






A LOD Backend Infrastructure for Scientific Search Portals

Benjamin Zapilko^(✉), Katarina Boland, and Dagmar Kern

GESIS - Leibniz-Institute for the Social Sciences, Cologne, Germany
{benjamin.zapilko,katarina.boland,dagmar.kern}@gesis.org

Abstract. In recent years, Linked Data became a key technology for organizations in order to publish their data collections on the web and to connect it with other data sources on the web. With the ongoing change in the research infrastructure landscape where an integrated search for comprehensive research information gains importance, organizations are challenged to connect their historically unconnected databases with each other. In this article, we present a Linked Open Data based backend infrastructure for a scientific search portal which is set as an additional layer between unconnected non-RDF data collections and makes the links between datasets visible and usable for retrieval. In addition, Linked Data technologies are used in order to organize different versions and aggregations of datasets. We evaluate the in-use application of our approach in a scientific search portal for the social sciences by investigating the benefit of links between different data sources in a user study.

1 Introduction

The landscape of research infrastructures like libraries, archives and research data centers is undergoing significant changes [5,9,13], which is also reflected in the research agendas of international and national funding agencies. Were data collections and databases providing scientific information and research data originally unconnected due to historically grown organizational structures, there is now a demand for an integrated and connected provision of this information which is also justified through the ongoing Open Science discussion. In an online survey with 337 social science researchers in Germany, we found evidence that researchers are interested in links between information of different types and from different sources. As a result, not only data collections should be connected with each other, but there is also a need for integrated search functionalities.

However, in current scientific portals these user needs are often not yet reflected. For example, publications and research data are typically held in separate data collections, represented in different metadata schemas and different data formats. They have to be accessed differently, e.g. via different search portals. Even when pushing these data collections into one single database with a search functionality on top, the challenge of connecting datasets of different collections with each other remains. This problem does not only involve the identification of links, but also the treatment of vague links, i.e. if the identifier of a

linked dataset is unknown and several similar datasets as candidates exist. It is also necessary to keep track of provenance information, i.e. where the datasets come from and how the link has been created. Additionally, there is a problem of disambiguation, since different data collections may contain duplicates. Finally, it has to be decided whether new infrastructures should be built or whether historically grown infrastructures can be reused and extended.

For publishing and connecting data on the Web, Linked Open Data (LOD) [8] has become a popular method in recent years [17, 20]. Numerous institutions have started with efforts into that direction like several libraries (e.g. German National Library¹, the French National Library [19], Library of Congress², Europeana [10]) but also archives and museums³, or organizations which hold statistical data like Eurostat⁴, World Bank⁵, and smaller data providers. Especially in the library sector, Linked Open Data has become a popular technique for publishing bibliographic metadata on the Web and connect it to other Linked Datasets [21] like authority data and persistent identifiers like VIAF⁶.

In this article, we present a Linked Open Data based backend infrastructure for scientific search portals which is set as an additional layer between unconnected data collections and makes the links between datasets visible and searchable. Historically built infrastructures are kept running. In this approach, Linked Data serves as backbone for connecting datasets of different data collections. Metadata of the original non-RDF data collections (including information about links to other datasets) is imported into a link database where connections between datasets are identified. The links between datasets are stored as a graph and made available in an Elasticsearch index for an efficient integration into search portals. In order to address occurring heterogeneity with vague links between datasets, a research data ontology is used for representing different versions and aggregations of research datasets. In contrast to existing approaches, our approach covers the full workflow from heterogeneous non-RDF data collections up to the provision in an efficient search index with the integrated and interlinked data. We evaluated the implementation of this approach in a real world scenario, a scientific search portal by conducting a user study where the benefit of links between different datasets is investigated.

The rest of the paper is structured as follows. In Sect. 2, we give an overview of the use case. We present concept and implementation of the LOD backend infrastructure in Sect. 3 and present an evaluation through a user study in Sect. 4. In Sect. 5, we provide an overview of related work and similar approaches. Finally, in Sect. 6, we conclude and give an outlook on future work.

¹ <http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkeddata.html>.

² <http://id.loc.gov/>.

³ <http://americanart.si.edu/collections/search/iod/about/>.

⁴ <http://ec.europa.eu/eurostat/de>.

⁵ <http://www.worldbank.org/>.

⁶ <https://viaf.org/>.

