publications

MDPI

*Article*

# Data Science Tools for Monitoring the Global Repository Eco-System and its Lines of Evolution

**Friedrich Summann *, Andreas Czerniak, Jochen Schirrwagen and Dirk Pieper**

LibTec Department, Bielefeld University Library, 33615 Bielefeld, Germany;
andreas.czerniak@uni-bielefeld.de (A.C.); jochen.schirrwagen@uni-bielefeld.de (J.S.);
dirk.pieper@uni-bielefeld.de (D.P.)
* Correspondence: friedrich.summann@uni-bielefeld.de; Tel.: +49-521-106-2631

Received: 15 February 2020; Accepted: 17 June 2020; Published: 22 June 2020

check for updates

**Abstract:** The global network of scholarly repositories for the publication and dissemination of scientific publications and related materials can already look back on a history of more than twenty years. During this period, there have been many developments in terms of technical optimization and the increase of content. It is crucial to observe and analyze this evolution in order to draw conclusions for the further development of repositories. The basis for such an analysis is data. The Open Archives Initiative (OAI) service provider Bielefeld Academic Search Engine (BASE) started indexing repositories in 2004 and has collected metadata also on repositories. This paper presents the main features of a planned repository monitoring system. Data have been collected since 2004 and includes basic repository metadata as well as publication metadata of a repository. This information allows an in-depth analysis of many indicators in different logical combinations. This paper outlines the systems approach and the integration of data science techniques. It describes the intended monitoring system and shows the first results.

**Keywords:** repository network; monitoring; repository observer; data science

## 1. Introduction

The repository landscape and the emerging associated community began its development around 2001 in the context of the journal crisis, the Berlin Declaration, and in some respects, the first definition of the Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH) protocol in 2001 [1]. From our librarianship point of view, our first contact with repositories was when we were looking for academic metadata to set up a search engine for scientific content. To be honest, it was in some ways a happy coincidence that Bielefeld Academic Search Engine (BASE) decided to opt for OAI-PMH and skipped crawling websites of academic institutions. In fact, the number of these OAI repositories increased significantly in the first phase, and in comparison, it became clear that repositories provided a much better quality of bibliographic information than crawled websites. This was the main reason why BASE (Bielefeld Academic Search Engine)—originally designed as a library meta-search-system—switched to the upcoming repository infrastructure and started collecting metadata not only about publications but also about the data sources from which they originated (attributes of the repositories).

### 1.1. BASE as Repository Aggregator System

BASE started its first development activities in 2001 [2] and launched its search interface in production in June 2004, with a modest number of ten repository resources based on OAI-PMH [3] and ten other data resources supplied by file delivery. The first index included a number of 500,000 publications. At that time, the use of the OAI-PMH opened the possibility to switch to bibliographic metadata with a certain level of quality and with a focus on metadata records describing publications. During this phase it was not yet clear whether this strategy would be successful. Nevertheless, BASE became a registered OAI Service Provider in 2005. But now, more than 15 years later, BASE indexes about 165 million records from more than 8000 repositories worldwide, making BASE a large, global academic search engine and repository aggregator.

Before indexing can begin, data must be harvested and aggregated, currently mainly via OAI-PMH. The technical environment for this consists of an open source harvester adopted very early which is still doing its work in a widely original state. It had to be slightly modified and adopted for specific behavior of resources and in this way to circumvent or fix various problems and peculiarities. These include errors in the use of character sets and protocols as well as communication interruptions. The returned files are stored in the file-system. Currently there are more than 11,000 repositories harvested with 8,020 repositories active. This generates an amount of 378 Mill. records with 227 Mill. of them unique. The size of data binds 2.5 terabyte of storage space (see Table 1 for more details). Because of the size and global coverage, the data form a specific kind of treasure, ready for all forms of metadata analysis. To automate the harvesting process for the proven stable repositories, a daily cron job controls the harvest of 6,200 repositories, most of them in a weekly turn.

**Table 1.** BASE environment profile (April 2020).

| | **Size** |
|---|---|
| **Indexed Documents** | 166.2 mil |
| **Open Access Documents** | 76.4 mil |
| **Documents with CC Licences** | 29.6 mil |
| **Documents with DOIs** | 49.6 mil |
| **Documents with ORCID** | 0.22 mil |
| **Sources** | 7888 |
| **Countries** | 135 |
| **Storage Size Index** | 2.5 Terabyte |
| **Storage Size Metadata Store** | 1.2 Terabyte |

BASE supports three ways of reusing its data. The end-user search interface can be integrated in external environments via http. The retrieval Application Programming Interface (API) (restricted to registered users) has been available since 2007. In addition, an OAI-PMH download (also for registered users only) was introduced in 2012. It was extended with a File transfer protocol (FTP) download of the prepared response files due to the amount of data. This service enables a stable and fast initial file download and provides a basis for subsequent incremental additions based on the OAI-PMH.

In 2011 BASE switched from commercial search engine software to an open source platform based on Lucene Solr for the backend and VuFind for the frontend. Other services that have been added include the boosting of Open Access publications (2014), the normalization of the Open Access status and license information (2016) and the optimized extraction of Digital Object Identifiers (DOIs) and Open Researcher and Contributor IDs (ORCIDs) from metadata (2019).

In 2019 a multi-node infrastructure based on Lucene Solr was implemented. This approach could significantly improve the performance of the system and overcome the limitations on the number of indexed metadata records. On this basis, work is currently underway to further integrate extensive publication references.

