SCIENTIFIC DATA

OPEN

A data discovery index for the social sciences

Thomas Krämer[1], Claus-Peter Klas[1] & Brigitte Hausstein[1]

Received: 17 October 2017
Accepted: 9 March 2018
Published: 10 April 2018

This paper describes a novel search index for social and economic research data, one that enables users to search up-to-date references for data holdings in these disciplines. The index can be used for comparative analysis of publication of datasets in different areas of social science. The core of the index is the da|ra registration agency's database for social and economic data, which contains high-quality searchable metadata from registered data publishers. Research data's metadata records are harvested from data providers around the world and included in the index. In this paper, we describe the currently available indices on social science datasets and their shortcomings. Next, we describe the motivation behind and the purpose for the data discovery index as a dedicated and curated platform for finding social science research data and *gesisDataSearch*, its user interface. Further, we explain the harvesting, filtering and indexing procedure and give usage instructions for the dataset index. Lastly, we show that the index is currently the most comprehensive and most accessible collection of social science data descriptions available.

[1]Knowledge Technologies for the Social Sciences, GESIS Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne 50667, Germany. Correspondence and requests for materials should be addressed to T.K. (email: thomas.kraemer@gesis.org).

Introduction

In information infrastructure projects and initiatives one aspiration is to develop data sharing as a common part of scientific culture and practice. Achieving this goal is largely dependent on having internationally compatible infrastructures that facilitate sustainable data references, as well as integrated search and retrieval capabilities within research data. There is an obvious need for a comprehensive service that unifies data sources and allows for retrieving relevant and reliable search results as quickly as possible. Archives and repositories, such as ICPSR (https://www.icpsr.umich.edu), GESIS (https://www.gesis.org), UK Data (http://www.data-archive.ac.uk/ and the Dataverse Project[1], provide social science and economic data on their websites, (https://dataverse.org/). For example, the r3data.org database lists 201 social science repositories and 146 economics repositories (http://www.re3data.org/browse/by-subject)[2]. When searching for appropriate data, social scientists must use distributed services that are based on different systems and retrieval techniques.

Dedicated, discipline-specific social sciences and the economics search facilities are still missing. There are some initiatives underway to support the research community in this respect, but none with its sole focus on social sciences datasets and none with the purpose of providing advanced searches on high quality metadata by means of a curated set of harvested repositories (see Table 1). Advanced searches include the choice of search term operators (or vs. and) as well as searching at the field level, in contrast to searching for term in all fields.

To address this demand, the *gesisDataSearch* project (http://datasearch.gesis.org/start) was initiated. Its purpose is to create a central search point, enabling social scientists to look up or filter potential datasets quickly, to access dataset metadata and decide on its relevance for their work, and for citation purposes or reusing a dataset. This aim is achieved through a faceted search interface.

The point of departure, and core of the project, was the DOI Registration agency for social and economic data database of the DOI Registration agency for social and economic data (da|ra, https://www.da-ra.de)[3] that already includes searchable metadata from registered data centers, among them the considerable holdings of the German GESIS Data Archive (https://www.gesis.org/en/services/data-analysis/), and the US American Data Archive ICPSR. Together with data references of other relevant international data providers the content of this database was included in the search index after a systematic assessment. For this assessment, we harvested metadata in both standards, Dublin Core (DC, see http://dublincore.org/documents/DCes/) and Data Documentation Initiative (DDI, see http://www.DDIalliance.org/).

The assessment of the availability and quality of metadata records on datasets in different metadata standards showed that the more detailed DDI standard is not yet adopted by many social science institutions, resulting in lower numbers compared to records available in Dublin Core.

To provide a user-friendly search of a comprehensive social science research data collection, the search scope is more important than its depth. Further, DDI offers hundreds of elements, which differ and do not necessarily overlap across different DDI versions. Concerning the search interface, the choice of facets as the least common denominator of all available representations would have required an additional

Source	Focus on social sciences	Focus on datasets	Advanced search	Multiple languages	API	Datasets in social sciences	Facets
Elsevier datasearch	No	No	No	No	Yes	18,073	Data file type Data source type Data sources Date
OpenAIRE	No	No	No	Yes	No	7,850	Funder Publication Year Access mode Type Language Data provider
EUDAT	No	Yes	No	Yes	Yes	6,794	Start time End time Publication year Communities Tags Creator Discipline Language Publisher
DataCite	No	Yes	No	No	Yes	70,128	Publication year Registration year Data centers
gesis DataSearch	Yes	Yes	Yes	Yes	Yes	76,600	Date Language Subject Type Creator Contributor Metadata provider Data provider Set

Table 1. Comparison of portals offering search for datasets.