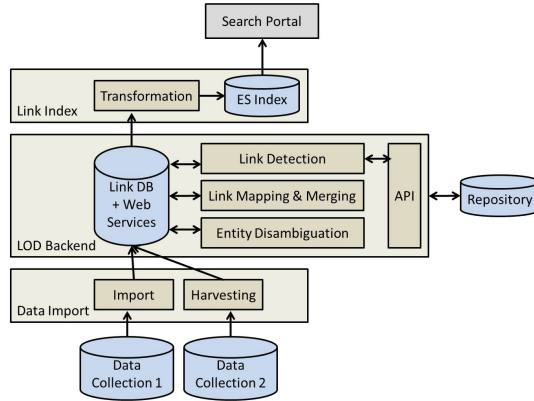


Fig. 1. Architecture of the LOD backend infrastructure

2 Use Case

The use case in this paper is centered on a research infrastructure organization, GESIS - Leibniz-Institute for the Social Sciences. GESIS offers a wide range of information and data, e.g. publications, research data, projects, and others, in various specialized portals. These portals are characterized by a high degree of heterogeneity in terms of architecture, data management, technical implementation, and the data itself. The data in these portals are poorly interlinked.

In a survey, we collected information needs of 337 social science researchers in Germany with an online questionnaire. We found that researchers are interested in links between information of different types and from different sources. For example, about 40% of the participants stated that “I’m looking for research data mentioned in a paper” is one of their own information needs. Therefore, the individual information objects in these data collections need to be integrated and interlinked.

3 Linked Open Data Backend Infrastructure

In this section, we present our approach for a LOD backend infrastructure. We describe the architecture and its components in detail and explain how the datasets of the original data collections are imported, linked, enriched and provided for the search portals via an Elasticsearch index.

3.1 Architecture

The architecture of our approach is set up additionally to existing infrastructures, i.e. original databases and portals as well as attached workflows remain the same. In Fig. 1, an overview of the architecture is shown.

Depending on the accessibility of the data (e.g. through a search index or as a dump file), the data is either imported directly or harvested from the different data sources and updated in an automated procedure (see Sect. 3.2). At import, several steps of data enrichment are performed on the data like mapping of IDs, entity disambiguation and link merging (see Sects. 3.4 to 3.6). Additional links are generated automatically (see Sect. 3.4). The enriched data is stored in the link database (see Sect. 3.7) together with detailed provenance information about the creation process (e.g. times and parameters of all executed algorithms creating or modifying the data). Finally, the data is transformed and pushed into an Elasticsearch index, which allows for efficient querying of the data (see Sect. 3.8). Since our use case does not require complex queries, we decided to use Elasticsearch. However, when more complex queries are desired we provide RDF through an API which can be queried using SPARQL. The above described processes are executed offline. Once the data is pushed to the index, no other processes are slowing down the search system's performance. Currently, the index holds 108435 documents with 277678 links between them.

Parts of this infrastructure have been developed in the DFG-funded projects InFoLiS I and II⁷ and have been extended for our purposes. The overall architecture is generalized, i.e. the infrastructure can be reused for different use cases. Only the import and possibly required mappings of IDs have to be adjusted when including different data sources. Some components are using GESIS portals to lookup metadata or IDs⁸. However, for a general applicability of the infrastructure these portals can be replaced with others depending on the data sources and domains (see Repository in Fig. 1).

3.2 Data Sources

The infrastructure uses easily extendible JAVA routines for harvesting and import to allow processing of different data formats. We imported a variety of sources relevant for our use case. These include publications, research data, research projects, institutions and scales (survey instruments) as interlinked entities. The data is partly provided by the scientific community, partly created by GESIS staff; it is provided in different data formats either via dumps or via a Solr interface. Figure 2 gives a more structured overview of this heterogeneous input data. All data is transformed into the InFoLiS link format on import (this format is described in more detail in Sect. 3.3). Additionally, we use the embedded InFoLiS web service framework for automatically detecting links to research data in full texts of scientific publications.

⁷ www.infolis.gesis.org.

⁸ <https://dbk.gesis.org/dbksearch/>, <http://zis.gesis.org/>, <https://www.da-ra.de/>, <http://datasearch.gesis.org/>, <http://sowiport.gesis.org/> and <http://www.ssoar.info/>.

Name	Description	Format
GESIS Bibliographies	Bibliographies for research data	BibTeX
ZIS	Bibliographies for scales	BibTeX
SOFISWiki	projects, publications, data, institutions	custom (Solr index)
GESIS Data Catalog	Research data to literature links	custom (Solr index)
GESIS Library	Research data to literature links	custom (Solr index)
automatically created links	Research data to literature links	native

Fig. 2. Overview of data sources

3.3 Data Format

Figure 3 illustrates how entities and links are stored in the database using the respective InFoLiS format. As an entity may represent a publication, dataset, research project, institution, scale or data reference (here: citedData), the format includes a wealth of different bibliographic metadata fields such as collection titles, editors or versioning information. For better comprehensiveness, the list of bibliographic metadata is abbreviated here⁹.

