EXPRESSIVE RESOURCE DESCRIPTIONS FOR ONTOLOGY-BASED INFORMATION RETRIEVAL

Thanh Tran, Stephan Bloehdorn, Philipp Cimiano and Peter Haase
Institute AIFB, University of Karlsruhe, D-76128 Karlsruhe, Germany
dtr@aifb.uni-karlsruhe.de, sbl@aifb.uni-karlsruhe.de, pci@aifb.uni-karlsruhe.de, pha@aifb.uni-karlsruhe.de

Keywords: Information Retrieval, Resource Data Model, Ontology

Abstract: In this paper, we introduce an expressive ontology-based model for representing resources with respect to a domain ontology. Our resource model is based on semantic web standards as well as established ontologies and metadata schemas such as SUMO, MPEG-7 and Dublin Core to provide a reference model for ontology-based information retrieval. Based on this expressive resource model, the user can directly specify his information need at an enhanced level of expressiveness. In particular, it does not restrict the description of resources to keywords but allows for the description of resources in terms of factual and terminological axioms as well as events and complex situations. We show that with the proposed resource description model, a large set of different retrieval functionalities can be supported to address complex information needs.

1 INTRODUCTION

The aim of an Information Retrieval (IR) system is to retrieve resources (often synonymously referred to as documents) relevant to a given user query while retrieving as few non-relevant documents as possible (Baeza-Yates and Ribeiro-Neto, 1999). Since the early days of IR research, different IR models describing documents and queries as well as capturing the relation between information and resources have been defined. Regardless of the IR model adopted, one needs to 'interpret' the documents' content and queries w.r.t to the model chosen, i.e. as Baeza-Yates and Ribeiro-Neto (1999) have pointed out: *"To be effective in its attempt to satisfy the user information need, the IR system must somehow 'interpret´ the contents of the information items (documents) in a collection and rank them according to a degree of relevance to the user query."* The crucial question is certainly, how expressive the IR model is and thus how much interpretation is indeed required. So far, IR systems have rarely aimed for a real interpretation of resource content but often adopted the so called fulltext document view (or variants thereof). In this view, a (textual) resource and the information need are simply represented by the set of terms. Since the beginning of IR research, users and developers have envisioned alternative paradigms that allow the user to provide expressive descriptions of his information need and are able to match them against the system resources. We argue that in order to address more complex information needs, it is necessary to move towards a more fine-grained representation of the semantic content of documents. Consider for example the following information need:

Example. *A user is searching the publications of the research institute AIFB using the information portal* `http://www.aifb.uni-karlsruhe.de`. *He might look for a publication that (i) was written by an author of the knowledge management research group, (ii) deals with the topic of information retrieval and (iii) describes a question answering system that has been deployed in a corporate setting.*

In fact, answering such an information need requires capturing the resources at a much more fine-grained level than done in current IR mod-

els. The move towards an elaborated resource description model obviously blurs the predominant distinction between an information retrieval and a data retrieval system as postulated by van Rijsbergen (1979). Having an expressive description of the resource content implies that the system storing these descriptions needs to be queried, which can be seen as a data retrieval task. In essence, moving towards more expressive IR models means also rephrasing the IR as a data retrieval task in which the documents matching a certain complex description are retrieved from a certain database.

Towards this end, one needs to define what and how the content of resources should be described. In this paper, we propose an ontology-based resource model that captures different aspects of the resources in a way that can address such complex information needs. In particular, all these different aspects of the resource (and the example information need) are addressed: (i) retrieval based on standard resource metadata, (ii) retrieval based on the content's topic classification, (iii) and retrieval based on complex descriptions of the actual resource content. While (i) and (ii) are more or less widely supported in existing IR systems (especially those for accessing digital libraries), aspect (iii) is currently only available as fulltext or index-term based retrieval, while the specification of complex descriptions of the content is still an open research question. Our model does not constrain the resource description (and the description of the information need) to keywords or named entities but allows to specify arbitrary factual and terminological axioms to describe relationships, events and complex phenomena with respect to a domain ontology. Our proposal is generic and can be flexibly extended and tuned.

We formalize these different aspects in a specific, logic-based, instantiation of the classical IR model formulated by Baeza-Yates and Ribeiro-Neto (1999), which we refer to as Ontology-based Information Retrieval (OIR). In this formalization, OIR involves the user query model, an ontology-based system resource model and the system query model as core elements, and query translation and query answering as tasks that need to be performed by OIR systems. Based on the previous example information need, we give concrete instances of the resource model. We discuss how these resource descriptions can be developed and demonstrate how the formalized OIR models and tasks can be managed to address complex information needs.

The paper is organized as follows. We start with related work in Section 2. We review the classical information retrieval model in Section 3, and present an instantiation of this model in the light of ontology-based information retrieval in Section 4, where we discuss the involved elements and tasks. As the main contribution, Section 5 presents a concrete resource model that is developed based on the basis of OWL (Web Ontology Language) and several other existing ontology and metadata standards. In Section 6, we discuss the application context of our framework, both with respect to mechanisms to obtain resource descriptions and to interpret user queries. Also, we illustrate how these example resource descriptions allow for the matching of queries to complex information needs. Finally, we conclude in Section 7, among others with pointers to open issues and future work.

2 RELATED WORK

The OIR model we present in this paper has its roots in the formalization of relevance of a document w.r.t. a user query as a logical implication (van Rijsbergen, 1986). In particular, our formalization is close to the work of Meghini et al. (1993), who use a terminological logic to model the retrieval problem. While these are two examples, there are many other approaches which advocate the use of logical formalisms to represent documents and to consider query answering as a problem of determining logical implication. However, the formalization proposed here is more centered on ontologies, which are expressed by means of logical formalisms. Thus, this also implies an inherently logic-based view on . However, our formalization suggests that besides query answering, the knowledge formalized in the ontologies (and referred to in the document descriptions) can also be exploited for enhancing and translating the user queries to system queries.