Box 1 | Glossary

Harvesting
Harvesting in the context of this paper is the process of retrieving metadata from data repositories describing the records available in these repositories.

Interoperability
Interoperability in this context is the ability of multiple systems with different structures to interact and exchange data without data loss.

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
OAI-PMH is a widely used protocol for harvesting metadata. Most popular repository software provide support for this protocol. For details see http://www.openarchives.org/OAI/openarchivesprotocol.html.

Dublin Core DC
Dublin Core denotes widespread metadata standards to describe resources with the purpose to increase findability and interoperability. In most cases, DC refers to the Dublin Core Metadata Element Set, version 1.1
It contains the following fields

- Contributor: An entity responsible for making contributions to the resource.
- Coverage: The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
- Creator: An entity primarily responsible for making the resource.
- Date: Time associated with an event in the lifecycle of the resource.
- Description: Description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource.
- Format: The file format, physical medium, or dimensions of the resource.
- Identifier: An unambiguous reference to the resource within a given context.
- Language: A language of the resource.
- Publisher: An entity responsible for making the resource available.
- Relation: A related resource.
- Rights: Information about rights held in and over the resource.
- Source: A related resource from which the described resource is derived.
- Subject: The topic of the resource.
- Title: A name given to the resource.
- Type: The nature or genre of the resource.

See http://dublincore.org/documents/dces/ for details.
Data Documentation Initiative DDI

DC is generic and its terms are broadly defined. For the domain of the social sciences, the DDI initiative has created a range of specific metadata standards for describing data produced by surveys and other methods in social and economic sciences, and that are used for the documentation, discovery and interpolation.
Please refer to https://www.ddialliance.org/ for details on the various standards.

abstraction from various DDI versions and DC to a sensible set of aggregations for the search interface. We therefore use the metadata standard Dublin Core element set version 1.1, as well as three additional fields: OAI set (http://www.openarchives.org/OAI/openarchivesprotocol.html#Set), data provider and metadata provider.

Results

Metadata assessment

We started the metadata collection via the OAI-protocol and collected both the Dublin Core and Data Documentation Initiative metadata formats in the latest versions available (Box 1). This first harvesting (January 2016) resulted in about 470,000 DC v 1.1 and about 67,500 DDI metadata records in versions ranging from 1.1 to 3.1. The harvesting routine for DC metadata took about seven days to complete one full harvesting.

Results also revealed that DDI records often contained only little, if any, more detail than DC records. This is because adopting DDI standards and produced rich metadata using DDI requires more time, effort, and expertise; additionally tool support for DDI is still in its infancy. In order to include as many datasets as possible on the index while creating an index for a faceted search interface, DC was chosen as the minimal metadata standard.

Metadata Provider	Count
dalra (Registration agency for social science and economic data)	39,903
DataCite	13,459
DANS (Data Archiving and Networked Services)	8,346
Odum Institute Dataverse Network	3,534
FSD (Finnish Social Science Data Archive)	1,082
LiDA (Data Archive for Humanities and Social Sciences)	299
SADA (South African Data Archive)	184
The World Bank	5

Table 2. Number of datasets in the index by metadata provider.

Harvesting

At the time of writing, the *gesisDataSearch* production system periodically harvests DC metadata from 120 OAI-PMH sets from 58 different data providers, which distribute their metadata through eight different metadata providers (see Table 2, Table 3 (available online only), and Table 4 (available online only)). After a first review of the DC metadata records, some sets were excluded from harvesting as they describe few or no relevant datasets.
Harvesting is executed according to the following automated schedule:

1. Initial full harvesting of all OAI-PMH sets for all metadata providers after system setup
2. Daily incremental harvesting of metadata records recently added to the sets. A time range starting from the past 48 h to the moment of harvesting covers short-term corrections of erroneously published datasets.
3. Yearly full harvesting of all OAI-PMH set for all metadata providers

This produced a total of about 295,000 metadata records that were filtered during the following steps in the processing chain (Data Citation 1).