For entity links, the fields fromEntity and toEntity of a link specify the URIs of the origin and target entity of the link relation. The field linkReason gives the URI to a TextualReference, an entity containing a text snippet taken from the fromEntity containing the reference to the toEntity, i.e. it constitutes the reason why a link was established between the two entities. For automatically created links, this is the text snippet extracted using an extraction pattern. Some manually created links also feature a text snippet explaining the relationship between the linked entities. The linkView shows the reference to the toEntity. For automatically created links, this is the reference extracted from the text snippet in the linkReason field. For manually created links, this is the name of the linked entity, if given in the source data. The field entityRelations specifies the relation of the reference in linkView to the toEntity. When the fromEntity is of type publication and the toEntity of type citedData, the entityRelation typically is “references”. When the fromEntity is of type citedData and the toEntity of type dataset, the entityRelations specifies the match of granularity of citation and linked dataset, i.e. whether the linked dataset holds exactly the cited data or only a subset or a superset of it.

It is important to note the structure of the links: given a publication which references a dataset and a dataset being described by that reference, these relations are represented using two links: one link from a publication entity to a citedData entity plus one link from the citedData entity to a dataset entity. This way of modelling relations has the advantage that matchings from references to actual datasets can be updated easily, e.g. when new datasets are entered in the

⁹ For a full overview of the format, see <https://github.com/infolis/infolis-web/blob/master/data/infolis.tson>.

entityType	entityView	entityProvenance	name	year	authors	...	entityReliability
*Type of the entity	Citation string	Source of the entity	bibliographic metadata: Title	bibliographic metadata: Year	bibliographic metadata: Authors	further bibliographic metadata	Reliability score: 1 for manually created entities, <1 for automatically generated data
dataset	Schupp, Jürgen; Goebel, Jan; Kroh, Martin et al. (2017): Sozio-oekonomisches Panel (SOEP), Daten der Jahre 1984-2015 (internationale Version) 32i.1. Dataset. http://doi.org/10.5684/soep.v32i.1	datasearch	Sozio-oekonomisches Panel (SOEP), Daten der Jahre 1984-2015 (internationale Version)	2017	Schupp, Jürgen; Goebel, Jan; Kroh, Martin; Schröder, Carsten; Bartels, Charlotte; Erhardt, Klaudia; Fedorets, Alexandra; (...)	...	1
citedData	Mikrozensus 1982	InfoLink	Mikrozensus	1982	(empty)	(empty)	0.3

* information type: publication | dataset | project | institution | instrument | citedData

EntityLink

fromEntity	toEntity	entityRelations	provenance	linkView	linkReason	confidence
URI of entity1	URI of entity2	+*Relation of entity2 to the cited entity	Source of the link	*Reference of entity2 used for creating the link	**URI of the Textual Reference entity representing a text snippet containing the reference to entity2	Reliability score: 1 for manually created links, <1 for automatically generated data
http://example.foo/en/ty/150a5a30-63211a7-bd8a-3010	http://example.foo/en/ty/037a2ae8e11611e7-bd8a-87c3	part_of_temporal	InfoLink	SOEP 1995	http://example.foo/textualReference/0132e2a1-e573-12d7-ba8a-88a0	0.53
http://example.foo/en/ty/150a5301-377600d9-ad0b-5389	http://example.foo/en/ty/57321a79-990f11a7-9c5b-d5d4	references	InfoLink	Mikrozensus 1982	http://example.foo/textualReference/b1ae4260-98cb-11e7-91ad-2742	0.3
http://example.foo/en/ty/29e0ba01-676910c9-da2a-3409	http://example.foo/en/ty/0735a790-a95703d6-2d3a-439e	(empty)	DBK	USIA, Washington (1960): Internationale Beziehungen (Februar 1960)	(empty)	1

+ used values: references | part_of_temporal | superset_of_temporal | same-as_temporal
 * available for automatically generated links
 ** available for all automatically and some of the manually generated links

Fig. 3. Format of entities and links in the link database

repository used for matching. However, this model makes querying links from publications to datasets more costly as more queries are required.

3.4 Link Detection

We employ the following mechanisms for link detection: 1. extraction and lookup of DOIs; 2. pattern-based reference extraction and linking; 3. term-based reference extraction and linking. All of them are implemented as extensions to the InFoLiS framework and are thus executable as web services.

Extraction and lookup of DOIs. DOIs are extracted from full texts and looked up in a research data repository (da|ra in our case) to retrieve further metadata. Any DOI not found in the repository is ignored as DOIs can be used to identify other entities such as publications that are not of interest here.

Pattern-based reference extraction and linking. Many research data citations to date lack persistent identifiers. Instead, often a more colloquial way of referencing research data is followed [1]. To identify these research data references, we use a method based on the semi-automatic generation of extraction patterns which is included in the InFoLiS framework as an enhanced version of the algorithm described in [1].