Besides, the main contribution of this paper actually lies in the proposal of an ontology-based resource model, which is a particular element of our OIR formalization. Thus, more specifically related are logic-based approaches to that make more detailed assumptions on the model of the underlying resources. For instance, Fuhr (1995) has dealt with how to model resources in Datalog. In his proposal, resources are represented through the concept `document`, which has (meta-

data) properties such `title` and `author`, and in particular `docTerm`, which points to terms of a thesaurus. In the model of Fuhr, the modeling of document structure and content as well as of terminological knowledge in the form of Datalog clauses allows for drawing non-trivial inferences at retrieval time. However, as Fuhr himself acknowledges, a more expressive model cannot be achieved with Datalog, but with a "terminological logic". A fuzzy version of the well-known terminological logic $\mathcal{ALC}$ is for example used by Meghini et al. (2001) for modeling the retrieval of multimedia resources. Among other properties, such a model can refer to instances of fuzzy $\mathcal{ALC}$ concepts. In more recent approaches, ontologies are explicitly used. For instance, resources are described through ontology elements such as `annotation`s, which carry two properties `instance` and `document` by which document entities and other ontology entities are related (Vallet et al., 2005). A resource description comprises a document and instances of the class `domain concept` representing entities referred to in the document. Popov et al. (2003) undertake a different approach, i.e. a document description can also be described by a set of instances of `lexical resource`s, i.e. terms. The semantics of these instances is established by the property `hasAlias`, which relates these lexical resources to instances of entities of a domain ontology.

The resource model proposed in this paper is different in the sense that it specifically distinguishes the different aspects of resources, namely content, structure and presentation. Most importantly, it is distinct in the level of expressivity. In all the above mentioned approaches, a resource is described simply by a set of elements, e.g. terms of a taxonomy (Fuhr) and concept instances (Meghini et al., Vallet et al., Popov et al.). Besides instances, resource descriptions in our approach can also refer to concepts and any complex expressions expressible in the Ontology Web Language (OWL). That is, our model allows assertional axioms (descriptions at the instance level) as well as terminological axioms (descriptions at the concept level) to be the subject of content. For instance, the model allows to specify that a content describes Philipp Cimiano, who is a researcher working at the Knowledge Management Department, which is part of the institute AIFB, i.e. to refer to axioms such as `works_at(Philipp Cimiano, KM Group)` and these entities are further described in the domain ontology through the axioms `Researcher(Philipp Cimiano)`, and `partOf(KM Group,AIFB)`.

In addition, we also spell out in details how existing metadata schemas and ontology standards such as the Dublin Core (DC) (Weibel, 2000), MPEG-7 (Hunter, 2001) as well as the IEEE standard for foundational ontology SUMO (Niles and Pease, 2001) can be smoothly integrated to make them compatible with the proposed resource model. In our view, this alignment with existing standards is a key aspect for this resource model to be widely accepted.

3 CLASSICAL INFORMATION RETRIEVAL MODELS

In this section we begin with a formalization of different Information Retrieval (IR) models. We start with a recapitulation of the classical information retrieval model and its vector-based variant as one well-established instantiation. Subsequently, in Section 4 we describe our definition of an Ontology-based Information Retrieval (OIR) as a novel instantiation. In one of the classic references, Baeza-Yates and Ribeiro-Neto (1999) formalize an IR system as follows:

Definition 1 (Information Retrieval Model)**.** *An information retrieval model is a quadruple* $\langle \mathcal{D}, \mathcal{Q}, \mathcal{F}, R(q_i, d_j) \rangle$ *where*

1. $\mathcal{D}$ *is a set composed of views (representations) for the resources (documents) in the collection.*
2. $\mathcal{Q}$ *is a set composed of views (representations) for the user information needs. These are called queries.*
3. $\mathcal{F}$ *is a framework for modeling resource representations, queries and their relationships.*
4. $R(q_i, d_j)$ *is a ranking function which associates a real number with a query* $q_i \in \mathcal{Q}$ *and a document representation* $d_j \in \mathcal{D}$*. Such ranking defines an ordering among the documents with regard to the query* q_i*.*

As an example, a common variant of vector-based fulltext retrieval can be formalized as a particular instantiation of this generic IR model as follows:

Definition 2 (Vector-based Information Retrieval Model)**.** *The Vector-based Retrieval Model is a quadruple* $\langle \mathcal{D}_\mathcal{V}, \mathcal{Q}_\mathcal{V}, \mathcal{F}_V, R_V(q_i, d_j) \rangle$ *where*

1. $\mathcal{D}_\mathcal{V}$ *is the so called bag-of-words-model in which a document is represented through a set*

of words contained in the document with associated word weights

2. $\mathcal{Q}_\mathcal{V}$ *consists of sets of keywords*
3. $\mathcal{F}_\mathcal{V}$ *is a vector-based framework in which documents and queries are represented as vectors* $\vec{q}$ *and* $\vec{d}$ *in a t-dimensional space whereby the dimensions correspond to words appearing in the full text representation of documents. The available operations are then operations in the t-dimensional vector space (e.g. the dot product or the cosine of the angle between two vectors etc.)*
4. $R_V(q_i, d_j)$ *is defined as the cosine between the document and the query vectors respectively, i.e.* $R(q_i, d_j) := cos(\vec{d_j}, \vec{q_i})$.

While illustrating the main concepts of the IR model, this example also serves te purpose to show the differences to our alternative.

4 Ontology-based Information Retrieval

Before digging into the details, we start by explaining our notion of an ontology. Here, we follow a description logics-based view on ontologies (Baader et al., 2003). In description logics, the important notions of a domain are described by means of concept descriptions that are built from different *ontology entities* called *concepts* (also referred to as *classes*), *roles* (also referred to as *properties* or *relations*), denoting relationships between things, and *individuals* (also referred to as *instances*). Entities can be related to each other and constrained by means of axioms. Terminological axioms make statements about how concepts or roles are related to each other, assertional axioms (sometimes also called *facts*) make statements about the properties of individuals of the domain. The types of available axioms and their structure vary depending on the specific description logic under consideration. By *ontology elements* we refer to the ontology entities together with the axioms.