Indexing

Dublin Core v1.1 is a universal metadata standard aiming at maximum interoperability. It can be applied in various ways to describe objects. Different providers comply differently with the implementation guidelines. Not all providers follow the recommendations and use controlled vocabularies. Others provide substructures, such as key-value pairs in simple text fields. These variations had to be addressed during the creation of the *gesisDataSearch* index, and are further explained below.

Filtering datasets

The selection of OAI-PMH sets is the first step in filtering metadata records that describe datasets in the social sciences and related fields. Many of the selected sets also contain metadata on objects other than datasets, such as documents, audio files, etc. Therefore, we applied the second level of filtering by excluding those metadata records from our index, that have at least dc:type element (http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=elements#elements-type) with a value matching terms on a curated exclusion list. The list of values excluded by default (currently 483 terms; https://bitbucket.org/cessda/cessda.pasc.indexer/src/e5941c0d9bc4ab5cec86b4bf9c7285de6cf688b8/src/main/resources/application.yml?at=master&fileviewer=file-view-default#application.yml-562) can be extended during runtime using the web-based admin interface. Combined with a re-indexing of parts of the metadata or the whole corpus, the index can be iteratively curated.

Handling multiple languages

DC v.1.1 includes a 'dc:language' element http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=elements#language) that should name the language of the resource; the described dataset in this case. Some providers, however, use the 'language' element to indicate the language of the metadata.

Further, each element might contain a 'lang' attribute, indicating the language of the value of that particular field (http://dublincore.org/schemas/xmls/simpledc20021212.xsd).

As *gesisDataSearch* should contain as much information as possible, we applied a simple procedure for handling language in our index:

- If a 'lang' attribute of any DC element indicates a language, save the element content as sub-field (e.g., title.en = x, title.fr = y)
- If no 'lang' attribute is given, store the element content as 'nn' e.g., title.nn
- Store all elements' contents in an additional 'all' field e.g., title.all

