Where are the Datasets? A case study on the German Academic Web Archive.

Yousef Younes
Yousef.Younes@gesis.org
GESIS – Leibniz institut for Social Sciences
Cologne, Germany

Sebastian Tiesler
sebastian.tiesler@hu-berlin.de
Humboldt-Universität zu Berlin
Berlin, Germany

Robert Jäschke
robert.jaeschke@hu-berlin.de
Humboldt-Universität zu Berlin
Berlin, Germany

Brigitte Mathiak
brigitte.mathiak@gesis.org
GESIS – Leibniz institut for Social Sciences
Cologne, Germany

ABSTRACT

The German Academic Web (GAW) is a longitudinal archive of websites from German academic institutions, mainly universities. It can support answering research questions about academia in Germany. Recent discussions about reproducible research have brought the availability and sharing of research data into focus. Collecting, linking, and providing metadata about research data is thus an important task for infrastructure facilities. In this work, we examine how existing datasets are linked and referenced on German academic web pages using the GAW archive. For that, we use the social sciences and economics datasets registered at da|ra as our case study. The results show that academic web pages as presented in GAW are not a good foundation to answer dataset-related questions. But from the few results found, it was obvious that da|ra datasets are usually mentioned using their DOIs and not their URLs.

KEYWORDS

research data, data findability, web archiving, German Academic Web

ACM Reference Format:
Yousef Younes, Sebastian Tiesler, Robert Jäschke, and Brigitte Mathiak. 2018. Where are the Datasets? A case study on the German Academic Web Archive.. In *Proceedings of Web Archiving and Digital Libraries Workshop (WADL).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Re-using research data has become an important driver of scientific innovation. Aggregating, replicating, or applying different methods to existing data leads to new insights and increases the quality of the underlying research results, while lowering costs [6]. As research data becomes more important, so does archiving information on research data. From an observation study, we know that relevant information on research data is not limited to the data itself, but also on a variety of web sites, such as project websites [8].

The web is also an important information source for researchers looking for data. According to a survey conducted among 1,637 researchers from all disciplines [4] 59 % use web search engines to find data often. Other surveys, for example, among social scientists [3] confirm that web search is a very important part of their data discovery process. As such, archiving the research data itself is not enough, its traces in the web need to be archived as well to get a full picture.

This leads to the following research questions:

(1) How can a web archive be used to find references to research datasets?
(2) Which identifiers for datasets can be found?
(3) How does the volume of referenced datasets change over time?

As these questions in their totality cannot be answered easily, due to scaling effects, we are instead focusing on a specific subset: datasets from social sciences and economics as registered through the registration agency da|ra[1]. The web archive The German Academic Web[2] is employed to answer these questions.

This paper is structured as follows: In Section 2 we discuss related work and in Section 3 we explain the methods we have used for our experiments. The results are presented in Section 4, followed by a discussion in Section 5.

2 RELATED WORK

Research data is traditionally stored in databases or data repositories and only recently opening up to web infrastructure, for instance, through the application of the FAIR principles [14]. Schema.org added *Dataset* as an entity type in 2013, which can be used to provide metadata on research data through markup. This is used, for example, by Brickley et al. to build catalogues [2]. But they observed that metadata markup, although it is rather simple, it needs proper curation, as not every *Dataset* entity is describing a dataset [1]. However, not all data repositories adhere to this recommendation yet [10]. Instead, metadata is represented on plain web pages. Thompson et al. did a longitudinal analysis of Common Crawl data [13] to find out about the use of persistent identifiers. They suggest

[1] https://www.da-ra.de/
[2] https://german-academic-web.de/

WADL, June 20, 2022, Virtual

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

to use DOIs over URIs to identify scholarly publications on the web. Like metadata, persistent identifiers come with their own set of problems. For DOIs it is critical to maintain the mapping between DOI and resource location over time and to deliver a consistent response to DOI queries [7].

3 METHOD

In this work, we track mentions of dataset identifiers in an academic web crawl. Particularly, we search for datasets from social sciences and economics registered at the da|ra registration agency in the German Academic Web (GAW). The search process is performed using two dataset identifiers against multiple GAW snapshots. First, we describe the GAW data and how it is collected (Section 3.1). Then we introduce the da|ra system (Section 3.2). Finally, the experimental setup is explained (Section 3.3).