We now define OIR as another instantiation of the general definition of the IR model.

Definition 3 (Ontology-based Information Retrieval Model)**.** *The Ontology-based Information Retrieval Model is a quadruple* $\langle \mathcal{D}_\mathcal{O}, \mathcal{Q}_\mathcal{O}, \mathcal{F}_\mathcal{O}, R_O(q_i, d_j) \rangle$ *where*

1. $\mathcal{D}_\mathcal{O}$ *is the ontology-based model in which a resource is represented through a set of ontology elements* $o \in \mathcal{O}$*. For this, we assume a function* $f_\mathcal{T} : \mathcal{D} \rightarrow \mathcal{D}_\mathcal{O}$*, which transforms a resource* $d \in \mathcal{D}$ *into an ontology-based representation* $\mathcal{D}_\mathcal{O}$.
2. $\mathcal{Q}_\mathcal{O}$ *is a set of elements that represent the user information needs. We assume a correspondence between the elements of* $\mathcal{Q}_\mathcal{O}$ *with the ontology elements in* $\mathcal{O}$*. This correspondence allows to represent* $\mathcal{Q}_\mathcal{O}$ *in an ontology-based representation* $\mathcal{Q}'_\mathcal{O}$.
3. $\mathcal{F}_\mathcal{O}$ *is an ontology-based framework in which resources and queries are represented as ontology elements. An entailment operation checks whether the ontology-based representation of the resource entails the ontology-based representation of the information need, i.e. if for a given information need* q_i *and resource* d_j *the entailment relation* $d_j \models_O q_i$ *holds (query answering)*[1].
4. $R_O(q_i, d_j)$ *is a ranking function defined with* $R(q_i, d_j) \in (0, 1]$ *iff* $d_j \models_O q_i$ *and* $R(q_i, d_j) = 0$ *otherwise.*

We will now continue with a more detailed elaboration on our framework for OIR, discussing the components of the OIR model one after another. In particular, we also discuss how the rather abstract elements of the model can be instantiated with concrete formalisms.

4.1 System Resource Model

In our model, system resources are described according to an ontology that we refer to as the System Resource Model (SRM), which we present in its details in Section 5. Resource descriptions in $\mathcal{D}_O$ comprise a resource entity (representing system resources), domain entities, and axioms making statements about them, i.e. defining relations among them.

In our realization of the SRM, we rely on OWL (Web Ontology Language), an expressive description logic-based language standardized by the World Wide Web Consoritum (W3C) (Bechhofer et al., 2003). In particular, we make use of the extended annotation and meta-modeling features available in OWL 1.1 which can be exploited for modeling expressive resource models. In Section 5.2, we show how meta-modeling allows a resource to refer also to concepts and the

[1] The $\models$ symbol refers to logical consequence in the sense that it holds in all interpretations. For alternatives, the interested reader is referred to (Sebastiani, 1996)

axiom annotation feature allows a subject to be described by an arbitrary OWL axioms.

4.2 Query Model

The Query Model $\mathcal{Q}_O$ consists of elements that correspond to ontology elements in $\mathcal{O}$. For the queries we distinguish between user queries (expressed in a language $\mathcal{L}_U$) that are posed by the end user and system queries (expressed in a language $\mathcal{L}_S$) that are used for actual evaluation of the query by the system.

User Query Model While the system query is expressed in terms of elements of the ontology language, we do not further constrain the representation of the user query. The user query can for example be represented as keywords (Clarke et al., 2000) or a natural language question. Yet, we assume the correspondence between elements of the user query with the ontology elements in order to be able to translate the user query into a logic-based system query (query interpretation). With respect to the models defined above, the interpretation of a query can be defined as a mapping from the user query to the system query. In Section 6, we will discuss how keywords-based and natural language questions can be translated into a logic-based system query. Note that the more related the syntax and semantics of the user query language $\mathcal{L}_U$ and the system query language $\mathcal{L}_S$, the more straightforward is the mapping. Clearly, when $\mathcal{L}_U$ is the same as $\mathcal{L}_S$, such a translation is not required (this is the case for sets of keywords as queries).

System Query Model The second task after query interpretation is answering the system query. In doing so, the system query is evaluated via an entailment between the document descriptions and the information need: $d_j \models_O q_i$. Our notion of entailment is held abstract on purpose. As we will discuss, it can for example be realized using standard description logic reasoning tasks. In essence we thus reduce the IR problem to an instance retrieval problem and in particular to entailment between the logical representation of the document and the one of the query. A similar logic-based view on IR has already been presented by van Rijsbergen (1986) and later by Meghini et al. (2001). In our concrete realization, for the system queries we rely on conjunctive queries, a common language for querying DL-based ontologies. A conjunctive query is defined as a conjunction of terms of the form $C(x)$ or $R(x, y)$, where C is a concept, R is a role, and x, y are variables or individuals. In other words, this query language allows to constrain the result set to individuals of some specific types, interrelated via specific relations or carrying specific attributes. As a concrete syntax for encoding conjunctive queries, we rely on SPARQL, again a standard proposed by the W3C [2].

As a result, query elements refer to a full-fledge ontology and the query engine can exploit entailment relations to infer new knowledge, e.g. to classify content resources, to exploit the concept hierarchy for query expansion and to exploit concept description for disambiguation. We give examples of such queries in Section 6.

Ranking Our model also abstracts from a specific ranking function. In the most simple case, our model corresponds to the so-called boolean document retrieval in case a standard entailment without any notion of relevance is used. Yet, we can also apply a different ranking functions by either relying on non-crisp entailments, e.g. by making use of a logical language that allow for ranking using a relevance terminological logic (Meghini et al., 2001) or by relying on ranking mechanisms such as defined by Ding et al. (2005) or by Siberski et al. (2006).