### *1.2. BASE and Its Role in the Repository Community*

The further development of BASE [4] has increasingly focused on repositories and adjacent systems and has thus accompanied the rapid development of the repository network. BASE itself has become a part of the repository community and has built up a lot of expertise in this area. This has been accompanied by participation in a number of national (Automatic Enrichment of OAI Metadata [5], ORCID-DE [6]) and European projects (DRIVER [7], EuropeanaCloud [8] and OpenAIRE [9]). This is the main reason why the information collected by BASE can be used as a mirror of the repository network environment from the beginning [10]. In retrospect, the most important step for this approach was to create and maintain snapshots of the key information describing this environment on a periodical rate. BASE data have been continuously maintained since 2004 (Figure 1 and thus the stored data describe and reflects the timeline of changes and developments in this area over a long period of time. Other key players in the repository network such as OpenDOAR [11] follow similar approaches. What makes BASE unique, however, is its temporal coverage, global approach and the large amount of related publication metadata available for data science-based analysis methods. BASE is not a registry like openarchives.org, OpenDOAR, or re3data that lists repositories and their attributes. BASE also lists repositories, but on a larger and broader scale, and this feature is a by-product of end-user functionality. BASE includes only resources with a functional technical interface, which must meet certain technical requirements and must provide correct and complete metadata. All types of scientific resources are accepted, i.e., BASE includes journals and journal portals, digital collections and other types of resources. Repositories are constantly changing their general properties and technical conditions, and therefore BASE must constantly track these changes to ensure that data harvesting and subsequent end-user presentations are correct. This permanent process results in a much broader range of descriptive information. Currently, OAI-PMH is still clearly the predominant protocol for harvesting metadata. However, recent developments such as resourceSync can be flexibly added by adding a harvesting plugin to the BASE workflow. The example of the Crossref API has shown that such an integration can be realized quickly and easily. A strong reason for the comprehensive coverage of the global repository landscape are the different channels of data resource ingest. If possible, automatic processes are installed to use the established repository registries such as openarchives.org, OpenDOAR, ROAR and re3data to detect new or deleted repositories or updated attributes for repositories that are already included. In a similar way many different information resources are observed. These include national repository hosting services from Sweden, Norway, France and Japan and technical platform communities such as DSpace, Islandora, OJS, Digital Commons, and many others. All these providers run web-sites and provide lists of their members in different formats and these have to be processed individually with different solutions. Finally, the processing of new data with already indexed repositories is automatic. Table 1 lists some descriptive data that show the information space that BASE works with.

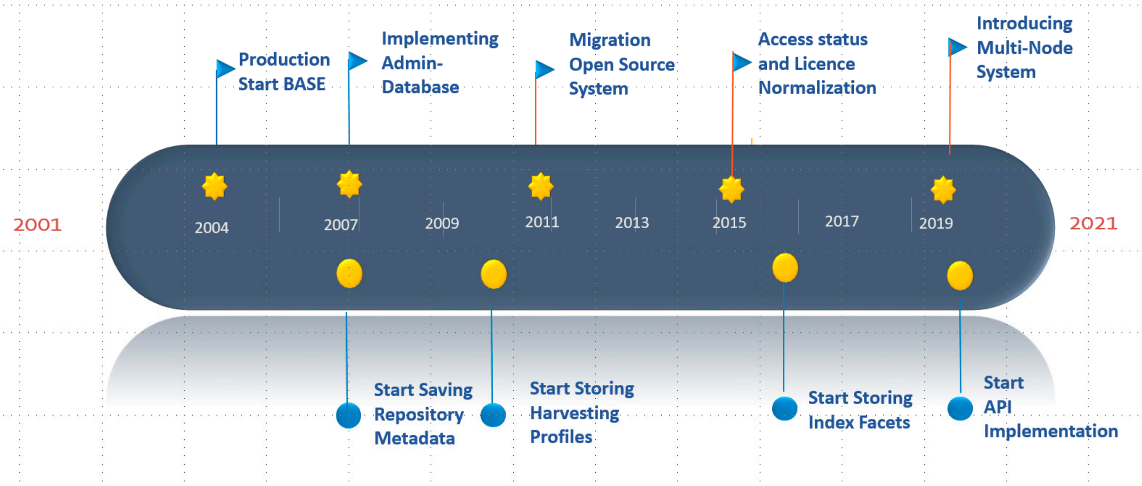

Production Start BASE
Implementing Admin-Database
Migration Open Source System
Access status and Licence Normalization
Introducing Multi-Node System
2001
2004
2007
2009
2011
2013
2015
2017
2019
2021
Start Saving Repository Metadata
Start Storing Harvesting Profiles
Start Storing Index Facets
Start API Implementation

**Figure 1.** Main events of the BASE history.

Crucial to feeding the monitoring infrastructure was the extraction of the actual substantial figures from the wide range of available data sources. This should include the scope and quality of metadata, but also metadata describing the resources (usually the repositories).

Another relevant topic is the creation of analytical profiles of the technical backbone of the communication, especially with regard to the technical behavior of the interface for metadata delivery (in most cases OAI-PMH [12,13]).

In this context, it is particularly valuable that many of the above-mentioned data are stored over a long period of time, so that temporal (time-based) developments and changes can also be identified, described and derived. In this sense, it has been a fortunate coincidence that repository metadata has been collected since 2007. In connection with the introduction of the normalized Open Access status and licensing information in OAI metadata [14], a monthly storage of repository metadata and index facets has been carried out on a country and repository level, since 2016. The aspect of the timeline is extremely relevant for the calculation of numbers describing developments of different aspects and views.

A fundamental lesson learned from these activities is: the earlier and the more comprehensive the better. Often, we have found later that data should have been stored much earlier in order to prepare for new information needs.

The fundamental question in this context is: why is monitoring the repository network valuable and to which questions can monitoring provide answers or at least deliver analytical figures for further consideration and evaluation?

The activities in BASE began with the preparation of conference presentations within the repository community, which required profound figures to describe and support certain issues, especially the analysis of the repository situation at the global or country level. Certainly, this information is able to provide concrete insights into developments and strategic decisions that are discussed and prepared, especially at country and repository level. For example, strategic approaches such as the COAR Roadmap Future Direction for Repository Interoperability [15,16] or Plan S [17,18] could be prepared, monitored and reviewed in a variety of ways. The transparency of this information offers, among other things, further possibilities for deriving strategies for improvements in the area of Open Access transformation and for better topic classification. The quality of communication and publication metadata is also a good indicator of the need for improvements.

