Making Data Citable
A Web-based System for the Registration of Social and Economics Science Data

Dimitar Dimitrov, Erdal Baran and Dennis Wegener
GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany

Keywords: Citable Research Data, DOI, Registration Agency, Information System.

Abstract: Reliable identification and citation of research datasets, used to answer particular research questions, is currently limited. Even with well agreed standards, classic bibliographic methods of data citation have their limits when research datasets have been used several times or are stored in different locations. In this paper, we present the web-based information system da|ra, which aims at addressing this issue. We introduce a technical architecture which allows the registration of metadata of research datasets. The system allows a user to get a DOI name for these datasets, to search for registered datasets and to resolve DOI names. Today, our system is used by 12 publication agents and includes more than 6,000 research datasets that can be searched and cited using their DOI name.

1 INTRODUCTION

Today, we are experiencing an explosion of data, since producing data has become very easy and inexpensive. This is not only true for the common Internet user who, e.g., writes blogs and publishes photos, but also for the scientific community. Different tools support producing research datasets of various types. However, the exact citation and referencing of these datasets poses a problem for researchers and thus limits, e.g., the reproducibility of experiments on scientific data. This is surprising, as the possibility of getting citations for datasets would intrinsically encourage researchers to publish their data to earn reputation and acknowledgement.

A solution to this problem is the use of persistent identifiers for datasets as, e.g., offered by the DOI (Digital Object Identifier) system (DOI Foundation, 2012). In cooperation with DataCite (DataCite, 2012), an international initiative to establish easier access to research data, the Leibniz institutions GESIS (GESIS, 2012) and ZBW (ZWB, 2012) offer DOI registration for social science and economic data in Germany. The registration agency for social and economics science data da|ra (http://www.da-ra.de) aims at developing an infrastructure to attach DOI names to research data and make them findable and citable (Hausstein and Zenk-Möltgen, 2011).

In this paper, we present the current web-based information system for the registration and retrieval of social and economic science data, which was developed in the da|ra project. The key objective of the system is to capture metadata of social and economic datasets, making them citable by registering a DOI names as persistent identifier, and making them searchable on the Web. The data centres that register data, also called publication agents, are responsible for providing the metadata and specifying the correct landing page for resolution, as well as for taking care of their up-to-dateness. da|ra is responsible for the DOI registration and the metadata maintenance.

The remainder of the paper is as follows: An overview over prior and related work is given in Section 2. Section 3 introduces the domain-specific metadata schema of the da|ra system. In Section 4 we present the da|ra system architecture and introduce the process for the registration of a dataset with da|ra. Finally, Section 5 concludes and presents an outlook.

2 PRIOR & RELATED WORK

Data available on the Web can be very dynamic and often changes over time. However, this can be a curse if we want the data to be accessible and reusable. Thus, we can attach a reference to the data that can be resolved and points to the recent location of the data. Such a reference is called persistent identifier (PID). The most important properties of a

Dimitrov D., Baran E. and Wegener D..
Making Data Citable - A Web-based System for the Registration of Social and Economics Science Data.
DOI: 10.5220/0004371601550159
In *Proceedings of the 9th International Conference on Web Information Systems and Technologies* (WEBIST-2013), pages 155-159
ISBN: 978-989-8565-54-9

persistent identifier are uniqueness, which can be addressed by defining namespaces or using special identifier generation strategies, and resolvability, which means that the identifier can be resolved persistently. Further important properties in the context of PID systems are, e.g. the association of metadata with the identifier, the ability to incorporate legacy identifiers or identifiers of other types, or the handling of versioning, granularity and management of the PIDs (Ball and Duke, 2012). In general, we can distinguish two categories among the systems assigning persistent identifiers: systems that store metadata associated with the PID and systems that do not store metadata. The main part of the systems storing metadata has a basic metadata schema, which often consists of Dublin Core elements.

The DOI Foundation provides a managed resolution system for identifiers. A DOI name may be represented as a URL by prefacing the string http://dx.doi.org/ to the DOI of the document (e.g., the DOI name 10.4232/1.11380, can be resolved by http://dx.doi.org/10.4232/1.11380).