5 AN EXPRESSIVE SYSTEM RESOURCE MODEL

In this section, we describe in detail the adopted system resource model which is formalized relying on the Web Ontology Language OWL. The main hierarchy of concepts of the corresponding OWL ontology is shown in Figure 1[3], where the black arrows indicate that there are some concepts being excluded. In essence, the ontology constrains the relations as well as their domain and range which can be defined between individuals denoting entities and resources, which are represented as a `content bearing object` (CBO). The distinction between actual `content` and `content bearing object` is in fact a crucial design choice in our resource model. In what follows, we briefly describe these concepts

[2] `http://www.w3.org/TR/rdf-sparql-query/`

[3] The complete ontology can be downloaded at `http://ontoware.org/frs/download.php/315/oir.owl`

in more detail and then provide a set of examples which illustrate their usage.[4] Note that all ontology elements are identified with a Uniform Resource Identifier (URI). In particular, for the sake of conciseness, we use abbreviated URIs using namespace prefixes, e.g. `oir` abbreviates the *http://www.aifb.org/2007/05/oir/*, which is the prefix of all the elements of the OIR ontology. In order to ensure compatibility with existing standards, many elements defined in standardized vocabularies such as XML Schema, the Dublin Core (DC) schemas as well as the Suggested Upper Merge Ontology (SUMO), the MPEG-7 ontology and Simple Knowledge Organisation Systems[5] (SKOS) have been reused. These elements are imported into the proposed ontology and prefixed by "*xsd:*", "*dc:*", '*sumo:*", "*mpeg:*" and "*skos:*", respectively. We will discuss how the use of these standards can facilitate interoperability of resource models across applications and domains in Section 5.3.

5.1 Definition of the Resource Model

For the description of resources in OIR systems, we distinguish three different aspects that are relevant, namely the *content*, the *structure*, and the *presentation*. Given a description containing all these aspects, resources can be retrieved based on structural properties, content-related as well as presentation-related information. To capture these aspects, we employ a conceptual distinction: a resource is actually modeled through two entities, i.e. an instance of `content` and an instance of `content bearing object` (`CBO`). While `CBO` captures presentation-related information, `content` contains information related to the resource's content, e.g. the subject and the topic.

In the following, we define these concepts through a set of axioms using the standard DL syntax (Horrocks and Patel-Schneider, 2003). While the actual ontology contains many more axioms, the concepts we present here simply define the specific relations, i.e. object and data properties, which may be instantiated between a resource individual and other domain entities. As shown in Formula 1 for instance, for the class `CBO`, using existential quantification we define that an instance of `CBO` is required to have a minimal set of properties, i.e. it should contain at least some information of the type `content` (∃oir:contains_information.oir:Content). Using universal quantification, we specify that if an instance of `CBO` has a particular property, this must have a particular range, e.g. a `CBO` can have only a `CBO` as part (∀oir:has_part.oir:CBO).

The Content Bearing Object In our model, content is assumed to be abstract in the sense that it can be materialized in different media types such as audio and text, using different layouts, color schemes etc. The `CBO` concept is introduced to describe the physical properties of the resource that bears the content. In particular, `CBO` is concerned with all presentation-related aspects of the resource in question. The ranges of these properties are descriptors as specified in the MPEG-7 standard and modeled in the MPEG-7 ontology (Hunter, 2001). A CBO is related to an (abstract) content object through the property `contains_information`. It may be further described by a `title`, by the `language` it is expressed in, by its `publisher` and associated rights (e.g. intellectual property and access rights modeled in the form of `Permission` and `Credential`). Besides these standard metadata and presentation-related information, this concept also captures structural information through the properties `has part` and `is part of` which together define a resource as a complex object which can have subparts.

$$
\begin{aligned}
\text{oir:CBO} \sqsubseteq\ & \\
& \exists \text{oir:contains\_information.oir:Content} \sqcap \\
& \exists \text{oir:size.xsd:byte} \sqcap \\
& \exists \text{oir:format.oir:Format} \sqcap \\
& \forall \text{dc:publisher.sumo:Agent} \sqcap \\
& \forall \text{oir:creation\_date.xsd:date} \sqcap \\
& \forall \text{dc:language.xsd:language} \sqcap \\
& \forall \text{dc:title.xsd:string} \sqcap \\
& \forall \text{oir:has\_part.oir:CBO} \sqcap \\
& \forall \text{oir:is\_part.oir:CBO} \sqcap \\
& \forall \text{oir:color.mpeg:Color\_Descriptor} \sqcap \\
& \forall \text{oir:shape.mpeg:Shape\_Descriptor} \sqcap \\
& \forall \text{oir:texture.mpeg:Texture\_Descriptor} \sqcap \\
& \forall \text{dc:rights.sumo:Permission} \sqcap \\
& \forall \text{dc:access\_rights.oir:Credential}
\end{aligned}
\tag{1}
$$

The Content While `CBO` primarily captures presentation-related information, the (abstract) content itself is represented by the `content` concept. Besides standard content metadata, the content is mainly defined through two different aspects, i.e. the *content's subject* and the *content's topic*.

[4] We refer the interested to the actual ontology available on the web for the explicit formalization of these and other concepts

[5] `http://www.w3.org/2004/02/skos/`

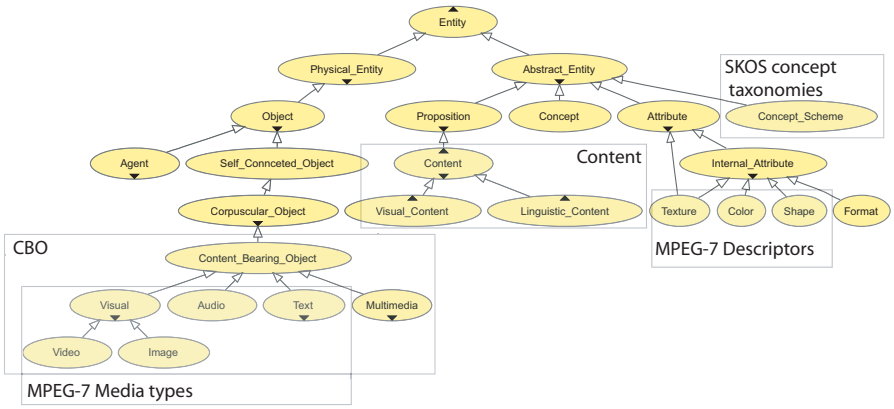

Figure 1: OIRonto Concept Hierarchy

As specified in Formula 2, a `content` resource must be embodied in some `CBO`. The *topic* of a `content` is defined through the property `topic`, which relates instances of `content` to instances of some `concept`. By this, we can describe content via any taxonomy or classification hierarchy which specify hierarchical relations between concepts as specified by the `broader` and `narrower` properties defined in the SKOS vocabulary.

