Helbig, Kerstin; Hausstein, Brigitte; Toepfer, Ralf

Article

Supporting Data Citation: Experiences and Best Practices of a DOI Allocation Agency for Social Sciences

Journal of Librarianship and Scholarly Communication

Suggested Citation: Helbig, Kerstin; Hausstein, Brigitte; Toepfer, Ralf (2015) : Supporting Data Citation: Experiences and Best Practices of a DOI Allocation Agency for Social Sciences, Journal of Librarianship and Scholarly Communication, ISSN 2162-3309, Pacific University Libraries, Forest Grove, Vol. 3, Iss. 2, pp. eP1220,
https://doi.org/10.7710/2162-3309.1220

This Version is available at:
http://hdl.handle.net/11108/215

JLSC

ISSN 2162-3309 | JLSC is published by the Pacific University Libraries | http://jlsc-pub.org

Volume 3, Issue 2 (2015)

Supporting Data Citation: Experiences and Best Practices of a DOI Allocation Agency for Social Sciences

Kerstin Helbig, Brigitte Hausstein, Ralf Toepfer

Helbig, K., Hausstein, B., & Toepfer, R. (2015). Supporting Data Citation: Experiences and Best Practices of a DOI Allocation Agency for Social Sciences. *Journal of Librarianship and Scholarly Communication,* 3(2), eP1220. http://dx.doi.org/10.7710/2162-3309.1220

ISSN 2162-3309 10.7710/2162-3309.1220

PRACTICE

Supporting Data Citation: Experiences and Best Practices of a DOI Allocation Agency for Social Sciences

Kerstin Helbig
Research Data Management Coordinator, Humboldt-Universität zu Berlin
Brigitte Hausstein
Head, da|ra, GESIS Leibniz Institute for the Social Sciences
Ralf Toepfer
Deputy Head of Electronic Publishing, ZBW Leibniz Information Centre for Economics

INTRODUCTION As more and more research data becomes better and more easily available, data citation gains in importance. The management of research data has been high on the agenda in academia for more than five years. Nevertheless, not all data policies include data citation, and problems like versioning and granularity remain. **SERVICE DESCRIPTION** da|ra operates as an allocation agency for DataCite and offers the registration service for social and economic research data in Germany. The service is jointly run by GESIS and ZBW, thereby merging experiences on the fields of Social Sciences and Economics. The authors answer questions pertaining to the most frequent aspects of research data registration like versioning and granularity as well as recommend the use of persistent identifiers linked with enriched metadata at the landing page. **NEXT STEPS** The promotion of data sharing and the development of a citation culture among the scientific community are future challenges. Interoperability becomes increasingly important for publishers and infrastructure providers. The already existent heterogeneity of services demands solutions for better user guidance. Building information competence is an asset of libraries, which can and should be expanded to research data.

Received: 02/18/2015 Accepted: 06/30/2015

Correspondence: Brigitte Hausstein, GESIS Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany, Brigitte.Hausstein@gesis.org

INTRODUCTION

Research data are an increasingly important basis for scientific knowledge, not only in science, technology and medicine but also in social sciences and economics. A major reason for the increasing importance lies in the better and easier availability of research data. Thus, the possibilities for re-analysis of existing datasets have improved significantly through the Internet. The ever growing technical capabilities enable more complex analyses (King, 2011). Furthermore, since the mid-1980s the share of empirical papers in top economic journals has climbed to more than 70%, and a substantial number of these papers (34%) use empirical data that have been gathered by the authors (Hamermesh, 2013, p. 168). As a reaction to this development funders and publishers started to implement data policies that demand proper data management. Not all of these policies cover data citation, though.

The management of research data has been high on the agenda in academia for more than five years. An ever increasing number of universities, libraries, and research institutions have started to support their researchers to cope with various aspects of research data management issues (Ludwig & Enke, 2013). As the German Data Forum (2010) points out:

> Historically, researchers paid little attention to the quality of metadata surrounding their work; archiving was left to archivists. This mind-set is changing. There have been rapid advances in the development and implementation of high-quality metadata standards, standards which apply to datasets throughout their life cycle from initial collection through to secondary use. (p. 20)