Term-based reference extraction and linking. The aforementioned algorithm yields a list of research data citations. From this list, we compiled a set

of names used to refer to research datasets. We curated this list to remove any false entries and enhance them with additional relevant dataset names. This list cannot directly be extracted by crawling a research data repository such as da|ra because the exact titles given in the metadata are rarely cited in publications. Authors prefer a more colloquial representation of the titles in their citations [1]. Each term in this list is searched in all available full texts of publications. The extracted references are then linked to research data records by mapping them to records in a research data repository. While this leads to duplicate links when used in addition to the pattern-based approach, this procedure generates additional links and increases the recall of the overall approach.

We applied the introduced link detection on all documents in the Open Access repository SSOAR and saved all resulting links directly in our database. Furthermore, we developed an automated workflow to apply automatic link detection on any new documents uploaded to SSOAR as an extension to the InFoLiS infrastructure. Any links created by this procedure are incorporated into the link database automatically. Finally, the link detection mechanism can be activated at any time to process any document and store the resulting links in the link database by using the respective web service.

Full texts of the manually linked publications are often not publicly available. Using SSOAR documents, we focus on a distinct set of publications for automatic link creation. Thus, the overlap between the manually and automatically created links is low which means that we gain valuable new information by our automated methods but at the same time cannot easily evaluate the full system's precision and recall. We are currently implementing a manual review phase for all links generated at document upload in SSOAR.

3.5 Entity Disambiguation and Link Merging

The links imported from the different sources come with varying metadata and varying degrees of granularity and exactness. Also, the input data may contain duplicates, both within and across sources and both on entity level and on link level: multiple data sources may contain the same entity, e.g. publication, with equal or diverging metadata and with equal or diverging links.

ID matching. In the ideal case, the linked entities in the source data are represented by IDs. However, different sources use different types of identifiers, e.g. DOIs, URNs, URLs or handles. In order to link information items inside a repository, the identifiers of the source dataset have to be matched with the identifiers used in that repository. Matching of IDs is also required for entity disambiguation. Thus, we perform a lookup in repositories for research data and for publications respectively. When a matching record is found, an entity is created in the link database representing this record/the linked entity. This entity is reused when a matching reference occurs again in the same or in another data source, i.e. at this step, publication and dataset entities are disambiguated.

Disambiguation. Some links in the source data, however, are only vague: they contain a reference to an object (e.g. a citation string or the referenced objects

metadata) rather than to an ID. These references are mapped to an ID by querying repositories using the available metadata, a publication repository if the cited entity is a publication or a research data repository if the entity is a dataset. References that cannot be mapped to any record are assigned a URI as their distinct ID. In order to disambiguate such entities, we use normalized versions of titles, years and author names. Entities lacking any of these fields are ignored to minimize matching errors.

Merging. When a duplicate entity is found, the links of all copies are merged. For this, the link database is searched for all ingoing and outgoing links of the entity already present in the database. For every entity, there must be at least one link or else the entity would not have been included in database in the first place. Thus, for every found link, the following cases may occur:

1. the new entity shares the same link, i.e. it is linked to the same entity
2. the new entity does not have this link
3. the new entity is linked to the same entity but on a lower level of granularity
4. the new entity is linked to the same entity but on a higher level of granularity

Likewise, the new entity may contain new links hitherto not present in the database.

At this point, all entities are disambiguated including sources and targets of known and new links. With this information, we can determine whether links are equal, i.e. whether they link the same entities. For our use case, the direction of a link does not hold any meaning, all links are being treated as bidirectional because in information systems, it is usually desired to show links between resources for the title views of each resource involved, not just for one of them. Hence, links are considered equal when the linked entities are equal regardless of which of these entities is the source or the target of the link.

1. When the link already exists, provenance information of the new link is added to the provenance information of the known link. The confidence value is updated to the higher confidence value of both links. In case of conflicting metadata, the metadata with the higher confidence is kept. The new link is discarded and only the updated link remains in the database.
2. When the new entity does not have the current link, no action is required.
3. When the new link is the more coarse-grained version of a known link, the known link is kept and the new link is discarded.
4. When the new link is the more fine-grained version of a known link, the new link is added to the database with the disambiguated entity as source/target entity. The known link is deleted.
5. When the link is not yet known, it is added to the database with the disambiguated entity as source/target entity.

Whether a link is a more coarse- or fine-grained version of another is determined using the research data ontology described in Sect. 3.6. Lookup of the entities in the ontology is performed automatically while the ontology has been created manually.