$$\begin{aligned} &\text{sumo:Content} \sqsubseteq \\ &\quad \exists \text{sumo:embodied\_in.oir:CBO} \sqcap \\ &\quad \exists \text{oir:author.sumo:Cognitive\_Agent} \sqcap \\ &\quad \exists \text{dc:subject.sumo:Entity} \sqcap \\ &\quad \forall \text{oir:topic.skos : Concept} \sqcap \\ &\quad \forall \text{dc:source.sumo:Content} \sqcap \\ &\quad \forall \text{oir:authoring\_date.xsd:date} \end{aligned} \tag{2}$$

The `subject` of a content object is an instance of `entity`, which might refer to an individual of the domain ontology or even to a concept or more complex axiom. Thus, while a content could have as subject a specific individual such as the researcher *Philipp Cimiano*, it could also refer to the class of researchers in general and thus refer to the concept `researcher`. This is where our model is unique as previous work on using terminological logics to model IR do not allow to refer to non-individuals to describe the content of a resource. To some extent, this might be due to the fact that, if done in the wrong way, allowing to talk about concepts as first class citizens can lead to undecidability (see (Motik, 2005)). In our model we are able to talk directly about concepts via the meta-modeling capabilities of OWL 1.1. In particular, meta-modeling is realized in OWL 1.1 via a technique known as punning (defined in (Motik, 2005)) which does not render the underlying logic undecidable. With punning, concepts are also asserted as being instances of `concept`, e.g. the domain concept `people` is asserted as instance of `concept`. Thus, the subject of a content object can be captured by an object property assertion that relates a `content` instance with a `concept` instance.

In addition, the subject of a content object can be described by any complex axiom, e.g. a concept definition (terminological axiom), or an assertion about specific individuals (assertional axiom). Thus, any axiom can be annotated as being the subject of a particular `content` individual. In our framework, we use the property `is_subject_of` to point from within axiom annotations to the described resources. Thus, the subject of a content object can refer to specific individuals denoting domain entities, concepts or even complex axioms which can be expressed in the OWL language.

Having described the basic design choices and characteristics of our resource model, we now discuss a specific example which is in line with our introductory scenario.

5.2 Example Resource Description

We now illustrate the previous descriptions with an intuitive example. In addition to OIRonto (prefix "*oir:*"), which also contains imported elements of SUMO, DC and SKOS, the examples involve two further ontologies. These are the SWRC ontology (prefix "*swrc:*"), which is available as part of the AIFB Portal metadata[6]) and a fictional domain ontology (prefix "*dom:*"). The SWRC ontology contains a publication that

[6] `http://www.aifb.uni-karlsruhe.de/about.html`

matches the information need described in section 1, which is described as follows:

Example (The publication pub1942)**.** *The individual* ***pub1942*** *is a publication with the name "Ontology-based Question Answering for Digital Libraries". The author of* ***pub1942*** *is "Philipp Cimiano", who is member of a research group with the name "Knowledge Management". Further,* ***pub1942*** *is also described by further standard metadata such as language and creation date. Let us assume that the topics of* ***pub1942*** *have been specified to "Information Retrieval" and "Question Answering". The content of* ***pub1942*** *deals with "Orakel", a question answering system, that has been deployed at a company with the name "British Telecom".*

Now, we will show how the different aspects of this resource can be specified using the concepts defined above. While we focus on the resource's content, structure and presentation-related aspects can be modeled along the same lines.

Modeling the Resource's Metadata The example resource is captured through the individuals pub1492 and pub1492c. As defined below, the former is a CBO (3, 16), which is related to the latter, a content (4), through contains_information (5).

swrc:InProceedings(*pub1492*) (3)
oir:Content(*pub1492c*) (4)
oir:contains_information(*pub1492*, *pub1492c*) (5)

Through several assertions about these instances, the resource description is represented in terms of some property values. For instance, the title of the publication is "Ontology based..." (6), the language of the publication is English (7) and its creation date is 01/02/2007 (8).

dc:title(*pub1492*,"Ontology-based...") (6)
dc:language(*pub1492*,"English") (7)
oir:creation_date(*pub1492*,"01/02/2007") (8)
oir:author(*pub1492c*, *pers98*) (9)

Note that the author of the content part is the individual pers98 (9). The name of this individual is Philipp Cimiano (11), whose affiliation is an instance of a research group with the name Knowledge Management (12, 13 and 14). This knowledge about the author is described in the following axioms:

swrc:Person(*pers98*) (10)
swrc:name(*pers98*,"Philipp Cimiano") (11)
swrc:affiliation(*pers98*, *group3*) (12)
swrc:ResearchGroup(*group3*) (13)
swrc:name(*group3*,"Knowledge Management") (14)

Note that much of the knowledge about this resource is already described in the SWRC ontology. SWRC also contains some concepts to describe resources. SWRC resource models are made compatible with the resource model proposed here by mapping SWRC concepts and properties to elements of OIRonto. In particular, Formulas 16, 17 and 18 show how SWRC classes are integrated into OIRonto by a number of subclass mappings. For instance, (15) declares that the concept InProceedings is a subclass of CBO. Besides, further mappings asserting the equivalence of concepts and properties are also needed to make SWRC resource models fully compatible. In Section 5.3, we discuss in details further mappings that are required to integrate different standards into OIRonto.