Ultimately, the scope of the data collected in terms of time horizon, sources and details of the information, together with the usability of the interface provided, allows competition with similarly positioned systems such as OpenAIRE and CORE. BASE does not seem to have a bad starting position in this respect.

## 2. Materials and Background

### *2.1. The Data Basis and its Pre-Processing*

In order to describe the complete network infrastructure, comprehensive data coverage is the crucial point. According to our experience, every kind of data is valuable, and in addition, storing timeline-related information is valuable as well. The data include not only the metadata that is provided by the particular repository but also the data generated when establishing and processing a connection to a repository endpoint. These observations over time periods are very useful for evaluation purposes.

In fact, the data are spread among different resources and creativity is needed to bring them together in a universally applicable common data schema.

Currently, our statistical information is composed of

- Repository Metadata incl. common resource description;
- Harvested Metadata (for publications and related objects);
- Log Files of the harvesting processes;
- the BASE index.

The basic data listed above are available in different forms (related to formats, scope and time dimension).

BASE has been collecting repository metadata since 2007, when a database solution was implemented in order to use these data for displaying content provider information in search results and in the list of indexed resources. Initially these data based on daily database dumps were stored annually and since 2015 at monthly intervals. Included are country of origin, federal state (only for Germany, Switzerland, Austria), continent, technical platform, repository type (based on a vocabulary developed and used in BASE), date of first indexing in BASE, Open Access policy (on the repository level), number of documents, number of OA documents (after normalization), and status of indexing (besides 'indexed' as default value various special states as 'removed', 'to be checked' etc.)

The analysis of the harvested metadata has been carried out daily since 2007 using an automatic tool for monitoring the harvesting environment. The tool provides as a log the number of documents (complete and cleaned with corrections and deletions) and the content of some OAI-PMH response information—repository name, supported metadata formats and the deleting strategy per repository. The evaluation process must be flexible; new developments such as the emergence of ORCID information in metadata allows monitoring and recording of its occurrence in a similar way to license types, for example. The ORCID identifier is a good and up to date example of a development that can be observed through a monitoring process. The stored metadata can provide extended numbers that show from which resource ORCID information originates and in which geographical and technical background. It is obvious that timeline based figures provide valuable information on how to support further dissemination. Other similar examples from the past in which new developments have been documented are DOI and license information and their increasing use in repositories.

BASE as a globally operating search engine includes a Lucene index, which offers API access since 2007. This BASE search index can be used as a database for analyzing the current state of the repository landscape. The API allows all kinds of combinations of available search aspects in the API syntax to construct a highly complex search query. This allows the execution of complex cascaded queries and evaluations for the current status, which go far beyond the information stored in the database. The essential point is to find the most appropriate strategy, how and by what means the comprehensive descriptive information can be extracted from the current search index. Consequently, any combination of search aspects is possible in any nesting supported by the index. Unfortunately, it is not possible to save the complete index as a snapshot and to use it for later complex analyses at earlier points in time. The simple reason for this is that the index is so large (currently more than 1.5 TB of storage space is required) that its storage, even at reasonable intervals, far exceeds the storage capacity of our technical system. Another serious problem would be that BASE has changed the index structure several times and in parallel has updated the software of the core system very frequently which would result in incompatibility for the query behavior. To compensate for this deficit BASE has stored relevant facets of query responses (for single repositories and countries). Available attributes include language, type, publication date, Open Access status, license information and DDC class codes (this information can be used to assign topics). Vocabularies for repository and publication types were derived for the BASE normalization process and clear use of search facets. For this purpose, the terms of certain fields from the available metadata were sorted by frequency and then the content-matching values were defined. It was therefore logical that this expertise should support the subsequent definition activities at national (DINI) and international level (OpenAIRE, COAR).

The index facets at country and repository levels for the normalized contents of BASE index fields have been stored as snapshots for evaluation purposes on a monthly rate since 2016. This situation needs to be reconsidered in order to complement further useful issues and to extend the extraction of data with more complex requests.

A log file of the harvesting processes is available for the period since 2009. The log file data contain status and error messages of the harvesting processes and can provide additional information on the technical behavior of the interfaces (availability, stability, error situations, etc.).

For this point, properties describing the technical behavior of the OAI-PMH interface (performance measuring, comprehensiveness and quality of metadata, batch size) can be extracted from the harvesting response files. It is planned to integrate this information and store it at periodic intervals (monthly).

In order to be able to connect the evaluation figures for the repository network with a broader understanding of the academic infrastructure, especially at the country level it makes sense to relate them with figures describing the scientific infrastructure. Those indicators have to be discovered, prepared and then related to the basic repository-related metadata. Such key figures for the academic landscape include the general population size, the number of scientists and students and expenditures in research per country. Those data are available from different stakeholders, especially from international organizations such as the OECD. The normalization of these figures is not a simple task, since the periods of evaluation are different and the definition of the data collection differs from country to country. This starting point makes a careful and transparent normalization process necessary in order to make the comparison process fair and well-founded.

At this point, it must be mentioned that differences in the comprehensiveness of metadata, the normalization processes and quality deficits of metadata can produce a certain degree of vagueness in general.

### *2.2. The Data Store*

After the data have been determined, it is transferred to an appropriate data schema and then saved in this format. In addition to the question of the data schema, technical aspects such as the underlying database solution and the resulting data structure in combination define the access options. With this extensive amount of data, comprehensive analyses of the repository network are possible. In order to access this raw data efficiently, it is necessary to merge them into a generic data schema and store it in a suitable storage system, i.e., a relational or a NoSQL system. Since the data structure is relatively clear and the data volume is manageable, this situation could be a good argument for a NoSQL solution. The illustration in Figure 2 shows the different modules of the planned system.