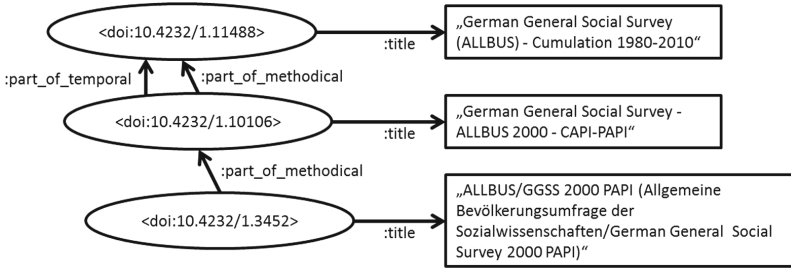


Fig. 4. Example of the research data ontology with different granularities

3.6 Research Data Ontology

There is often a mismatch between the granularity in which data is cited and the granularity in which data is registered in data repositories [14]. In order to address this problem, we developed a research data ontology that captures the relationships between many of the datasets relevant to our use case. The ontology models hierarchical relationships between research datasets, i.e. a dataset may consist of different data collections (e.g. taken place in different years) which may also include different versions of the dataset (e.g. different samples or with errata). In Fig. 4, an example of the research data ontology is shown which illustrates different granularities of research datasets. For comparing the granularity of links, the linked research data entities are compared. If the first link points to a higher level in the research data ontology, e.g. a cumulated data file, while the second points to a lower level of the same dataset, e.g. a specific subset of the cumulated data file, the first link is seen as being more coarse-grained than the second. Beside using the ontology for merging, it is also accessible in the link database so that for every research data entity in the database, its relations to other datasets can be retrieved.

3.7 Link Database

As described in Sect. 3.1, we reuse the InFoLiS infrastructure consisting of a Node.js based API backend which provides RESTful web services and an LOD representation of the data. Since we work with heterogeneous document-like data, we use a MongoDB for storage. Storing the data in graphs allows for easy representation of the links between items and storing additional, necessary information like provenance information. This is highly important in order to comprehend how and on which basis links have been generated. The infrastructure has been modularized and deployed in Docker containers which allows an execution with moderate resources.

3.8 Elasticsearch Index

For a fast and efficient search, the link database is pushed into an Elasticsearch index. While the link database features indirect links for facilitating updates

regarding reference matching (see Sect. 3.3), indirect links are not desired in the index queried by the information portals.



Fig. 5. Links before and after transformation

Thus, we implemented an algorithm to flatten links when necessary and push all links to the search index in a simplified format. Figure 5 illustrates a set of links before and after transformation. In this example, a publication holds two dataset references which are matched to one or more datasets. After flattening (right side of the figure), the publication is now linked to the datasets directly. Note that an entity can still be linked to a citedData entity if the latter is not linked to a dataset. Information on the data citations is not lost but instead added to the links’ metadata. The algorithms for transforming the links and pushing them to the index are implemented as an enhancement to the InFoLiS framework and can be invoked via the web service.

fromID	toID	fromType	fromView	toType	toView	linkReason
ID of entity1	ID of entity2	Information type* of entity1	Citation string of entity1	Information type*+ of entity2	Citation string of entity2	**Text snippet containing the reference to entity2
gegis-ss-ar-6762	datasearch-ht-pwww-da-ra-deoalp--oaoa-i-da-ra-de557591	publication	Erwerbsarbeit und Erwerbsbevölkerung im Wandel: Anpassungsprobleme einer alternden Gesellschaft. Frankfurt am Main, Campus Verl., 1998, 281 S., (Veröffentlichung aus dem Verbund Arbeits- und Innovationspotentiale im Wandel)	dataset	Schupp, Jürgen; Goebel, Jan; Kroh, Martin et. al. (2017): Sozio-oekonomisches Panel (SOEP), Daten der Jahre 1984-2015 (internationale Version) Version: 32i.1. Dataset. http://doi.org/10.5684/soep.v32i.1	Datenbasis ist das Sozio-oekonomische Panel (SOEP) (Projektgruppe Panel 1995).
gegis-ss-ar-20988	literaturpool-57321a70-990f-11e7-9c5b-d59dcbf11d82	publication	Hartmann, Peter H. (1990): Wie repräsentativ sind Bevölkerungsumfragen? Ein Vergleich des ALLBUS und des Mikrozensus	citedData	Mikrozensus 1982	Verwendet wurden Daten des Mikrozensus 1982
ZA2125	wzb-bib-b000028159	dataset	USIA, Washington (1960): Internationale Beziehungen (Februar 1960)	publication	Merritt, Richard L.; Puchala, Donald J. (Hrsg.): Western European Perspectives on International Affairs: Public Opinion Studies and Evaluations. New York: Praeger 1968.	(empty)

+ information type: publication | dataset | project | institution | instrument | citedData
 * available for automatically generated links
 ** available for all automatically and some of the manually generated links