swrc:InProceedings $\sqsubseteq$ oir:CBO (15)
swrc:Person $\sqsubseteq$ sumo:Cognitive_Agent (16)
swrc:ResearchGroup $\sqsubseteq$ sumo:Agent (17)

Modeling the Resource's Content As defined in the previous section, the resource content is described in terms of the topic it can be assigned to and the subject it deals with. The example below shows that the content instance pub1492c is related with two topic instances. Both these instances are described using SKOS, e.g., has preferred label Information Retrieval (18, 19 and 20) and Question Answering (21, 22 and 23).

oir:topic(*pub1492c*, *top152*) (18)
skos:Concept(*top152*) (19)
skos:prefLabel(*top152*,"Information Retrieval") (20)

oir:topic(*pub1492c*, *top153*) (21)
skos:Concept(*top153*) (22)
skos:prefLabel(*top153*,"Question Answering") (23)

dom:QASystem $\sqsubseteq$ sumo:Artifact (24)
dom:Corporation $\sqsubseteq$ sumo:Agent (25)

oir:subject(*pub1492c*, *dom:id555*) (26)
dom:QASystem(*dom:id555*) (27)
dom:name(*dom:id555*,"Orakel") (28)

oir:subject(*pub1492c*, *dom:id333*) (29)
dom:Corporation(*dom:id333*) (30)
dom:name(*dom:id333*,"British Telecom") (31)

For the subject description, we refer to concepts and instances of a fictional domain ontology, which in other scenarios might be about Health Care, Automobile or Politics. The knowledge of this domain is about `Orakel`, a question answering system (`QASystem`) (27 and 28) that is `deployed at British Telecom`, a `corporation` (30 and 31). In the above example, the subject is described simply by property assertions that refer to the instances `Orakel` and `British Telecom` (Formulas 26 and 29). In order assign more complex descriptions as subject to resources, axiom annotations can be used as shown in the example below.

dom:OIR $\sqsubseteq$ dom:QA $\sqcap$ dom:IR (32)
dom:deployedAt(*dom:id555*, *dom:id333*) (33)
oir:is_subject_of(
  dom:deployedAt(*dom:id555*, *dom:id333*),
  pub1492c) (34)
oir:is_subject_of(
  dom:OIR $\sqsubseteq$ dom:QA $\sqcap$ dom:IR,
  pub1492c) (35)

Axioms 32 and 33 specify that `OIR` is a combination (intersection) of question answering (`QA`) and `IR` and that `Orakel` is `deployed at British Telecom`. While the former represents a terminological axiom, the latter represents an assertional axiom. In fact, any axioms expressible in OWL can be annotated as the subject of a resource content.[7] The axiom annotations 35 and 36 for instance, assert that axioms 33 and 34 are subjects of `id1492_c`.

[7]We have introduced a special DL syntax for axiom annotation, for which only the abstract and XML syntaxes have been specified. Note that we do not annotate axioms locally as given in the specification but instead as external pointers to an axiom. Since the axiom has no URI, the structure of the axiom is used for this external reference.

5.3 On the Use of Standards

We have chosen OWL as the language for modeling resource descriptions. As shown previously, this language is sufficiently expressive to model many aspects of IR resources.

Furthermore, many existing ontologies and metadata standards have been incorporated. These standards can capture some of the aspects of system resources. When possible, they have been reused, aligned and combined to obtain an ontology-based resource model that can capture all the relevant aspects. We deem that this is a necessary step to establish a commonly agreed-upon model that is accepted by the community and can allow for interoperability among OIR systems. This interoperability is twofold: the interoperability with existing resource descriptions specified with these standards and the interoperability of the proposed ontology with domain ontologies.

The interoperability of resource descriptions is supported by alignments with the DC schema, the MPEG-7 ontology and SKOS. This means that we map concepts and properties from these standards to elements of the proposed ontology. This has been illustrated by the use of prefixes in the definitions of the resource model. For instance, elements with the prefix "mpeg" are MPEG-7 descriptors such as `texture`, `color` and `shape`. They have been reused to model presentation-related information of resources. Figure 1 shows that these concepts have been integrated into the ontology as subclasses of `internal attribute`. Also, other MPEG-7 concepts describing the different media types such as `audio`, `video`, `image` and `hypermedia` have been integrated in the same manner, i.e. specified as subclasses of `CBO`. While MPEG-7 is concerned with multimedia knowledge, SKOS allows to describe concept schemes such as thesauri, classification schemes, subject heading lists and taxonomies. The DC schema specifies a set of metadata attributes that can be used to describe resources. As indicated by the "dc" prefix, these DC attributes are directly reused as properties in the ontology, e.g. `title` and `language`.

Due to these alignments, it is possible for applications using our model to exchange information with other systems that support (one of) the mentioned standards. The example discussed in the previous section shows that the resource model is not only compatible with these standards, but is also sufficiently general to sup-

port proprietary metadata such as defined in the SWRC ontology. Note that due to the many axioms contained in our ontology, SWRC metadata have not been imported as simple data but "knowledge" that can be exploited by reasoners to derive new facts.

Since the resources are described by their relations to entities of domain ontologies via the `subject` property, interoperability across domains is also necessary. This means that all domain ontologies used in the resource models must be imported and integrated. For instance, the ontology with the prefix "dom" mentioned in the last section must be integrated for the example resource description to be processable by the system. This integration is expected to be facilitated by the use of the foundational ontology SUMO. The ontology proposed here is in fact an extension of SUMO. That is, all elements with the "oir" prefix extend `sumo:Entity`—and the subclasses `object`, `process` and `abstract entities` respectively (see Fig. 1). When domain ontologies are also such SUMO-compliant, i.e. share the same foundational conceptualization, conceptual mismatches are less likely. In this case, it is straightforward to map domain concepts to corresponding concepts of our ontology for cross-domain integration. In our example, for instance, the SWRC concept `person` has been mapped as a subclass of `cognitive agent` as specified in Formula 5.