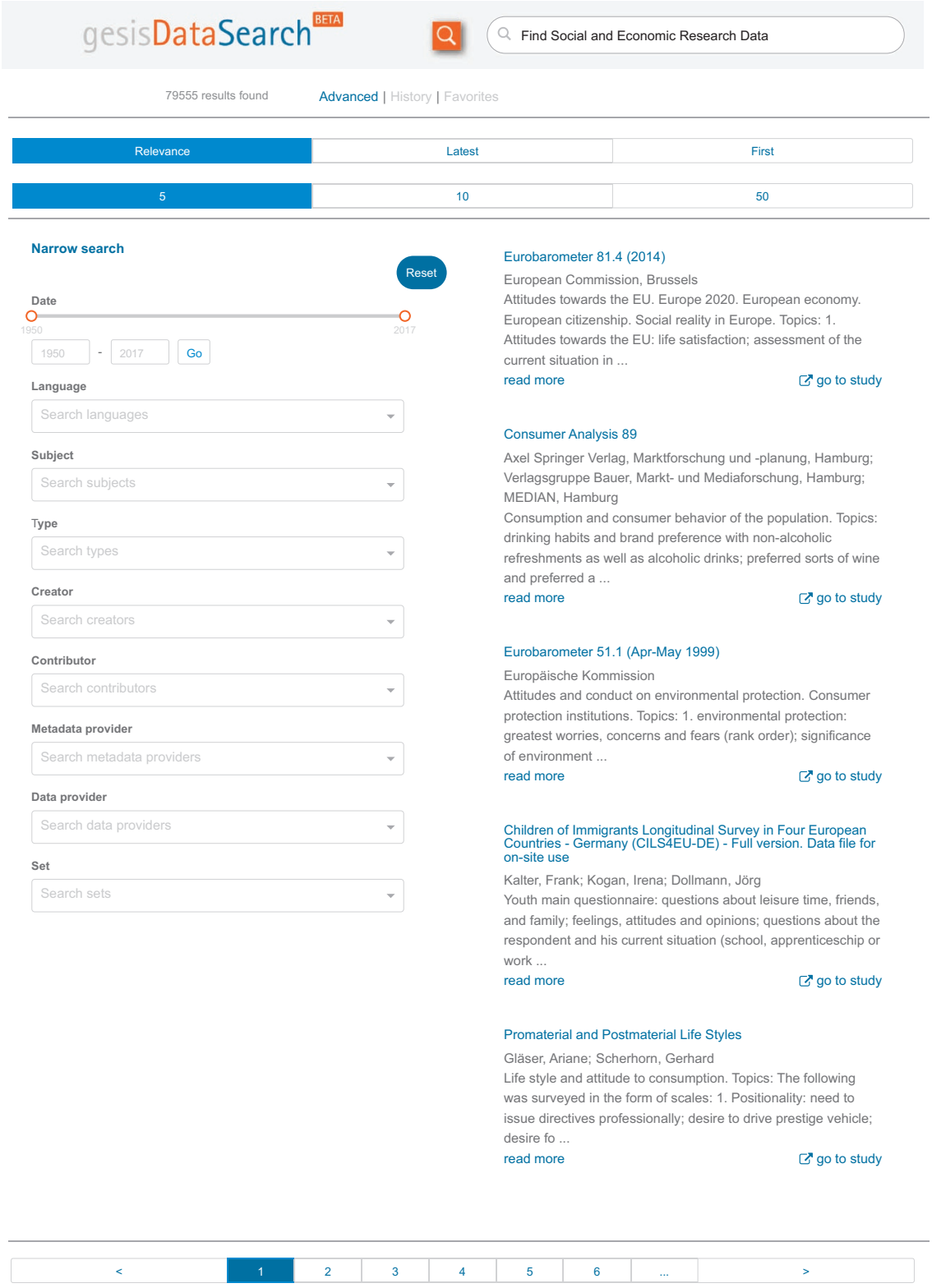

Figure 1. The gesisDataSearch user interface. The user interface to the social science data index is implemented as a faceted search portal. Free term search is available either across all fields of the index or within one field, such as subject. All facet search inputs include autocomplete functionality.

Dublin Core source element	Annotator	Elasticsearch field names	Elasticsearch field type
Title	CoreNLP	Countries	String
Coverage	CoreNLP	Countries	String
Subject	CoreNLP	Countries	String
Title	CoreNLP	AnydateYear Anydate If timespan was detected: Start End	Number Date
Coverage	CoreNLP	AnydateYear Anydate If timespan was detected: Start End	Number Date
Subject	CoreNLP	AnydateYear Anydate If timespan was detected: Start End	Number Date
Date	CoreNLP	AnydateYear Anydate If timespan was detected: Start End	Number Date
Title	CoreNLP & Photon	Points	geo_point
Coverage	CoreNLP & Photon	Points	geo_point
Subject	CoreNLP & Photon	Points	geo_point

Table 5. Experimental enrichment of metadata using named entity recognition and geocoding.

This makes it possible to let the faceted search interface users choose their preferred language and while still showing metadata content if it is not present in the desired language. The 'all' field is used for a per field search (Fig. 1).

Metadata enrichment

The DC elements 'dc:coverage' and 'dc:subject' have a high topical overlap (http://dublincore.org/usage/decisions/2012/dcterms-changes/); for instance, subject elements often contain location names such as countries. The usage of the 'coverage' element – intended to denote spatial and temporal applicability – is very diverse and ranges from standardized dates with milliseconds granularity to relative-time indications such as 'Early Middle Ages', and can contain both instants and time ranges.

We addressed this semantic problem by introducing a set of experimental, non-validated fields whose content is the result of a named entity recognition and geocoding. For named entity recognition, the Stanford CoreNLP library v3.6 was used[4]. Entities that are recognized as locations are forwarded to a geocoding service based on photon (https://github.com/komoot/photon), which uses OpenStreetMap data to provide coordinates to location names (Table 5). The current index contains 76,600 descriptions of datasets for the social sciences and related fields (Fig. 2).

Managing the processing chain

Processing involves a number of services, some of which were developed by GESIS. The harvester is responsible for fetching DC metadata records from various metadata providers via their OAI-PMH endpoints. The DC metadata is stored as XML files in a folder. The indexer application runs on the same machine and processes all files in the metadata folder. The CoreNLP entity recognition is embedded into the indexer application. The indexer further uses the photon geocoding service to retrieve geo coordinates from place names detected by the CoreNLP entity recognition and associates these geo coordinates to indexed metadata record. The Worldbank application integrates both, fetching data from the Worldbank Data Catalog API (http://api.worldbank.org/v2/datacatalog) and indexing into the search index following the same document model that is used by the indexer (Fig. 3).

As the index is continuously growing and being curated, we created a possibility to intervene in the procession chain when need, e.g., to get basic statistics, re-index or re-harvest particular OAI-PMH sets, to add or remove selected metadata from the index, or to change the execution schedule (Fig. 4).