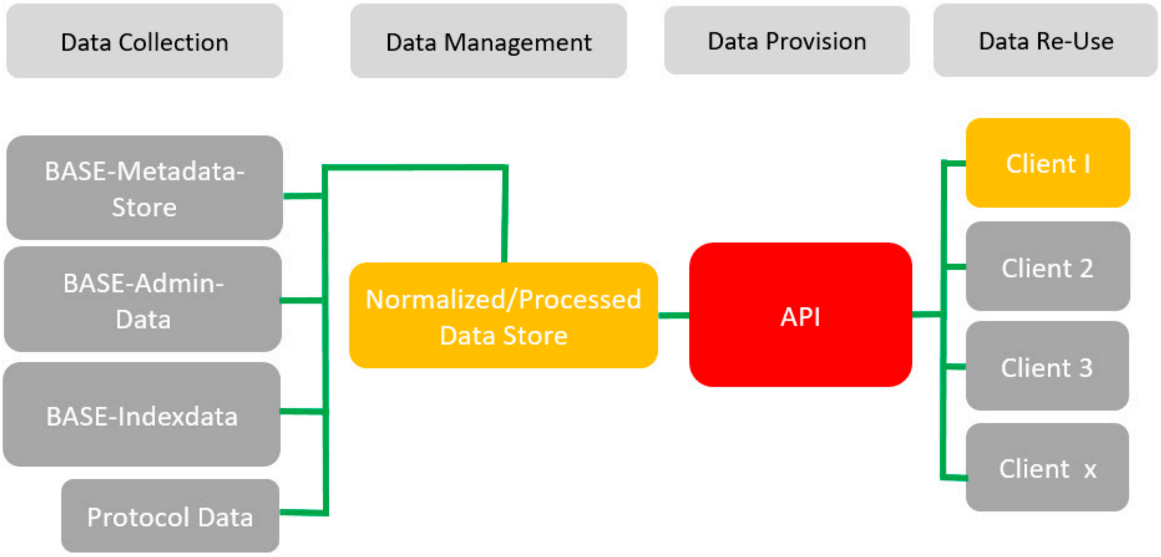


**Figure 2.** The data infrastructure of the monitoring system approach (n-tier structure).

Many of the system's components shown in the illustration already exist (the rectangles in orange), mostly in a rough state that needs to be optimized. The real challenge of the approach is to make the access robust and interoperable, and special emphasis must be placed on the new API core module (in red) as a door opener to return all the available figures to the repository community. The internal normalization and processing module need to be encapsulated, while the implementation of an internal client is useful for both BASE and the repository community. It has the potential to extend the BASE services and tests the robustness of the concept. In parallel, this prototype can demonstrate the capabilities of the approach as a prototype solution and provide valuable information without any further technical implementation effort for interested stakeholders. OpenDOAR and ROAR offer similar services but are obviously focused on the associated repository metadata they contain. BASE is quite uniquely conditioned by the amount and scope of content metadata it collects and the broad time span since 2004 in which it consistently stores essential analytical information from various sources.

### 2.3. *Visualization Tools*

The task of extracting and normalizing data is one side of the data science oriented activities described above. The other field in this area is the task of visualizing the available data infrastructure. Over time, a portfolio of solutions for different purposes has been developed. This has resulted in certain tools that have been further developed and for which extensive expertise has been built. It turned out to be advantageous that such approaches in the field of knowledge management where also available in other areas of Bielefeld University Library. This led to a synergetic exchange with other projects (especially in the area of Digital Humanities), which resulted in numerous further implementations.

In the first phase, the focus was on country-based evaluations. Because of the ease of use, Google Charts [19] was used, especially with regard to the output of maps. The results [Figures 3 and 4] proved to be quite usable, so that the possible analyses were increasingly extended and provided with a form-based application for internal purposes. This allows the setting of graphical parameters like color scale selection and value range, as well as the definition of content-based filter settings like OA status, repository type or percentage representation. The experience gained during the implementation has been directly fed back into the system design and its optimization.