6 THE RESOURCE MODEL IN PRACTICE

In this section, we elaborate on how our abstract notion of OIR can be used in practical applications. We start with a review of approaches for the automatic enhancement of documents with advanced descriptions such as envisioned in this paper. These ontology-based descriptions can then be queried to address a complex information need. As this information need might be specified with a formalism different than the final query language, we will discuss how ontology knowledge can also be used to translate such a resulting user query into a system query. Finally, we discuss the introductory example with respect to how it can be accounted for with the expressive system queries that are supported by the proposed resource model.

6.1 Developing Resource Descriptions

The proposed resource model comprises ontology elements that can be used to describe multimedia resources. In order to obtain expressive document descriptions based on this model, a manual approach can be undertaken. That is, the user specifies the metadata, the structure, the topic, the subject etc. in a manual way. However, there are a number of approaches which can support the automatic extraction of document descriptions with respect to an ontology-based model. While the extraction of document metadata and structure information seems feasible given the current state-of-the-art, the extraction of semantic content is indeed critical. Nevertheless, there are initial blueprints which show that capturing the documents' semantic content at a large scale might be feasible. First of all, in our own work we have shown that it is possible to train efficient classifiers which associate documents to a certain topic of a given taxonomy (Bloehdorn et al., 2007). Furthermore, several approaches have been recently presented to extract relations from large sets of documents such as Wikipedia (Blohm and Cimiano, 2007) and even from the Web, e.g. the Pronto (Blohm et al., 2007) and the TextRunner (Banko et al., 2007) systems. While the automatic identification of complex situations is still difficult, the above blueprints show that it is feasible to extract instances, relations (assertional axioms) and to assign documents to fine-grained topics. In all, this represents a good starting point towards the OIR as described in this paper where documents are enhanced with rich semantic descriptions.

6.2 Interpreting User Queries

In the formalization of OIR, we have deliberately abstracted from the way a user can specify an expressive query. However, in practice, formulating expressive queries in the formal language of the system is not an easy task, especially because most users are used to simple Google-like interfaces. There are different alternatives here. For instance, the user can be supported at the user interface level, e.g. through special forms.

Another possibility is to allow users to specify standard keyword queries which can then be transformed into SPARQL queries with respect to the ontology. Such an approach has been presented for example in (Tran et al., 2007). In a

different work, we have shown that full natural language questions can also be translated—with a reasonable performance between 70% and 100% of accuracy (Cimiano et al., 2007).

6.3 Answering User Queries

The generic query language formalized in our OIR model is now grounded to SPARQL. We now illustrate how the example resource description from section 5.2 can be queried to address the information need in the introductory section. Note that since this example does not contain any quantitative measures, e.g. confidence degree of description elements, the evaluation of the queries discussed here correspond to the basic boolean model such that the ranking function boils down to $R(q_i, d_j) = 1$ *iff* $d_O \models_O q'_O$ otherwise $R(q_i, d_j) = 0$.

Given the information need, the user knows (or specifies) only some of the aspects of the resources that might satisfy this need. For instance, the topic `Information Retrieval` is known and thus, might be part of the query in 37. This query simply returns all resources (`CBO`) associated to this topic.

```
SELECT ?r WHERE {
   ?r oir:contains_information ?c .
   ?c oir:topic ?t .                          (36)
   ?t skos:prefLabel 'Information Retrieval'
}
```

Given the same need, a different user might know only something about the author. This user is not able to name any author but just requires the author to be part of the `Knowledge Management` group. Also, the user specifically requires the returned results to be of the type `publication`. The corresponding query is given in 38. Note that answering this query already requires inference capabilities as in the example description the resource is specified to be of the type `CBO`. The engine must be able to infer that this resource is also an `entity`

```
SELECT ?r WHERE {
   ?r rdf:type sumo:Entity .
   ?r oir:contains_information ?c .
   ?c oir:author ?p .                         (37)
   ?p swrc: affiliation ?g .
   ?g swrc:name 'Knowledge Management'
}
```

Now, a further user might want to retrieve resources based on the content semantic. In standard IR systems, the "content semantic" is only available in form of a bag of keywords. Thus, the user would need to enter suitable keywords, e.g. "question" and "answering" to retrieve resources that deals with question answering. The SPARQL query in 39 would produce similar results, but in this case the returned resources are more specifically required to have question answering systems as subject. Note that the example resource model is also returned because its subject is `Orakel`, which can be inferred by the engine to be of the type `QASystem`.

```
SELECT ?r WHERE {
   ?r oir:contains_information ?c .
   ?c oir:subject ?s .                        (38)
   ?s rdf:type dom:QASystem
}
```

As a last example, we want to retrieve documents which describe some question answering system deployed at some corporation. Addressing this information need precisely is not straightforward and requires a more expressive resource model such as presented in this paper. The query that can achieve this result is given in 40, which exploits the axiom annotations of the example resource description. Note that in order to query these annotations, we need a mechanism to refer to the specific axiom dom:deployedAt(*dom:id555*, *dom:id333*). This can be achieved by so-called meta-ontologies (Vrandecic et al., 2006), in which ontology axioms are reified as instances, e.g. the axiom in this example is reified as an instance of `object property assertion`. The second part of the query in 40 illustrates how axioms (`?ax`) in these meta-ontologies can be addressed—note the prefix "*axns:*", which denotes references to meta-ontologies. For the interested reader, please refer to *MetaViews*, a recent proposal for such meta-ontologies (Motik et al., 2007).

```
SELECT ?r WHERE {
   # query knowledge in ontologies
   GRAPH <ontology> {
      ?r oir:contains_information ?c .
      ?sys rdf:type dom:QASystem .
      ?corp rdf:type ?dom:Corporation
   }
   # query knowledge in meta-ontologies
   GRAPH <ax:ontology> {                      (39)
      ?ax oir: is_subject_of ?c .
      ?ax rdf:type
         axns:ObjectPropertyAssertion .
      ?ax axns:objectProperty
         dom:deployedAt .
      ?ax axns:sourceIndividual ?sys .
      ?ax axns: targetIndividual ?corp .
   }
```

7 CONCLUSION

Many researchers have argued that in order to express and answer more complex information needs, we need to provide more expressive resource and query models allowing for a precise match between content and information needs. We have argued that moving to more expressive models requires to reformulate the IR task as a data retrieval task. Towards this end, in our work we build on earlier work on formalizing the retrieval problem as one of determining logical implication between a document and a query. Our OIR model, however, differs from earlier work in that it explicitly relies on domain knowledge captured in the form of ontologies which can be used at retrieval time to infer non-explicit relations. Our main contribution thus lies in an ontology that can be used to model expressive resource descriptions.