We developed a remote control (https://github.com/codecentric/spring-boot-admin) for the spring boot based microservices (Fig. 4) which allows us to:

- review log files and keep track of what is currently being harvested or indexed
- change configuration during runtime, e.g., add new metadata provider or change data provider labels
- get e-mails in case of problems

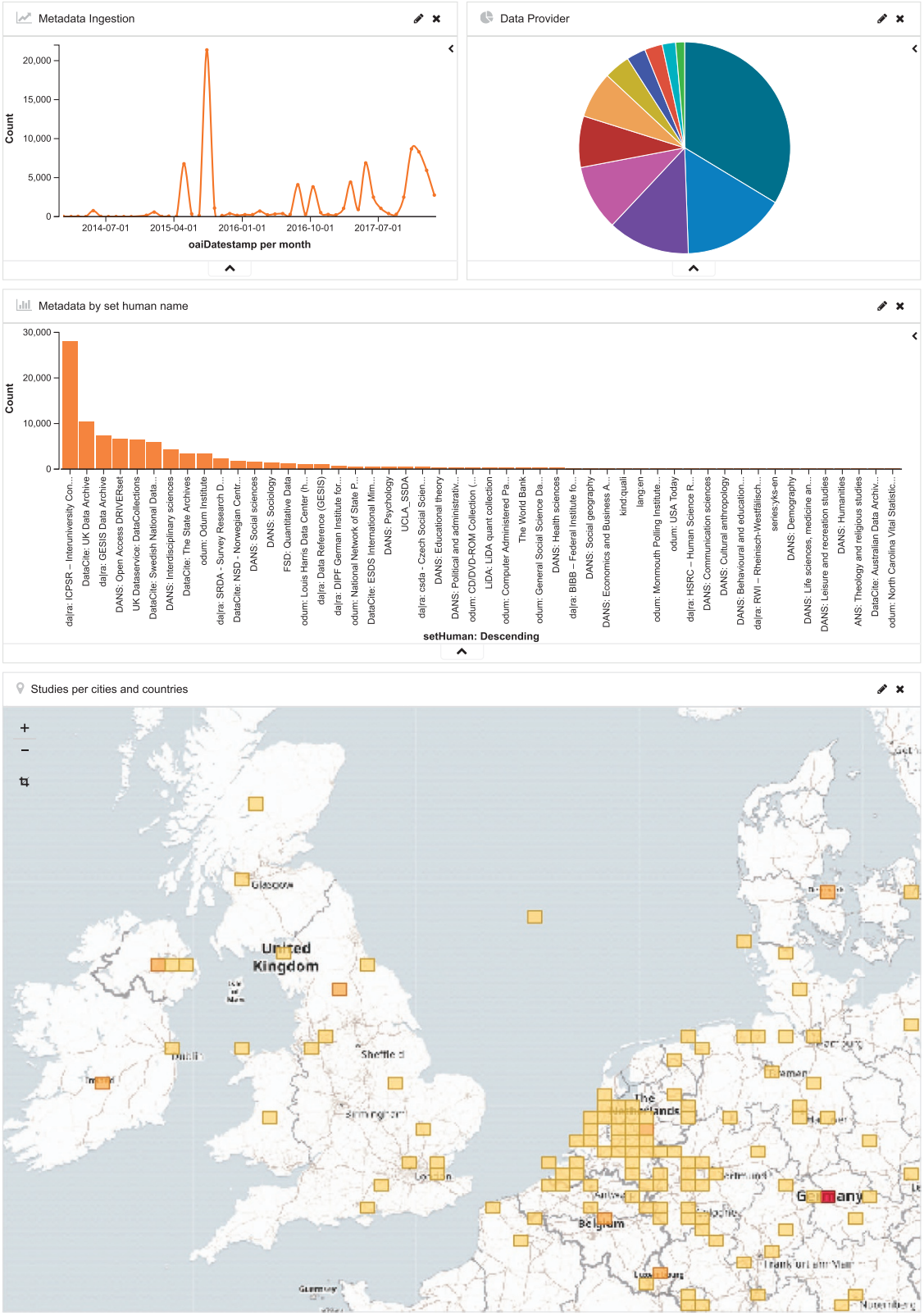

Figure 2. Visualisation of the contents of metadata. The visualization of the indexed social science metadata was helpful particularly in the early stages, when different metadata standards and versions were compared with respect to their application and use of fields across different metadata providers.