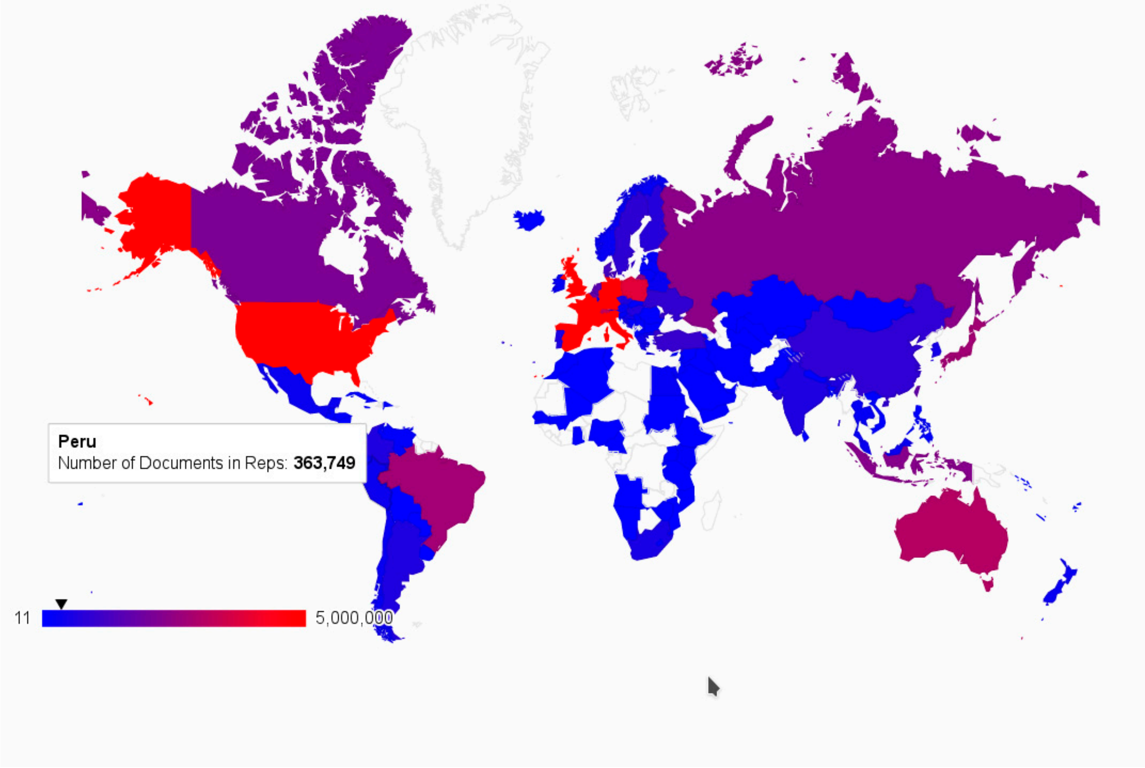


**Figure 3.** Total number of documents in repositories per country with the tooltip information for Peru.

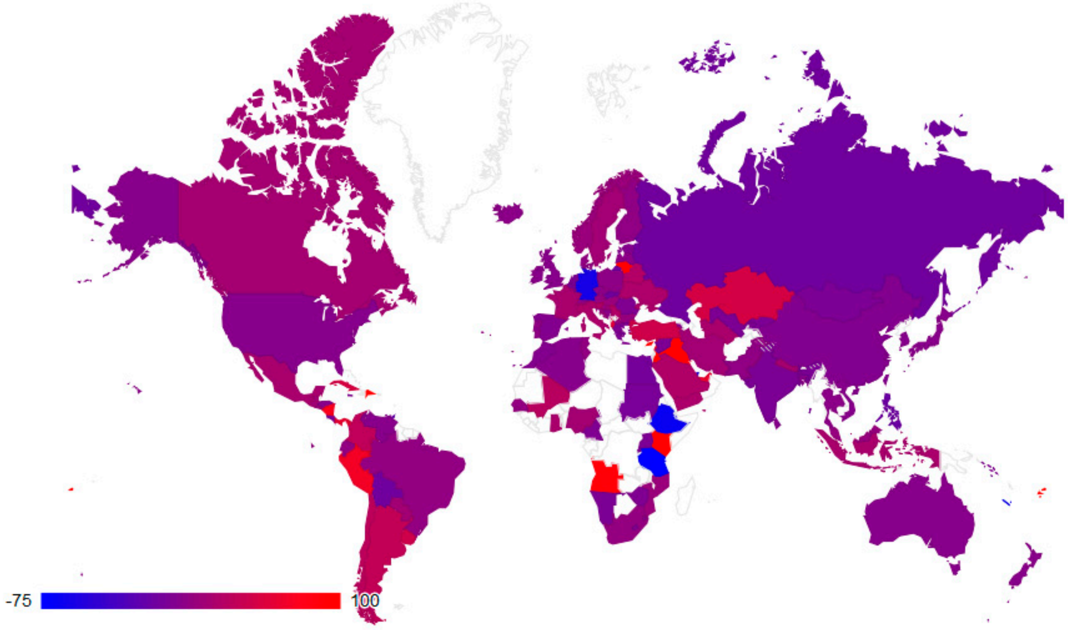
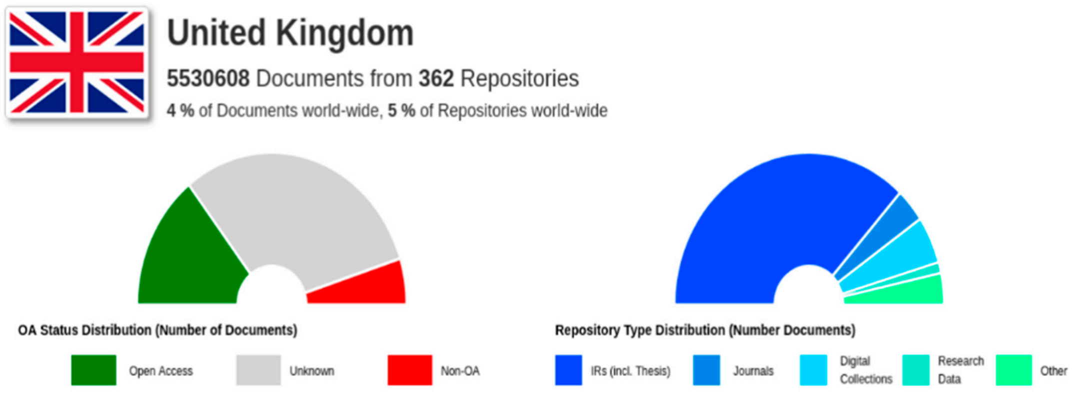


**Figure 4.** The in-/decrease of the number of repository documents per country in 2020 in percentage.

This work has been increasingly supplemented by efforts to create graphical analyses for different countries and for individual repositories. The aim is to create convincing visual implementations using appropriate techniques. The open source software D3js [20,21], a JavaScript framework with numerous well-documented examples for almost all use cases, has proven itself to be compatible with the existing script and template-based development environment. On this basis some adequate basic elements like pie charts, histograms, arcs and timelines extended with mouse-over and tooltip functionalities, have been implemented as documented in the examples of the following section.

The next world map (Figure 4) focuses on a much more demanding implementation since it covers the temporal change with regard to the decrease or increase of document numbers per country from January 2019 to January 2020. In contrast to the previous map, this map illustrates growth rates per country in percentages.

Based on metadata describing repositories, aggregating on the country level can give valuable insights. The following diagrams (Figures 5–7) use D3js as a visualization framework for descriptive statistics. Figure 5 shows a country profile for the United Kingdom repository landscape, displaying the number of repositories and documents. Additionally, the percentage of OA documents and the distribution of repository types is visualized.



United Kingdom

**5530608** Documents from **362** Repositories

**4 %** of Documents world-wide, **5 %** of Repositories world-wide

OA Status Distribution (Number of Documents)
Open Access
Unknown
Non-OA
Repository Type Distribution (Number Documents)
IRs (incl. Thesis)
Journals
Digital Collections
Research Data
Other

**Figure 5.** The national repository network profile for the UK, displaying some basic information such as number of documents and repositories, OA status and repository type distribution.

