quality Criteria
n_t	Number of Overall Knapsack Elements
$\perp$	Null Value
ω	Preference Vector
P	Ask Price
$\pi_Y(r)$	Projection of Relation r to the Attribute(s) Y
q	Quality Criterion
Q_a	Set of all Quality Criteria
Q_v	Set of all Versioning Quality Criteria
$Rel(X)$	Set of Relations over Attributes X
r	Relation
$R = (X, \Sigma_X)$	Relational Schema over Set of Attributes X with Constraints Σ_X
$\boldsymbol{R}$	Set of Relational Schemas

$\sigma_{Condition}(r)$	Selection of Tuples in Relation r that Match *Condition*
Σ_R	Set of Inter-Relational Constraints
Σ_X	Set of Intra-Relational Constraints
$Tup(X)$	Set of Tuples Restricted to Attributes X
$U = (X_U, \cdot)$	Universal Relational Schema over Attributes X_U with Unspecific Constraints
u	Universal Relation
V	Set of Automated Versioning Quality Criteria
W	Bid Price
w	Weight
X	Set of Excluded Quality Criteria
$\bowtie$	Inner Join
⟗	Full Outer Join

High-Quality Web Information Provisioning and Quality-Based Data Pricing

Florian Stahl

Today, information can be considered a production factor. This is attributed to the technological innovations the Internet and the Web have brought about. Now, a plethora of information is available making it hard to find the most relevant information. Subsequently, the issue of finding and purchasing high-quality data arises. Addressing these challenges, this work first examines how high-quality information provisioning can be achieved with an approach called WiPo that exploits the idea of curation, i. e., the selection, organisation, and provisioning of information with human involvement. The second part of this work investigates the issue that there is little understanding of what the value of data is and how it can be priced – despite the fact that it is already being traded on data marketplaces. To overcome this, a pricing approach based on the Multiple-Choice Knapsack Problem is proposed that allows for utility maximisation for customers and profit maximisation for vendors.

ISBN 978-3-8405-0133-3 EUR 20,10
02010
9 783840 501333