3.1 German Academic Web

The *German Academic Web* (GAW) [11] is a collection of snapshots of German academic institutions' web sites. It is a domain-specific longitudinal web archive and was created to preserve the websites of German academic institutions. By the time of this writing, GAW contains nineteen snapshots obtained by crawling on a biannual basis since 2013 in addition to one snapshot from 2012. Each of these snapshots occupies about 6-8 TB of storage and involves around 100 million breadth-first crawled web pages (text, PDF, and images) stored as WARC files which in turn contain several WARC records. Every crawl is performed using a recent version of the Heritrix[3] web crawler initialised with a seed list of 150 domains associated with all German academic institutions who have the right to award doctorates. The characteristics of the crawling process change over time, for example, a new domain could be added to the seed list if a new university is created or one URL could be retired if it was found to be out of scope (e.g., an e-learning system or a file repository) [11]. These changes are ignored in this paper, because one of the goals is to find which web pages are most likely to contain dataset mentions.

The experiments are conducted on a collection of crawls that consist of the mid-year crawls from 2016 to 2021. Earlier crawls were omitted as da|ra was not fully online prior to 2016. From these crawls, only web pages whose content is text and which were available at the time of crawling are selected. This is achieved by choosing WARC records with MIME type text/html and HTTP status code 200.

3.2 da|ra

The availability of research data is a precondition to make the research results reproducible. To help achieve this availability for social sciences and economics data, GESIS[4] (Leibniz Institute for the Social Sciences) and ZBW[5] (Leibniz Information Centre for Economics) launched da|ra, in 2014, as a registration agency [9]. Beside its registration and archiving services, da|ra offers the metadata of its registered datasets for harvesting through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[6]. It also provides a DOI resolver service.

In this work we are interested in quantifying identifier mentions of the da|ra registered datasets in the GAW archive. We use identifiers available in da|ra for its registered datasets as part of their metadata. da|ra offers multiple identifiers in their metadata. We can differentiate between two types of identifiers: *mandatory* and *optional*. The mandatory identifiers are available for every dataset at da|ra; while the optional ones are only available for some of them. The mandatory identifiers are:

DOI: Digital object identifiers are unique, permanent and case-insensitive strings of alphanumeric characters that are used to identify digital resources (books, research data, etc.) [12]. There are 26,298 unique DOIs registered at da|ra (e.g. 10.17886/RKI-History-0011) – one for each resource. Among these DOIs, there are 17,723 DOIs that are associated with datasets. The other DOIs are for different types of resources such as text, image, service, software, etc. and are not considered for our analysis. DOIs can be resolved to their associated URLs using a DOI resolver, for example, https://doi.org/.

URL: Every da|ra dataset has a URL associated with it. There are 25,312 unique URLs at da|ra. Among them 17,084 are dataset-associated URLs. While different versions of the same dataset have different DOIs, they have the same URL. This is why the number of URLs is lower than the number of DOIs, despite the fact that they are both mandatory.

Titles: For every dataset there is at least one title. For some of them there is more than one. These titles are mostly in English and German but alternative titles in other languages such Chinese, Arabic, etc. exist but they are very rare. Titles do not have to be unique and often contain additional information, such as year of collection, acronyms, which makes many of them rather unwieldy and hard to search for (e.g., "German General Social Survey (ALLBUScompact) - Cumulation 1980-2018").

The optional identifiers are:

URN: Uniform Resource Name is a unique and permanent identifier that uses the URN schema. Unlike DOI, URNs are only resolvable through the assigning institutions web page which make them useful for locally closed systems. da|ra has 345 items with such an identifier (e.g., urn:nbn:de:0168-ssoar-383499).

GESIS-specific identifiers: These identifiers come from the GESIS Archive. There are 6,428 datasets with such identifiers (e.g., ZA0790).

For the sake of this analysis, we focus on the da|ra resources with type *Dataset*. Each of these datasets has a couple of identifiers. Figure 1 shows a distribution of da|ra dataset identifiers. As we can see, every dataset has one URL and one DOI but one or more titles. About a third of the datasets are from GESIS and thus have its identifier. Based on that, we choose to use URL and DOI identifiers, since they are available and unique for every dataset.