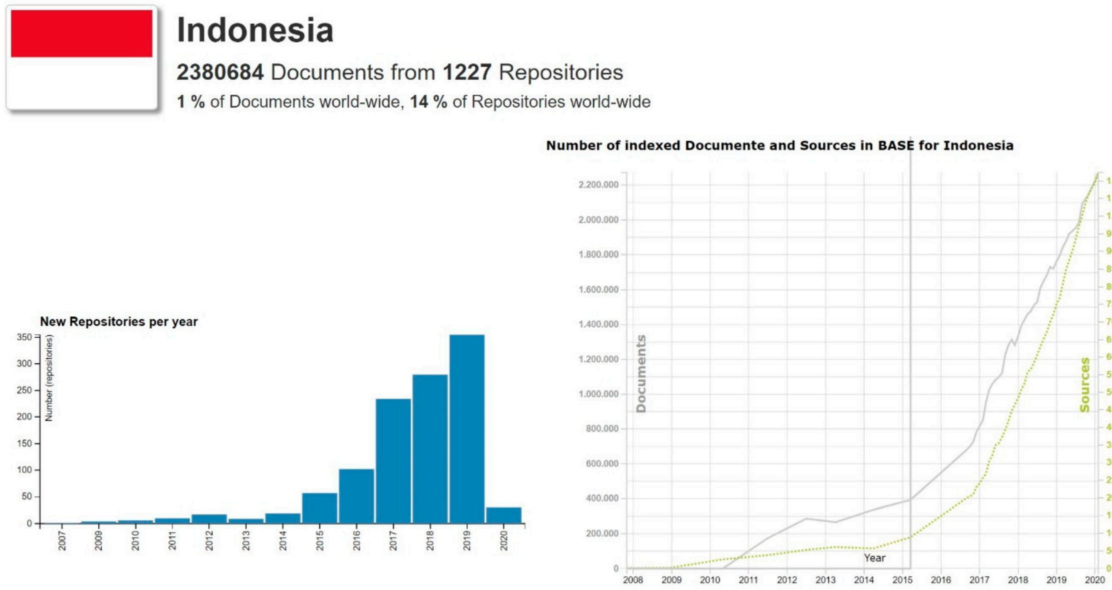


**Figure 6.** Timeline diagrams with the number of newly indexed repositories and the development of repositories and publications per year, in this case for Indonesia.

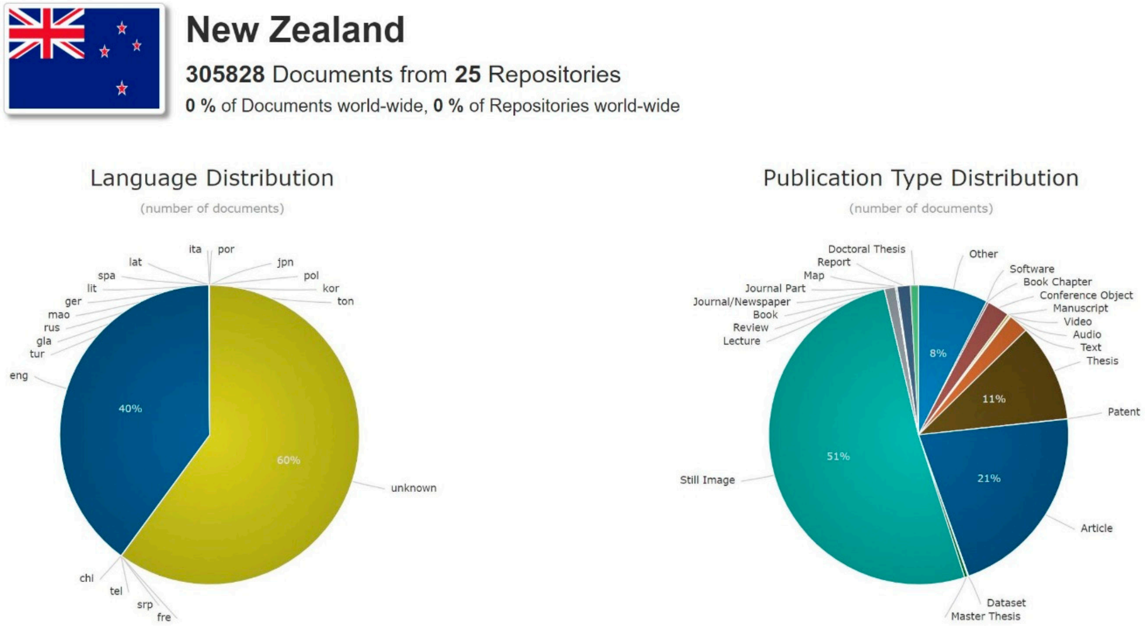
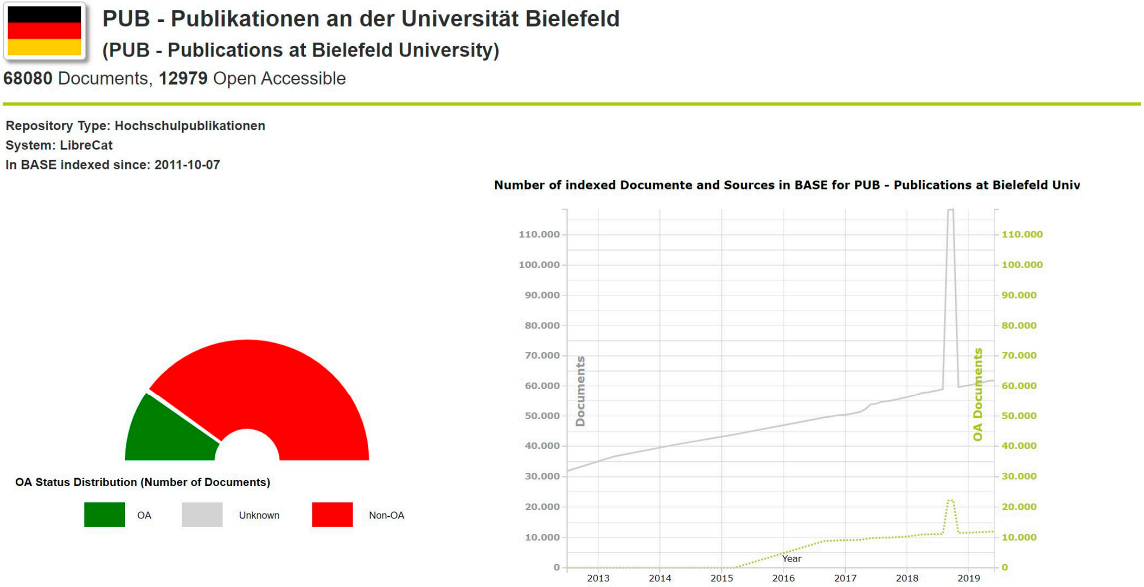