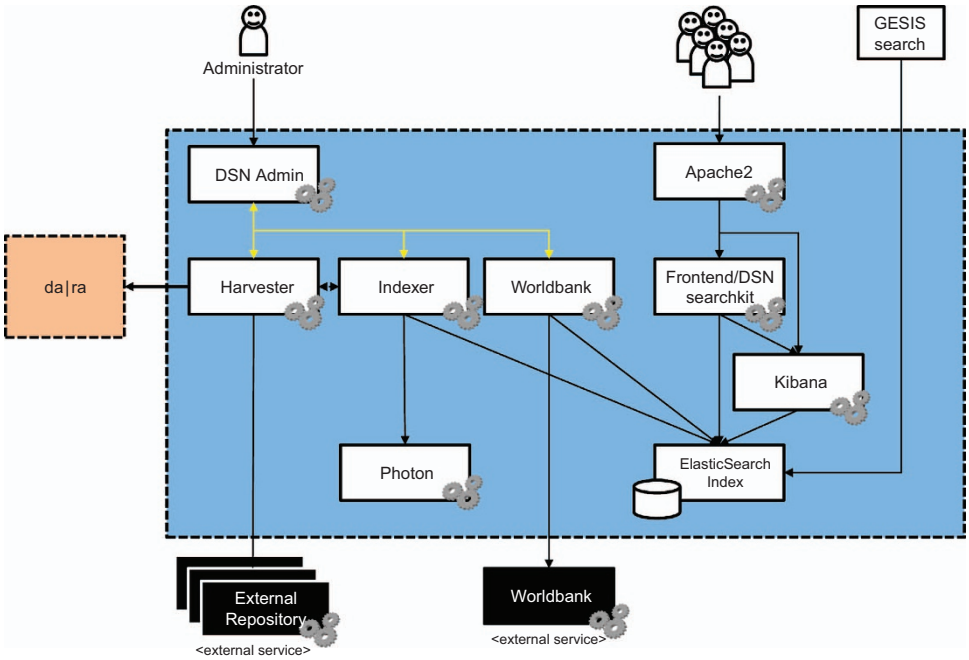

Figure 3. The gesisDataSearch architecture. The services required to create the index and provide the faceted search interface are either internal and developed or operated by GESIS (within the dotted line) or external (OAI endpoints or World Bank Data Catalog). While it is possible to run all required internal components on one physical machine, the architecture supports well scaling and operation on a cloud infrastructure. CESSDA operates the stack on Google Cloud Platform.

- review uptime of components
- execute dedicated functions, e.g., add new OAI-PMH repositories for harvesting or re-index all or selected metadata records during runtime

The index is created with elasticsearch. The open source technology stack elasticsearch v2.4.4 (https://www.elastic.co/downloads/past-releases/elasticsearch-2-4-4), based on Apache Lucene, was chosen as search engine framework, for being scalable and for its good tool support, e.g., with the spring-data-elasticsearch libraries (https://projects.spring.io/spring-data-elasticsearch/). The user interface datasearch.gesis.org (http://datasearch.gesis.org/start; Fig. 1) is based on searchkit (https://github.com/searchkit/searchkit), a collection of user interface components built using the react library (https://facebook.github.io/react/).

Usage Notes

The elasticsearch index 'DC' is available as elasticsearch snapshot, created with elasticsearch v 2.4.4 (Data Citation 1). It can be easily restored into an existing elasticsearch instance using the restore snapshot feature (https://www.elastic.co/guide/en/elasticsearch/reference/2.4/modules-snapshots.html).

Discussion

As of the time of writing (1/2018), the *gesisDataSearch* production system harvests DC metadata from 120 OAI-PMH sets from 58 different data providers, which distribute their metadata through eight different metadata providers. This results in about 295,000 metadata records. After filtering, the *gesisDataSearch* index provides 76,600 descriptions of datasets for the social sciences and related fields. This index is a comprehensive service that allows for obtaining relevant and reliable search results in one place.

Table 1 compares the *gesisDataSearch* index with alternative approaches to searching datasets in the social sciences. This shows that the presented index, accessible through datasearch.gesis.org, is currently the most comprehensive dataset focussing on the social sciences, with the most advanced search and filter possibilities combined with the possibility to review metadata on research data in different languages.

We tried to improve *gesisDataSearch* through several measures: First, by expanding the number of relevant data providers included in the harvesting process; second, enriching the DC metadata with elements from the DDI standard family; and finally, improving NLP accuracy for place names and date and time detection[5].