[3] https://github.com/internetarchive/heritrix3/wiki
[4] https://www.gesis.org/
[5] https://www.zbw.eu/
[6] https://www.da-ra.de/oaip/

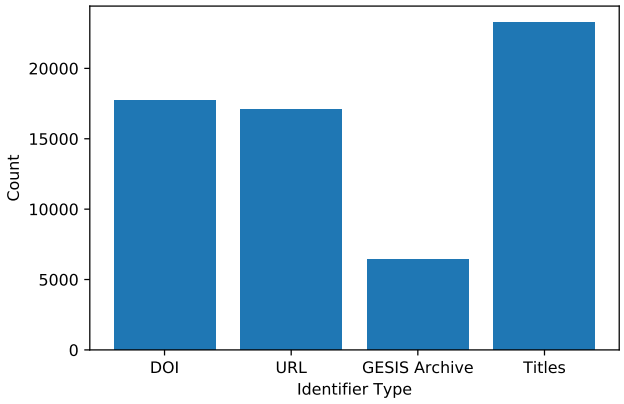


Figure 1: Distribution of the different types of dataset identifiers available in da|ra.

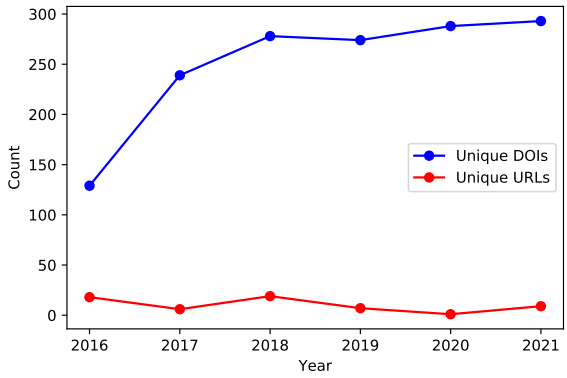


Figure 2: Distribution of the different types of identifiers in GAW over time.

3.3 Experimental Setup

Da|ra provides us with two lists of URL and DOI identifiers for its registered datasets. Our goal is to find web pages in GAW that mention da|ra datasets using the two chosen identifiers. After a preliminary analysis we performed some pre-processing to solve some of the problems we had identified.

First, we found that the research data repository hosted on madata.bib.uni-mannheim.de was archived in the GAW crawls. Since the repository contains a subset of da|ra datasets, we would trivially find all URLs from those da|ra datasets in that subset of GAW. As the repository should have been excluded by the crawl scope of GAW anyway, we exclude dataset URLs to that host from the list of da|ra dataset identifiers.

Then, similar to the results in [1], we also found that some da|ra dataset identifiers are not proper. It should be implicit that the URL identifier of a dataset points to a web resource containing the dataset but that is not always the case. For example, the DOI 10.5684/soep.v36-RV.RTBN2018 uses the landing page http://www.fdz-rv.de/ as resource for the dataset and a set of 48 different DOIs with prefix 10.25654 point to the same landing page https://www.hamburg.de/bsb/ifbq. Having such URLs and DOIs as dataset identifiers would erroneously increase the number of matching pages and thus we also excluded those URLs and DOIs from the list of da|ra dataset identifiers.

Finally, the URLs, DOIs, and crawled web pages have to be converted to lower case to prevent case sensitivity issues. The URLs also need to be normalised by removing common prefixes (https://www., http://www., https://, http//) and suffixes (.html, .htm). Then a simple string search is applied using ArchiveSpark [5] over the full text of the six selected web crawls for the two chosen identifiers. After that the results are analysed on different dimensions to quantify the unique identifiers, hosts, and pages to draw conclusion based on that.

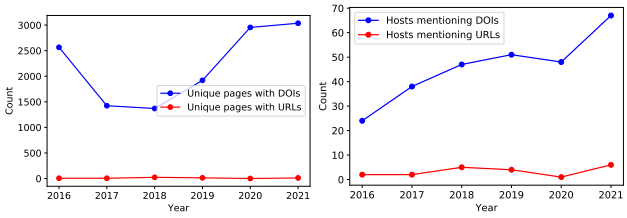
4 RESULTS

In Figure 2 we show the number of unique dataset URLs and DOIs found in six GAW crawls. Over the years, we observe only a small number of unique URLs and that number fluctuates in the range [1, 19], while the number of unique DOIs increases from 129 in 2016 to 293 in 2021. Although both numbers are small, it seems more common for GAW web pages to use DOIs when referring to datasets.