One of the biggest PID systems is Crossref (Crossref, 2012), which is mainly registering DOI names for different literature types. DataCite is registering DOI names, but their focus is on PIDs for datasets. DataCite also provides a very general metadata schema for datasets of all types. Furthermore, several institutions exist, e.g. national libraries, which allow registration of URNs (Daigle et al., 2002) for publications. We build our system on the services provided by DataCite, since the purpose of DataCite is to promote science and research, which perfectly matches our use cases. Thus, we use DOI names as PIDs (Hausstein, 2012).

3 METADATA SCHEMA

The main goal of the da|ra information system is the registration of scientific social and economic datasets and to allow for searching for metadata of research datasets. Typical data in social sciences is empirical primary data from survey research, historical social research and texts for content analyses. The typical economics data is statistical data collected with surveys of individuals, companies or states but also data representing experiment results.

The main requirements when developing the da|ra metadata schema to describe the data were the following: (1) Interoperability with other standards such as the DDI metadata specification (DDI, 2012) and the Dublin Core Metadata Initiative (DCMI); (2) Quality assurance of metadata; (3) Sustainability, e.g. the availability for semantic web applications.

The metadata schema of da|ra is implemented as XML Schema Definition (Hausstein et al., 2012) and is partially based on the metadata schema of the Metadata store of DataCite (Starr et al., 2011). As we are interfacing with the DataCite services, we incorporated all required fields of the metadata store schema in our schema, but also adapted and introduced new fields. The following fields are considered as the minimal set of fields required for a citation of a dataset: Title; Principal Investigator; Publication Agent; DOI; URL; Publication Date. Since da|ra does not store the data itself but only the metadata, the mandatory field 'Availability' additionally holds information about the access status of the dataset.

The da|ra schema includes 28 optional fields to give users the possibility to describe social and economic science data in detail, e.g. by fields such as Data Collector, Sampled Universe, Sampling, Temporal Coverage, Time Dimension, Collection Mode, Data, and Publication. These additional fields also increase the visibility of the datasets and make them easier to be found by a domain expert.

In the da|ra system, editing of metadata is supported by controlled vocabularies in order to support quality assurance and standardization. Hence, some fields of the da|ra metadata schema accept only values from controlled vocabularies from the social and economic sciences, such as TheSoz (Thesaurus Social Sciences) (Zapilko et al., 2012) or STW (Thesaurus for Economics) (Gastmeyer, 1998). For each controlled field there exists also a free text field to increase flexibility.

Versioning and granularity are issues in the context of persistent identifiers. In da|ra, we offer a comprehensive versioning mechanism and let the publication agents decide how to use it. For example, publication agents can register a new DOI name for each version of the metadata or update the existing metadata in order to, e.g., remove typos. Publication agents are also free to decide on the granularity of the datasets, which means that it is also possible to assign a DOI name for a package, e.g. a CD containing several datasets.

4 SYSTEM ARCHITECTURE

In this section, we give an overview over the architecture of the da|ra information system. The architecture of our system is visualized in Figure 1. On the left, we see the two types of user groups, *Publication Agents* and *Researchers*. The main difference be-