Figure 4. Web-based administration tool for spring boot based microservices. The web-based administration tools helps to manage and adapt the long running and repetitive processes. Users can see the resource consumption (network, storage, RAM, CPU), watch the log output to control operation, adapt log level, see and change configuration of each service at runtime, e.g., to adapt the harvesting schedule, to add new OAI endpoints to the harvesting process or to add or remove terms from filter lists.

Methods

To identify relevant metadata providers the available data resources were analysed. Our starting point was information on data archives in archival networks such as CESSDA European Research Infrastructure Consortium (ERIC) (https://www.cessda.eu). Furthermore, we consulted the inventory of data repositories 'r3data.org' as well as the metadata portals of DataCite, EUDAT, OpenAIRE, and Dataverse (https://dataverse.org/). As an outcome, 50 repositories were evaluated concerning technological premises, availability of metadata, used metadata formats and documentation level. Among them were archives and research centers, research projects, infrastructure projects, and service providers.

After a manual review of the wide variety of different technical systems for metadata publication a systematic assessment of the provided metadata was made on four levels:

- Metadata formats
- Quality of metadata
- Cross-disciplinary metadata
- Multilingualism of metadata

Further, the 'terms of use' of metadata also varied and needed to be taken into account when creating the index. We only included metadata that is publicly available. However, the referenced datasets themselves might be subject to other terms of use.

We decided to use the open archives initiative protocol for metadata harvesting (OAI-PM) for the retrieval of metadata from data providers. We dismissed alternatives to OAI-PMH, such as web scraping or API, which required both further clarifications of terms of use of scraped data and more human resources for implementing and operating many different APIs. One exception was the Worldbank Data Catalog that provides a basic API to its contents (http://api.worldbank.org/v2/datacatalog). It was included in the index because of the relevance of its content.

Source Code availability

The source code of the core components is publicly available at Bitbucket (https://bitbucket.org/cessda) as part of the CESSDA ERIC:

- admin https://bitbucket.org/cessda/cessda.pasc.admin
- harvester https://bitbucket.org/cessda/cessda.pasc.harvester
- indexer https://bitbucket.org/cessda/cessda.pasc.indexer
- worldbank https://bitbucket.org/cessda/cessda.pasc.worldbank
- searchkit https://bitbucket.org/cessda/cessda.pasc.searchkit

References

1. Crosas, M. The Dataverse Network: An Open-Source Application for Sharing, Discovering and Preserving Data. *D-Lib Magazine* **17, 1/2**. https://doi.org/10.1045/january2011-crosas (2011).
2. Kindling, M. The Landscape of Research Data Repositories in 2015: A re3data Analysis. *D-Lib Magazine* **23, 3/4**. https://doi.org/10.1045/march2017-kindling (2017).
3. Helbig, K., Hausstein, B. & Toepfer, R. Supporting Data Citation: Experiences and Best Practices of a DOI Allocation Agency for Social Sciences. *Journal of Librarianship and Scholarly Communication* **3,** eP1220. https://doi.org/10.7710/2162-3309.1220 (2015).
4. Manning, C. *et al.* The Stanford CoreNLP natural language processing toolkit. In: *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations.* 55-60 (2014).
5. Hausstein, B. *et al.* Knocking on data repositories' doors - How to build an integrated search index for social and economic data. Digital Infrastructures for Research 2016 https://www.digitalinfrastructures.eu/content/knocking-data-repositories%E2%80%99-doors-how-build-integrated-search-index-social-and-economic-data (Krakow, 28-30 September, 2016).

Data Citations

1. Krämer, T. *Zenodo* http://doi.org/10.5281/zenodo.896431 (2017).

Acknowledgements

The project was funded by the German Research Foundation (DFG) under the grant number SU 647/132. It runs from 2014–2018.

Author Contributions

T.K. and B.H. drafted and wrote the paper. T.K. implemented the code, with some contributions from C.-P.K. C.-P.K. is responsible for the technical project lead and participated in discussions on how to set up the harvester and indexer.

Additional Information

Tables 3 and 4 are available only in the online version of this paper.

Competing interests: The authors declare no competing interests.

How to cite this article: Krämer, T. *et al.* A data discovery index for the social sciences. *Sci. Data* 5:180064 doi: 10.1038/sdata.2018.64 (2018).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2018