We also looked at the unique pages (Figure 3(a)) and unique hosts (Figure 3(b)). By 'hosts' we mean the first part of the URL up until the first slash excluding the http and www parts. Again, the number of unique hosts and pages that mention DOIs is higher compared to the ones that mention URLs. Both figures show an upward trend for DOI usage. Dips in the graphs could be the result of changes in the crawl scope or web pages retiring or moving their service. An example for a web page retiring its service in 2016 is the host dszbo-portal.uni-bielefeld.de. It is the "Datenservicezentrum Betriebs- und Organisationsdaten" ("Data Service Center for Business and Organizational Data").[7] In 2017 the new host fdzbo-portal.uni-bielefeld.de for "Forschungsdatenzentrum Betriebs- und Organisationsdaten" ("Research Data Center for Business and Organizational Data (RDC-BO)") comes into existence until it becomes part of DIW Berlin in 2019[8] and leaves the scope of the crawl (see also Table 1). Generally speaking, the number of hosts increases over the years which means that datasets are getting more common because they are being mentioned by an increasing number of different web sites.

Since the number of results for the URL identifier is (close to) zero, we analyse the found DOIs in more depth. Additionally, as the figures plot unique results and to show a different dimension of the results, Table 1 shows the hosts with more than 50 matching pages. Additionally, we added the number of unique DOIs mentioned in the pages per host. With this information, we can further evaluate if a host contains interesting information regarding our research question. Furthermore, comparing the number of results in this table to the number of unique pages in Figure 3(a) gives an indication of the amount of duplicate pages for each year.

[7] https://web.archive.org/web/20160321191513/https://dszbo-portal.uni-bielefeld.de/
[8] https://www.diw.de/de/diw_01.c.670982.de/



(a) unique pages (b) unique hosts

Figure 3: Distribution of the number of unique pages and hosts containing dataset identifiers over time.

From the table, we see that the host name iqb.hu-berlin.de, which is associated with the Institute for Educational Quality Improvement in Germany, is among the top hosts in all involved crawls. This institute is involved in empirical educational research in Germany and also hosts a research data repository, so it references many of the social sciences datasets.

5 DISCUSSION

In this work we have searched for the social sciences and economical datasets registered at da|ra in GAW using two different identifiers. We were somewhat surprised to find only so little on datasets on German academic web pages, given that we know that people use web resources to find information on datasets extensively [8]. Nevertheless, there are lessons to be learned.

The takeaway points from this work can be summarised as follows. First, using DOIs to search for datasets produces better results than using URLs as identifiers. Second, there needs to be a curation mechanism for the da|ra metadata to validate whether, for example, URLs provided refer to those datasets and thus make the metadata more reliable. Third, since we were able to find only a few dataset mentions in GAW, we can say that either GAW is not including such pages or, which is more likely, that it is just not common to cite datasets on web pages. Future work should focus on finding a way to find or track datasets over the years and categorise them according to their importance. This suggests introducing a specialised crawl and the results obtained here could be used for that task.

ACKNOWLEDGEMENTS

Part of this research was funded by the DFG project *Unknown Data – Mining and consolidating research dataset metadata on the Web* (grant number 460676019).

Table 1: Top hosts with 50+ matching pages for DOI mentions over the years. The column *pages* shows the number of matching pages that come from the associated host. The column *uDOIs* shows the unique number of DOIs the host refers to.