**Figure 7.** Current distribution of document languages and publication type in publications from New Zealand.

Time series data open up particularly evaluation aspects as they capture developments and provide support for strategic conclusions. Figure 6 demonstrates the usage of a timeline-based visualization for Indonesia. The diagram to the left shows the number of newly indexed repositories per year since 2008, and the one to the right shows the development of the absolute number of repositories and the number of documents. These diagrams rely on aggregated repository figures and give a taste of the significance of timeline-related information. Those diagrams can be computed and displayed for all countries with visible and indexed repositories.

The database is fed with specific metadata fields drawn from the BASE index via storing the facets of specific automatically processed search requests. Figure 7 shows the current distribution of the index fields language and publication type for the repositories from New Zealand (as of 1 February 2020). Both fields have been normalized during the pre-processing step before indexing. For language, the ISO 639-3 standard [22] and for publication types, a specific internal vocabulary derived from the metadata content is used. The pie diagrams are designed with the D3js [20,21] framework and integrate tooltips for detailed figures per value as well.

A similar profile can be computed on the repository level for each resource. Figure 8 shows the basic information for the Bielefeld University repository PUB and demonstrates the distribution of OA and restricted documents and the timeline development for the absolute number of documents and the OA percentage. Obviously, such information together with content evaluation results on repository level can help repository managers getting insight information to optimize their system.



PUB - Publikationen an der Universität Bielefeld
(PUB - Publications at Bielefeld University)
**68080** Documents, **12979** Open Accessible

Repository Type: Hochschulpublikationen
System: LibreCat
In BASE indexed since: 2011-10-07

**Figure 8.** Repository profile for the Bielefeld University repository PUB.

## 3. Outlook and Refinement of the API Layer

Currently, data for analysis purposes are drawn from computed analytical data and then stored as JSON files. This can be used basically for the analyses presented, and it is essential to implement a generic, interoperable hub for data that can operate as the core module. The implementation of an API should follow modern concepts of Web Information Systems [23] and allow complex queries.

A modern, generic and comprehensive API will replace the currently provided internal API. Only two types of APIs can be considered as a appropriate replacement. The first API type based on the REpresentational State Transfer (REST) [24] has become a standard for designing web APIs. The second API type is the relatively new emerging GraphQL [25,26] technology. GraphQL is now widely used and closer to modern developers and their technologies. It is already used in some projects in scholarly communications [27], has a simpler query language, and with up-coming extensions, lower complexity than other languages, such as SparQL [28].

Examples for requests could be relatively straightforward as: the number of publications or the percentage of OA documents per repository/country/continent/globally. More complex issues could be the increase in the number of research data objects per country per year or the distribution of languages of publications of type book per country in percentages.

All these requests could be extended in combination with timeline support to show evolutionary aspects. To be able to support complex requests of this kind, particular attention must be paid to the design of an appropriate query language.

Since a large amount of relevant data is available and even essential components are already existing, there is a particular demand to make this resource open and reusable. BASE currently provides a search API and an OAI-PMH interface for selected partners. It is an obvious idea to extend the portfolio with an interface delivering information for monitoring purposes.

The basic entity for the monitoring efforts is the repository. The data can be also aggregated on different levels for analytical reasons, especially for

- Institutional/Organization (aggregating the repositories belonging to an institution);
- Country;
- Region (aggregating specific regional networks);
- Continent;
- Global.

Search aspects such as the repository platform or the repository type (institutional repositories, journal platforms, research data, digital collection, and others) should be available to refine the result.

The future plans for the API of the monitoring system include the implementation of GraphQL. The GraphQL specification describes an API query language and a way for fulfilling those requests and can serve as a concept for a modern and open architecture approach.

## 4. Conclusions

The described analytical approaches based on current data show that valuable opportunities for data extraction and evaluation are available with the already implemented components. Many and widely scattered data in different contexts are available for the entire life cycle of the repository network. It is now important to create a universal generic infrastructure based on the components already in place. A suitable data model and an appropriate interface (API) are also required to provide the data for further processing.

The API and the data about repositories offer a lot of potential use cases to discover a treasure such as monitoring of the integration of standards and policies in repositories (such as ORCID, Open Access mandates), bibliometric analysis (such as the Open Access share of publications) or the generation of subject specific sections of documents for integration in other portals.

BASE is one of the key players in the repositories network, and this makes it all the more natural to return the results of analysis as feedback to the community. For the Bielefeld University Library, it seems sensible to invest internal expertise in developing a prototype as a proof-of-concept, and to implement it in our own environment first to test the interface and verify the added value of the approach. The implementation of such a prototype offers a chance to optimize the core system and its quality level. On the other hand, it could be implemented as part of a national or international project as well. For the near future BASE is planning to open a basic API version which delivers a small set of valuable information on repository and country level.