Certainly, we still have a long way to go towards achieving full fledged OIR as described in this paper. In this direction, our work can be understood as a proposal towards a resource model that different systems can share and which integrates existing standards. In particular, we have provided a model which on the one hand builds on semantic web standards such as OWL 1.1 and the query language SPARQL and on the other hand integrates various accepted ontologies and schemas into our model, e.g. Dublin Core for standard metadata and MPEG-7 for multimedia aspects.

Individual parts of our OIR framework have already been implemented and successfully applied (Tran et al., 2007; Bloehdorn et al., 2007). In future work, we intend to further advance the integration of the different components to achieve end-to-end OIR with respect to highly expressive resource descriptions.

ACKNOWLEDGEMENTS

This work was partially supported by the European Commission under contract IST-FP6-026978 X-Media and IST-2006-027595 NeOn.

REFERENCES

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P., editors (2003). *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, Cambridge.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2670–2676.

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D., Patel-Schneider, P., Stein, L., et al. (2003). OWL Web Ontology Language Reference. W3C Candidate Recommendation 18 August 2003. http://www. w3. org/TR/2003/CR-owl-ref-20030818.

Bloehdorn, S., Cimiano, P., Duke, A., Haase, P., Heizmann, J., Thurlow, I., and Völker, J. (2007). Ontology-based question answering for digital libraries. In *Proceedings of the 11th European Conference on Research and Advaced Technologies for Digital Libraries, Budapest, Hungary, September 16-21 2007*. Springer.

Blohm, S. and Cimiano, P. (2007). Using the web to reduce data sparseness in pattern-based informatin extraction. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*. to appear.

Blohm, S., Cimiano, P., and Stemle, E. (2007). Harvesting relations from the web -quantifiying the impact of filtering functions. In *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI-07)*, pages 1316–1323.

Cimiano, P., Haase, P., and Heizmann, J. (2007). Porting natural language interfaces between domains – a case study with the ORAKEL system –. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, pages 180–189.

Clarke, C., Cormack, G., and Tudhope, E. (2000). Relevance ranking for one to three term queries. *Information Processing and Management*, 36(2):291–311.

Ding, L., Pan, R., Finin, T. W., Joshi, A., Peng, Y., and Kolari, P. (2005). Finding and ranking knowledge on the semantic web. In Gil, Y., Motta, E., Benjamins, V. R., and Musen, M. A., editors, *International Semantic Web Conference*, volume 3729 of *Lecture Notes in Computer Science*, pages 156–170. Springer.

Fuhr, N. (1995). Modelling Hypermedia Retrieval in Datalog. *Hypertext-Information Retrieval-Multimedia, Synergieeffekte elektronischer Informationssysteme*, 20:163–174.

Horrocks, I. and Patel-Schneider, P. (2003). Reducing OWL entailment to description logic satisfiability. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 17–29.

Hunter, J. (2001). Adding multimedia to the semantic web - building an mpeg-7 ontology. In *Proceedings of the First Semantic Web Working Symposium*.

Meghini, C., Sebastiani, F., and Straccia, U. (2001). A model of multimedia information retrieval. *Journal of the ACM*, 48(5):909–970.

Meghini, C., Sebastiani, F., Straccia, U., and Thanos, C. (1993). A model of information retrieval based on a terminological logic. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, pages 298–307.

Motik, B. (2005). On the properties of metamodelling in OWL. *Proceedings of the 4th International Semantic Web Conference*.

Motik, B., Grau, B. C., and Horrocks, I. (2007). Making metalogical information in ontologies logical using metaviews. Technical report. Submitted for publication.

Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems FOIS'01*.

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., and Goranov, M. (2003). KIM–Semantic Annotation Platform. In *Proceedings of the 2nd International Semantic Web Conference (ISWC)*.

Sebastiani, F. (1996). A note on logic and information retrieval. In *Proceedings of the Workshop on Multimedia Information Retrieval (MIRO´95)*.

Siberski, W., Pan, J. Z., and Thaden, U. (2006). Querying the semantic web with preferences. In *Proceedings of the 5th International Semantic Web Conference (ISWC)*, pages 612–624.

Tran, T., Cimiano, P., Rudolph, S., and Studer, R. (2007). Ontology-based interpretation of keywords for semantic search. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*. to appear.

Vallet, D., Fernández, M., and Castells, P. (2005). An ontology-based information retrieval model. In *Proceedings of the 2nd European Semantic Web Conference (ESWC)*, pages 455–470.

van Rijsbergen, C. (1986). A new theoretical framework for information retrieval. *ACM SIGIR Forum*, 21(1-2):23–29.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths, London, 2 edition.

Vrandecic, D., Völker, J., Haase, P., Tran, D. T., and Cimiano, P. (2006). A metamodel for annotations of ontology elements in owl dl. In Sure, Y., Brockmans, S., and Jung, J., editors, *Proceedings of the 2nd Workshop on Ontologies and Meta-Modeling*, Karlsruhe, Germany. GI Gesellschaft für Informatik.

Weibel, S. (2000). The Dublin Core Metadata Initiative. *D-Lib Magazine*, 6:12.