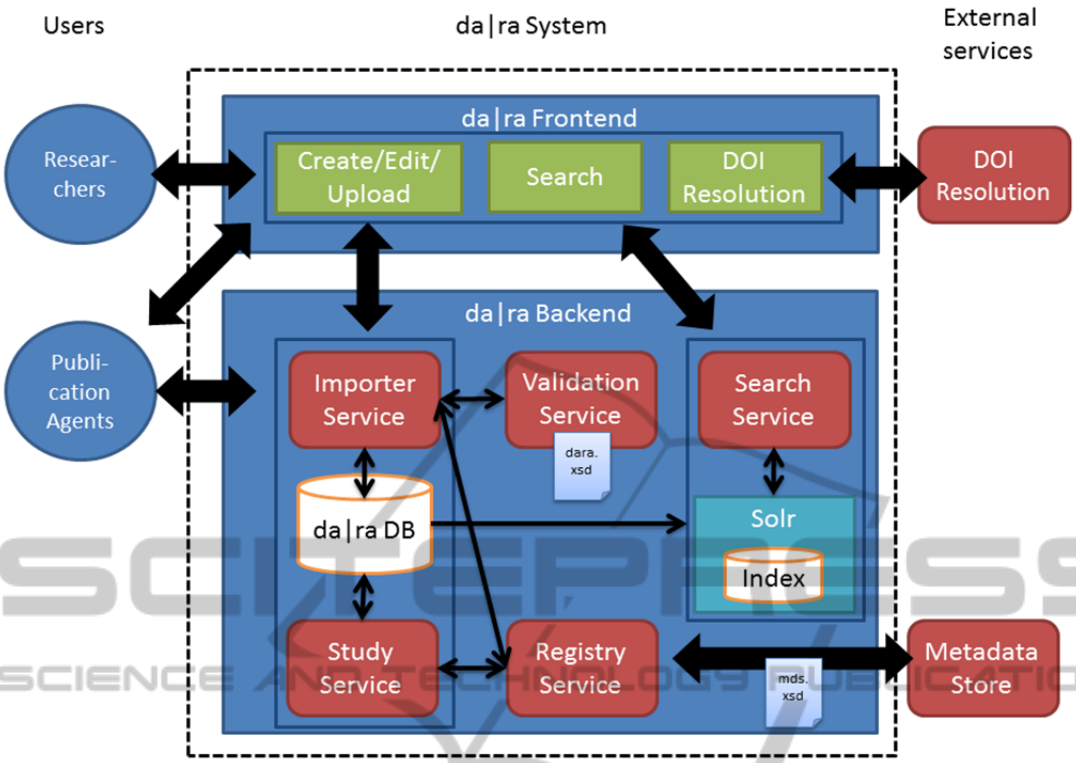

Figure 1: Technical overview of the da|ra information system.

tween these is that the researcher cannot create or edit a dataset. On the right, the external services interacting with da|ra - the *DOI resolution* and *Metadata Store* services - are visualized. In the middle, the details of the frontend and the backend of our system are visualized. The main frontend components are *Search*, *Create/Edit/Upload*, and *DOI resolution*. In the following, we present the details of the internal components and how they interface with the external services.

4.1 Create, Edit & Upload

A publication agent can use the da|ra frontend to create and then edit a dataset. To create a dataset one has to fill out the create form, which covers the basic metadata fields. After that the data is saved as a dataset and, if chosen by the publication agent, registered. Then, the dataset can be edited and the additional domain-specific metadata fields can be filled out in the edit form. When creating a dataset, one can choose whether to provide a DOI proposal or to let the system generate one. As mentioned before, versioning is crucial in the context of PIDs. With our system, a publication agent selects either to use only a 'Study ID' and let the system generate the 'Version', or to provide both manually, or to let the system generate both.

The importer and study service are the only services that have reading and writing access to the da|ra database. These two services are supporting the 'create/edit and upload' components offered in the frontend. The study service supports the form-based data manipulation. The importer service takes an XML based data description valid with the dara.xsd as input. This request will create an entry in our database if there does not yet exist an entry with the given 'Study ID' and 'Version'. Otherwise, it will update the existing entry. The importer service can be used via the frontend by a form for XML upload, or via the service API.

The DOI registration is performed by the registry service. The registry service acts as a proxy of the services of the metadata store. These provide the functionality for 'DOI registration' and 'metadata upload'. In addition, they can be used to set the status of a DOI name 'inactive' to deactivate a DOI, e.g. if the landing page of the DOI is unreachable. When registering a DOI, two key-value pairs with DOI name and URL have to be passed. If the DOI name already exists, it will be reminted. Otherwise it will be registered. The metadata upload stores a new version of the metadata for a given DOI. It takes an

XML, which is valid against the Metadata Store schema, as input.

4.2 Search & Doi Resolution

The search component is supported by the search service through the Solr framework. We offer quick and advanced search forms. The quick search is performed over all metadata fields. In the advanced search, one can narrow the search by stating explicitly the title, DOI, version, principal investigator, publication date or data centre (publication agent). Furthermore, the search result of a request can be filtered/narrowed with facets, which allows faster and easier finding of specific information about a dataset. We defined seven facets: Data Center, Principal Investigator, Data Collector, Collection Mode, Keywords, Availability and Publication Date.