**Author Contributions:** Conceptualization, F.S., A.C. and J.S.; Data Curation, F.S.; Supervision, D.P.; Visualization, F.S.; Writing—Original Draft Preparation, F.S. and A.C.; Writing—Review & Editing: A.C., J.S. and D.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| API | Application Programming Interface |
| BASE | Bielefeld Academic Search Engine |
| CC | Creative Commons |
| COAR | Confederation of Open Access Repositories |
| D3 | Data-driven documents |
| DDC | Dewey Decimal Classification |
| DINI | German Initiative for Networked Information |
| DOI | Digital Object Identifier |
| DRIVER | Digital Repository Infrastructure Vision for European Research |
| ISO | International Organization for Standardization |
| OA | Open Access |
| OAI-PMH | Open Archives Initiative Protocol for Metadata Harvesting |
| OECD | Organisation for Economic Co-operation and Development |
| OpenAIRE | Open Access Infrastructure for Research in Europe |
| OpenDOAR | Directory of Open Access Repositories |
| ORCID | Open Researcher and Contributor iD |
| REST | Representational State Transfer |

## References

1. Open Archives Initiative Protocol for Metadata Harvesting. Available online: http://www.openarchives.org/pmh/ (accessed on 13 February 2020).
2. Summann, F.; Lossau, N. Search Engine Technology and Digital Libraries. *D Lib Mag.* **2004**, *10*, 10. [CrossRef]
3. Pieper, D.; Summann, F. Bielefeld Academic Search Engine (BASE) An end-user oriented institutional repository search service. *Libr. Hi Tech* **2006**, *24*, 614–619. [CrossRef]
4. Pieper, D.; Summann, F. 10 years of Bielefeld Academic Search Engine (BASE): Looking at the past and future of the world wide repository landscape from a service providers perspective. In Proceedings of the OR2015, 10th International Conference on Open Repositories, Indianapolis, IN, USA, 8–11 June 2015; Available online: https://pub.uni-bielefeld.de/download/2766308/2766316/or2015_base_unibi.pdf (accessed on 14 February 2020).
5. Lösch, M.; Waltinger, U.; Horstmann, W.; Mehler, A. Building a DDC-annotated Corpus from OAI Metadata. *J. Digit. Inf.* **2011**, *12*. Available online: https://journals.tdl.org/jodi/index.php/jodi/article/view/1765/1767 (accessed on 20 June 2020).
6. Dreyer, B.; Hagemann-Wilholt, S.; Vierkant, P.; Strecker, D.; Glagla-Dietz, S.; Summann, F.; Pampel, H.; Burger, M. Die Rolle der ORCID iD in der Wissenschaftskommunikation: Der Beitrag des ORCID-Deutschland-Konsortiums und das ORCID-DE-Projekt. *ABI Tech* **2019**, *39*, 112–121. [CrossRef]
7. Feijen, M.; Horstmann, W.; Manghi, P.; Robinson, M.; Russell, R. DRIVER: Building the Network for Accessing Digital Repositories across Europe. Ariadne 2007. Available online: http://www.ariadne.ac.uk/issue/53/feijen-et-al/ (accessed on 17 March 2020).
8. Europeana Cloud Project. Available online: https://pro.europeana.eu/project/europeana-cloud (accessed on 12 February 2020).
9. OpenAIRE, an EC Funded Project, Grant Agreement No 777541. Available online: https://www.openaire.eu (accessed on 27 January 2020).
10. Gusenbauer, M. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics* **2018**, *118*, 177–214. [CrossRef]
11. Directory of Open Access Repositories. Available online: https://v2.sherpa.ac.uk/opendoar/ (accessed on 21 January 2020).
12. Tennant, R. Bitter Harvest: Problems & Suggested Solutions for OAI-PMH Data & Service Providers. Available online: http://roytennant.com/bitter_harvest.html (accessed on 6 February 2020).
13. Tennant, R. Specifications for Metadata Processing Tools. Available online: http://roytennant.com/metadata_tools.pdf (accessed on 6 February 2020).

14. Broschinski, C. Rechtenormalisierung in BASE: Ergebnisse aus dem EuropeanaCloud-Projekt. *Kolloqu. Wissensinfastruktur* **2015**. [CrossRef]
15. Summann, F.; Shearer, K. COAR Roadmap Future Directions for Repository Interoperability. In Proceedings of the COAR Confederation of Open Access Repositories, Göttingen, Germany, 14–16 April 2015. [CrossRef]
16. Rodrigues, E.M.; Bollini, A.; Cabezas, A. Next Generation Repositories. Behaviours and Technical Recommendations of the COAR Next Generation Repositories Working Group. 2017. Available online: https://doi.org/10.5281/zenodo.1215014 (accessed on 14 February 2020).
17. 'Plan S' and 'cOAlition S'—Accelerating the Transition to Full and Immediate Open Access to Scientific Publications. Available online: https://www.coalition-s.org/ (accessed on 12 February 2020).
18. Principles and Implementation|Plan, S. Available online: https://www.coalition-s.org/addendum-to-the-coalition-s-guidance-on-the-implementation-of-plan-s/principles-and-implementation/ (accessed on 12 December 2019).
19. Charts|Google Developers. Available online: https://developers.google.com/chart (accessed on 11 February 2020).
20. Bostock, M.; Ogievetsky, V.; Heer, J. D$^3$ Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 2301–2309. [CrossRef] [PubMed]
21. Bostock, M. D3.js-Data-Driven Documents. Available online: https://d3js.org/ (accessed on 12 February 2020).
22. ISO 639 Code Tables. Available online: https://iso639-3.sil.org/code_tables/639/data (accessed on 14 February 2020).
23. Thalheim, B.; Schewe, K.D. Codesign of Web Information Systems. In *Book Correct Software in Web Applications*; Thalheim, B., Schewe, K.D., Prinz, A., Buchberger, B., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 293–332.
24. Fielding, R.T.; Taylor, R.N. Architectural Styles and the Design of Network-Based Software Architectures. Ph.D. Thesis, University of California, Irvine, CA, USA, 2000.
25. Hartig, O.; Pérez, J. Semantics and Complexity of GraphQL. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 1155–1164.
26. DataCite GraphQL API Guide. Available online: https://support.datacite.org/docs/datacite-graphql-api-guide (accessed on 8 April 2020).
27. Querying DBpedia with GraphQL. Available online: https://medium.com/@sklarman/querying-linked-data-with-graphql-959e28aa8013 (accessed on 8 April 2020).
28. SPARQL 1.1 Query Language for RDF. Available online: https://www.w3.org/TR/sparql11-query/ (accessed on 8 April 2020).