Fig. 6. Additional or different fields of links in the Elasticsearch index

In Fig. 6, additional fields added to the links before pushing to the index and fields with different content are displayed. In addition to the fields shown in

Fig. 3, the links in the index feature the fields fromID and toID, which contain the persistent identifiers of the respective entities instead of the entities' URIs. The fromEntity and toEntity fields are, however, still present so that every entity in the index is connected to the respective entity in the link database. The fields fromView and toView contain the content of the entities' entityView fields and fromType and toType their entityType field content. The field linkReason contains the text snippet as a string instead of the URI of the TextualReference entry. Hereby, each link instance contains all information needed to display the link in an search portal's result list without needing additional queries to the index. The metadata in the link database and index can be used to enhance the presentation of the information in the portals. For example, the linkView field can be used to group links by their name. By displaying the linkReason text snippets, users can get a glimpse on how an entity is referenced in the linked entity, i.e. in what relation precisely they stand. The entityRelations can be used to improve ranking, i.e. to give priority to exact matches over partly matching dataset records. Information about the source of an entity or link in the field provenance can be used to filter searches. Figure 7 shows how the fields linkView and linkReason are displayed in the GESIS search portal.

Erwerbsarbeit und Erwerbsbevölkerung im Wandel : Anpassungsprobleme einer alternden Gesellschaft

Frankfurt am Main, Campus Verl., 1998, 281 S., (Veröffentlichung aus dem Verbund Arbeits- und Innovationspotentiale im Wandel)

Abstract: "In der Öffentlichkeit wird die künftige demographische Entwicklung - namentlich die sich abzeichnende Überalterung der deutschen Bevölkerung - vor allem unter zwei gegenläufigen Gesichtspunkten als problematisch wahrgenommen. Auf der einen Seite sieht man die Finanzierung der sozialen Sicherung durch einen erheblichen Rückgang der Beitragszahler strukturell gefährdet. Auf der anderen Seite drohe die absehbare Schrumpfung der Bevölkerung im erwerbsfähigen Alter zu einem Fachkräftemangel zu führen. Dieses Szenario einer demographisch bedingten Umkehrung der gegenwärtigen..." [more](#)

Institution(s): Internationales Institut für Empirische Sozialökonomie gGmbH (INIFES), Institut für Sozialwissenschaftliche Forschung e.V. ISF München, SÖSTRA Institut für Sozialökonomische Strukturanalysen GmbH

Keywords: [demographische Lage](#), [sozialer Wandel](#), [Erwerbsarbeit](#), [Bevölkerung](#), [Beschäftigung](#), [Arbeitsmarkt](#), [Anpassung](#), [Lebensalter](#)

Document type: Buch

Database: SSOAR - Social Science Open Access Repository

Full text

[Link](#)

[URN](#)

Actions

[Cite](#)

[search in Google Scholar](#)

[search in Google Books](#)

References (354)
▼

Data citation for: "SOEP 1995"
▲

The following text passage (s) in the publication with the mention "SOEP 1995" indicate that one or more of the research data listed below have been used to produce the publication:

"Datenbasis ist das Sozio- oekonomische Panel (SOEP) (Projektgruppe Panel 1995)."

[Sozio-oekonomisches Panel \(SOEP\), Daten der Jahre 1984-2015 \(internationale Version\)](#)

[Schupp, Jürgen; Goebel, Jan; Kroh, Martin](#)

Abstract: International Science Use Version der SOEP-Daten (95%-Version des Datensatzes <http://dx.doi.org/10.5684/soep.v32>). Dieser Datensatz ist zur weltweiten Nutzung freigegeben.

Actions

[Cite](#)

Fig. 7. Presentation of linkView and linkReason in the GESIS search

4 Evaluation

The LOD infrastructure is used productively in the new GESIS search portal that provides access to a range of different linked social science information in an integrated way. To evaluate the user experience of linked information between different information types, we performed a user study with 17 participants (7 female, 10 male, average age 33.35 years (SD = 10.04)). All participants work at German universities, three as professors, four as postdocs, nine as research associates, and one as a student assistant. They were recruited by email invitations. To date of the user study, the portal included publications, scales, projects, and institutions. The study is based on our use case introduced in Sect. 2. It was conducted in two steps: (1) a prescribed evaluation scenario to familiarize participants with interlinked information and (2) a free exploration phase where participants had time to use the prototype in the context of their own research interests.

In the evaluation scenario the users had to perform the following actions:

1. They were provided with a detailed information view of a literature entry (as shown in Fig. 7) as a starting point and were asked to find information about the research data that are cited in the paper.
2. After following the link to one of the research data sets and getting to the corresponding detailed information view, they were asked to find information about the project in which context the data set was created and which other studies had also applied the same survey instrument.
3. After checking the project information, they clicked on the corresponding link to the survey instrument entry and saw on the detailed information view a list of other research data that have also applied this survey instrument.