year	host	pages	uDOIs
2016	dszbo-portal.uni-bielefeld.de	2500	59
	iqb.hu-berlin.de	180	23
	uni-bielefeld.de	113	35
2017	fdzbo-portal.uni-bielefeld.de	1297	62
	fb03.uni-frankfurt.de	348	77
	goethe-university-frankfurt.de	174	77
	iqb.hu-berlin.de	139	38
	uni-bielefeld.de	114	36
2018	fdzbo-portal.uni-bielefeld.de	1125	63
	fb03.uni-frankfurt.de	324	85
	iqb.hu-berlin.de	163	45
	goethe-university-frankfurt.de	162	85
	uni-bielefeld.de	123	36
2019	iqb.hu-berlin.de	1667	52
	fb03.uni-frankfurt.de	439	89
	mzes.uni-mannheim.de	197	48
	goethe-university-frankfurt.de	171	88
2020	iqb.hu-berlin.de	2755	56
	fb03.uni-frankfurt.de	204	93
	mzes.uni-mannheim.de	185	48
	goethe-university-frankfurt.de	101	94
2021	iqb.hu-berlin.de	2770	18
	fb03.uni-frankfurt.de	222	102
	mzes.uni-mannheim.de	200	48
	goethe-university-frankfurt.de	110	103
	uni-bielefeld.de	64	25

REFERENCES

[1] Tarfah Alrashed, Dimitris Paparas, Omar Benjelloun, Ying Sheng, and Natasha Noy. 2021. Dataset or Not? A Study on the Veracity of Semantic Markup for Dataset Pages. In *The Semantic Web – ISWC 2021*, Andreas Hotho, Eva Blomqvist, Stefan Dietze, Achille Fokoue, Ying Ding, Payam Barnaghi, Armin Haller, Mauro Dragoni, and Harith Alani (Eds.). Springer International Publishing, Cham, 338–356.

[2] Dan Brickley, Matthew Burgess, and Natasha Noy. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *The World Wide Web Conference*. 1365–1375. https://doi.org/10.1145/3308558.3313685

[3] Tanja Friedrich. 2020. *Looking for data.* Ph. D. Dissertation. Humboldt-Universität zu Berlin, Philosophische Fakultät. https://doi.org/10.18452/22173

[4] K Gregory, P Groth, A Scharnhorst, and S Wyatt. 2020. Lost or Found? Discovering Data Needed for Research. *Harvard Data Science Review* 2, 2.2 (2020).

[5] Helge Holzmann, Vinay Goel, and Avishek Anand. 2016. ArchiveSpark. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. ACM. https://doi.org/10.1145/2910896.2910902

[6] Jonathan M Jeschke, Sophie Lokatis, Isabelle Bartram, and Klement Tockner. 2019. Knowledge in the dark: scientific challenges and ways forward. , 423–441 pages.

[7] Martin Klein and Lyudmila Balakireva. 2021. An extended analysis of the persistence of persistent identifiers of the scholarly web. *International Journal on Digital Libraries* (10 2021), 1–13. https://doi.org/10.1007/s00799-021-00315-w

[8] Thomas Krämer, Andrea Papenmeier, Zeljko Carevic, Dagmar Kern, and Brigitte Mathiak. 2021. Data-Seeking Behaviour in the Social Sciences. *International Journal on Digital Libraries* 22, 2 (2021), 175–195.

[9] Thomas Krämer, Claus-Peter Klas, and Brigitte Hausstein. 2018. A data discovery index for the social sciences. *Scientific Data* 5 (April 2018), 180064. https://doi.org/10.1038/sdata.2018.64

[10] Fidan Limani, Yousef Younes, Valentina Hiseni, Janete Saldanha Bach, Peter Mutschke, and Brigitte Mathiak. 2021. KonsortSWD Task Area 5 Measure 2 Report Scope: Milestones 1, 2, and 3. https://doi.org/10.5281/zenodo.5901207 Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of NFDI - 442494171.

[11] Michael Paris and Robert Jäschke. 2020. How to Assess the Exhaustiveness of Longitudinal Web Archives: A Case Study of the German Academic Web. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media (HT '20)*. ACM, New York, NY, USA. https://doi.org/10.1145/3372923.3404836

[12] Norman Paskin. 2010. Digital object identifier (DOI®) system. *Encyclopedia of library and information sciences* 3 (2010), 1586–1592.

[13] Henry S. Thompson and Jian Tong. 2018. Can Common Crawl reliably track persistent identifier (PID) use over time? https://doi.org/10.48550/ARXIV.1802.01424

[14] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3 (2016). https://doi.org/10.1038/sdata.2016.18