The structure of the Solr index is given through the Solr indexing schema. In this schema, we defined the advanced and faceted search fields as static fields and all other as dynamic fields. In addition, for the quick search, we defined also a static field called default search. All dynamic and static fields are copied to this field to perform a quick overall search.

Every new dataset created by da|ra is integrated into the Solr index by the search service. In addition, the search service manages updates. Depending on whether logged-in or not, the user gets different search results. For a non-logged user, we present only the registered datasets. The logged-in user gets additionally his/her own not yet registered datasets as search result.

The DOI resolution component is based on a service that is used to resolve a given DOI name, provided by the DOI Foundation.

5 CONCLUSIONS & OUTLOOK

Reproducibility of research processes is essential for every science discipline. Often, the reproduction of a research process is impossible without the primary data that was used in the process. The da|ra system supports the demand for the ability to find and to precisely cite primary data. In this paper, we presented the architecture of the da|ra information system, which allows registering and citing datasets by using DOIs as persistent identifiers. It is based on the da|ra metadata schema, which matches the needs of the social science and economics to describe their datasets. The da|ra information system is implemented using the Grails framework and is publicly available at http://www.da-ra.de (see Figure 2 for a screenshot). Today, our system is already used by 12 publication agents and includes more than 6,000 registered research datasets that can be cited using their DOI. Based on the resolution statistics from DataCite, we can see that the DOIs are frequently used: in total, about 10700 registered DOIs of the datasets in our system are resolved per month; these covered 4170 unique DOIs (average over July-September 2012).In future work, we will focus on linking the datasets in our system with other repositories, e.g. literature repositories. For doing so, techniques from InFoLiS (Boland et al., 2012) could be integrated.

Figure 2: Screenshot of the da|ra system.

ACKNOWLEDGEMENTS

This work is jointly funded by the German Research Foundation (DFG) and GESIS as part of the da|ra project. We would like to thank the members of the da|ra team for their comments and support.

REFERENCES

Ball, A. and Duke, M. (2012). Data Citation and Linking. DCC Briefing Papers. Digital Curation Centre.

Boland K., Ritze D, Eckert K. and Mathiak B. (2012), Identifying References to Datasets in Publications. TPDL 2012: *Theory and Practice of Digital Libraries.* Paphos, 23.-27.09. 2012.

Crossref (2012). Crossref. http://www.crossref.org

Daigle, L. et. al. (2002). URN Namespace Definition Mechanisms. The Internet Society.

DataCite (2012). DataCite e.V. – International Data Citation. http://datacite.org.

DDI Data Documentation Initiative (2012). http://www.ddialliance.org/

DOI Foundation (2012). DOI Handbook. doi:10.1000/182.

Gastmeyer, M. (1998). *Standard-Thesaurus Wirtschaft*. Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Kiel.

GESIS (2012). Leibniz Institute for the Social Science. http://www.gesis.org

Hausstein, B., Zenk-Möltgen, W. (2011). da|ra – Ein Service der GESIS für die Zitation sozialwissenschaftlicher Daten. In: Digitale Wissenschaft: Stand und Entwicklung digital vernetzter Forschung in Deutschland. Beiträge der Tagung vom 20./21. September 2010, Köln, pp. 139-147.

Hausstein, B. (2012). Die Vergabe von DOI-Namen für Sozial- und Wirtschaftsdaten - Serviceleistungen der Registrierungsagentur da|ra. SSRN Electronic Journal. DOI: 10.2139/ssrn.2008192

Hausstein, B.; Quitzsch, N., Jeude, K., Zenk-Möltgen, W., Schleinstein, N., (2012). da|ra Metadatenschema Version 2.2.1

Starr, J., Ashton, J.,Brase, J., Bracke, P., Gastl, A., Gillet, J., Heller, A., Krog, B., McAvoy, L., Morgenroth, K., Newbold, E., de Smaele, M., Wilde, A., Yeadon, S., Zenk-Möltgen, W. and Ziedorn, F. (2011). DataCite Metadata Schema for the Publication and Citation of Research Data. doi:10.5438/0005.

Zapilko, B., Schaible, J., Mayr, P. and Mathiak, B. (2012). *TheSoz: A SKOS Representation of the Thesaurus for the Social Sciences*. in: Semantic Web Journal.

ZBW (2012). Leibniz Information Centre for Economics. http://www.zwb.eu