We encouraged the participants to think-aloud during both steps to get their direct feedback and asked at the end for their assessments regarding usefulness, trust in the provided links and completeness of the linked information.

12 participants found the links very useful, four useful and one found it neither useful nor not useful (collected through a Likert-scale ranging from 1 “not useful at all” to 5 “very useful”, mean = 4.65, SD = 0.5). One participant stated that she was enabled to get connections that otherwise would be very hard to find. 14 participants indicated they trusted that the provided links would lead them to the right information. Yet, some of them mentioned that their confidence was based on their good experience with GESIS so far. Completeness of the links was expected by five participants. The other twelve subjects appreciated the additional information provided but didn’t expect any system developed with a reasonable amount of effort to be able to show all connections. At the end, we asked participants if knowledge about the origin of the links is important for them, especially whether they have been created manually or automatically. For 14 participants the origin of the links was not important. 5 persons added that it was more important that the links lead to the correct information than how the link was created. However, three participants indicated they liked to know

the origin of the links. Two of them said it would increase their confidence in the information provided.

Besides the mostly positive feedback, there are also some challenges that arise on the user interaction side. The chance to get lost after following a couple of links is high. We observed this especially in the free exploration task. Participants had problems to get back to their starting point. We are currently exploring visualization techniques that should help to keep the overview. Furthermore, after following a couple of links, the relation to the original information need gets lower: participants had problems to understand that all publications of a project were listed and not only those that were related to their initial search query. In this case, a ranking of the linked information items according to the information need or an appropriate labelling might be helpful.

5 Related Work

There are similar approaches considering the integration and linking of heterogeneous Linked Data. The tool Karma [12] allows data integration from a variety of data formats, their conversion and mapping as well as to push them into a search index. However, Karma does not address data enrichment tasks like entity disambiguation and link merging which was necessary for our use case. Built on Karma is the approach of [7] which uses Karma in order to allow an on-the-fly integration of static and dynamic Linked Datasets. The framework LDIF [18] converts heterogeneous Linked Data sources into one representation which can then be further processed. When the original data is not RDF, it has to be converted beforehand. The output is either N-Quads or N-Triples which needs to be transformed again in order to push them into a search index.

Established linking tools like Limes [15] and Silk [22] enable to find links between Linked Datasets. In contrast, our approach currently focuses on already existing, manually created links between datasets or on identification of implicit links between datasets by parsing pdf documents. In our approach, linking tools could be used in order to find links to additional Linked Datasets in the LOD cloud.

The Open PHACTS Discovery Platform [6] allow users to perform complex queries over a variety of integrated and linked RDF sources of the pharmacological and physiochemical domain. Once the data is available in RDF format it can be integrated and is linked by the Large Knowledge Collider (LarKC) [3] using its Identity Mapping Service and ConceptWiki. The data is available via an API for further use, e.g. in search platforms. A similar approach was developed for educational resources [4]. In this approach education resources available on the web and datasets from the LOD cloud have been integrated and linked and made available via an API. However, no data enrichments tasks comparable to our use case have been conducted. Similar approaches for other domains have been developed in [2,11]. The Semantic Web index Sindice [16] was a lookup index for Semantic Web documents which allowed for searching over different, even unconnected Linked Datasets. Being a pure index, the challenges of data integration, linking and enrichment were not addressed.

None of the of the mentioned approaches covers the full workflow from non-Linked Data sources via integration and linking up to a provision via an efficient and searchable index.

6 Conclusion and Outlook

The presented LOD backend infrastructure has been developed for research information of the social sciences. However, although presented among the use case of GESIS, it can be applied and adjusted for similar use cases, since all components have been developed independently of any specific portal or metadata schema. The source code of the underlying InFoLiS infrastructure is available via GitHub¹⁰ for further reuse.

When integrating datasets from different data collections of a domain, one will most likely face the occurrence of equivalent person names in different collections. This problem can be solved with author disambiguation algorithms. Our basic entity disambiguation methods do not yet make use of these. Another potential extension is the linking to external Linked Datasets, e.g. in the LOD cloud. In the context of research information, thesauri and authority data with persistent identifiers may lead to a benefit for users and improved retrieval. Moreover, they can serve as a linking hub in order to find related datasets.

Acknowledgements. Parts of the approach presented in this paper have been developed in the InFoLiS project funded by DFG (SU 647/2-1 and MA 5334/1-2). The concept of the LOD backend infrastructure has been created by Benjamin Zapilko and Katarina Boland in the internally funded project LOD infrastructure. All extensions have been developed by Katarina Boland. The user study has been conducted during the development of the GESIS search by Dagmar Kern. The GESIS search portal has been developed by Daniel Hienert.

References

1. Boland, K., Ritze, D., Eckert, K., Mathiak, B.: Identifying references to datasets in publications. In: Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F. (eds.) TPD 2012. LNCS, vol. 7489, pp. 150–161. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33290-6_17
2. Celli, F., Keizer, J., Jaques, Y., Konstantopoulos, S., Vudragovic, D.: Discovering, indexing and interlinking information resources. *F1000Research* **4**, 432 (2015). [version 2; referees: 3 approved]
3. Cheptsov, A., Assel, M., Gallizo, G., Celino, I., DellAglio, D., Bradesko, L., Witbrock, M., Della Valle, E.: Targe knowledge collider. A service-oriented platform for large-scale semantic reasoning. In: Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS 2011), ACM International Conference Proceedings Series, Sogndal, Norway (2011)

¹⁰ <https://github.com/infolis/>.

4. Dietze, S., Yu, H.Q., Giordano, D., Kaldoudi, E., Dovrolis, N., Taibi, D.: Linked education: interlinking educational resources and the web of data. In: The 27th ACM Symposium on Applied Computing (SAC-2012), Special Track on Semantic Web and Applications, 25–29 March 2012, Trento, Italy (2012)
5. Dua, A., Nelle, D., Stock, G., Wagner, G.G.: Facing the future: European research infrastructures for the humanities and social sciences (2014)
6. Gray, A.J.G., Groth, P., Loizou, A., Askjaer, S., Brenninkmeijer, C., Burger, K., Chichester, C., Evelo, C.T., Goble, C., Harland, L., Pettifer, S., Thompson, M., Waagmeester, A., Williams, A.J.: Applying linked data approaches to pharmacology: architectural decisions and implementation. *Semant. Web* **5**(2), 101–113 (2014)
7. Harth, A., Knoblock, C., Stadtmüller, S., Studer, R., Szekely, P.: On-the-fly integration of static and dynamic sources. In: Proceedings of the Fourth International Workshop on Consuming Linked Data (COLD 2013) (2013)
8. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web Theory and Technology*, 1st edn, pp. 1–136. Morgan & Claypool, San Rafael (2011)
9. Hey, T., Tansley, S., Tolle, K.M.: *The Fourth Paradigm: Data-Intensive Scientific Discovery*, vol. 1. Microsoft Research, Redmond (2009)
10. Isaac, A., Haslhofer, B.: Europeana linked open data-data.europeana.eu. *Semant. Web* **4**, 2917 (2013)
11. Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S.M., Martin, M., Le Novere, N., Parkinson, H., Birney, E., Jenkinson, A.M.: The EBI RDF platform: linked open data for the life sciences. *Bioinformatics* **30**(9), 13381339 (2014)
12. Knoblock, C.A., Szekely, P., Ambite, J.L., Gupta, S., Goel, A., Muslea, M., Lerman, K., Taheriyani, M., Mallick, P.: Semi-automatically mapping structured sources into the semantic web. In: Proceedings of the Extended Semantic Web Conference, Crete, Greece (2012)
13. Lyon, L.: The informatics transform: re-engineering libraries for the data decade. *Int. J. Digit. Curation* **7**(1), 126–138 (2012)
14. Mathiak, B., Boland, K.: Challenges in matching dataset citation strings to datasets in social science. *D-Lib Mag.* **21**(1/2), 23–28 (2015)
15. Ngonga Ngomo, A.C., Auer, S.: LIMES: a time-efficient approach for large-scale link discovery on the web of data. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI 2011), vol. 3, pp. 2312–2317. AAAI Press (2011)
16. Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., Tummarello, G.: Sindice.com: a document-oriented lookup index for open linked data. *IJMSO* **3**, 37–52 (2008)
17. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: Mika, P., et al. (eds.) *ISWC 2014. LNCS*, vol. 8796, pp. 245–260. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11964-9_16
18. Schultz, A., Matteini, A., Isele, R., Mendes, P., Bizer, C., Becker, C.: LDIF - a framework for large-scale linked data integration. In: 21st International World Wide Web Conference (WWW 2012), Developers Track, Lyon, France (2012)

19. Simon, A., Wenz, R., Michel, V., Di Mascio, A.: Publishing bibliographic records on the web of data: opportunities for the BnF (French National Library). In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 563–577. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38288-8_38
20. Smith-Yoshimura, K.: Analysis of international linked data survey for implementers. *D-Lib Mag.* **22**(7/8), 110 (2016)
21. Van Hooland, S., Verborgh, R.: *Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish Your Metadata*. Facet Publishing, Croydon (2014)
22. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and maintaining links on the web of data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 650–665. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04930-9_41