The value of research data archiving, sharing, and citation is manifold. First and foremost, it enables a more transparent research process, re-analysis, reproducibility of research findings, verification, and validation (Mooney & Newton, 2012). In addition, initial studies show that research articles with shared research data are more often cited and therefore lead to a higher impact factor (see e.g. Piwowar & Vision, 2013). The sharing of research data holds advantages that cannot be dismissed. Nevertheless, researchers still have reservations regarding data sharing and research data management that have to be overcome. Fear of misuse or plagiarism as well as legal insecurity worry academics and prevent further development. Moreover, the lack of formal recognition seems to be the main barrier of making data available. Researchers widely agree that it is beneficial for scientific progress to share data, and they agree that others should publish their data. However, only a minority of researchers actually share their data publicly. Fecher, Friesike, Hebing, Linek and Sauermann (2015) conclude that "[…] academia is a reputation economy, an exchange system that is driven by individual reputation beyond money and status. In this regard, data sharing will only see widespread adoption among research professionals if it pays in the form of reputation" (p. 3). To overcome this situation, appropriate reward structures for data sharing and easy and ready-to-use technical infrastructure are needed: "Robust data citation practices and infrastructure will play a critical role in the widespread adoption of data citation and in the promotion of data sharing and its benefits" (Altman & Crosas, 2013, p. 63). Against this background, GESIS–Leibniz-Institute for the Social Sciences (http://www.gesis.org) and ZBW–Leibniz Information Centre for Economics (http://www.zbw.eu) built up the registration agency[1] for social and economic data da|ra (http://www.da-ra.de).

da|ra offers information and publishes best practices on data citation and management with the aim to accelerate and support scholarly communication as well as the building of standards. Through the allocation of Digital Object Identifiers (DOI)—permanent, persistent identifiers used for citing and linking electronic resources—da|ra establishes the conditions for durable identification, localization and citation of research data. In this article we will concentrate on data citation because we believe that a proper data citation is a prerequisite for data publication and data sharing. Through our experiences with libraries, research data centers and other data providers in the social sciences and economics, we gathered knowledge about challenges of DOI usage. Discipline specific metadata, versioning, and granularity are therefore an essential part of this article. By sharing our knowledge, we hope to assist other disciplines that face similar challenges in data citation and management.

LITERATURE REVIEW

The history of research data citation with persistent identifiers is still quite young. The assigning of DOI names to scientific research data began in summer 2004, and DataCite, one of the registration agencies for DOI names, just recently celebrated its fifth birthday (Brase, Sens, & Lautenschlager, 2015). Consequently, literature of research data citation is still modest. Standards for the citation of research data were suggested quite early, and there is an increase in literature on this topic coinciding with the rise of interest in improved data management, though.

Data citation in the social sciences

One of the first to write about data citation was Sue Dodd (1979). During the expansion of machine-readable data files in the social sciences and the associated increased demands of archiving, she calls for bibliographic standards to cite quantitative social science data files properly. Dodd is already aware of the potential of secondary analysis of research data (1979,

[1] A DOI registration agency is officially accredited to register DOI names by the International DOI Foundation (IDF).

p. 77). The citation guidelines include authorship, title, subtitle (with geographic focus), edition, imprint (producer or distributor), series information, and note (e.g. resource type). Dodd also mentions possible variations in citation styles depending on journal standards (1979, p. 82). The only missing element in her bibliographic standard is the electronic location (Uniform Resource Locator (URL) or persistent identifier) of a data file, which is purely due to the time in which she wrote.

In 1995 Sieber and Trumbo examine secondary analyses and citations of General Social Survey datasets. In their study, they identify eight necessary elements for proper data citation. These include the citation of principal investigator(s), title, survey period or release date, producer or distributor including address, funder, resource type (machine readable or not), and presence or absence of a codebook (Sieber & Trumbo, 1995, p. 14). Thus, they reduce Dodd's citation standard by omitting edition and series information. Sieber and Trumbo (1995) conclude that data citation principles are still lacking among researchers of all disciplines and recommend the establishment of data citation norms as well as revised journal policies.

Another well-known recommendation for the citation of data in the social sciences is proposed by Altman and King (2007). Their suggestion aims to be universal for citing quantitative data, independent of field or journal. Altman and King deplore the then current situation in 2007, "Unfortunately, no such universal standards exist for citing quantitative data, and so all the problems […] exist now. Practices vary from field to field, archive to archive, and often from article to article" (p. 2). Altman and King (2007) suggest six minimum components required for proper data citation. These are author(s), publication date, title, unique global identifier, universal numeric fingerprint and a bridge service (URL) (Altman & King, 2007, p. 3). Most importantly, the unique global identifier should resolve not to the data itself but to a metadata page describing the research data (Altman & King, 2007, p. 5). In contrast to Dodd, Altman and King suggest publisher information and resource type as optional. They prefer to concentrate on technical information and a persistent resolution service to guarantee successful long-term citation. Altman and King do not include the subtitle and other cataloguing metadata.

Finally, Toby Green has to be mentioned. He criticizes the lack of rules to "publish, present, cite or otherwise catalogue datasets" (Green, 2009, p. 3). With Altman and King (2007) in mind, Green elaborates on and modifies their standard to adequately describe data from the Organisation for Economic Co-operation and Development (OECD) and other sources. These changes are based on bibliographic requirements, though many metadata fields are optional. Mandatory properties are author(s), publication date, main title, unique persistent global identifier (e.g. DOI), next publication date (if published regularly), periodicity, variable index (classification), short and long abstract, physical form (i.e. file format(s)), and copyright (i.e. legal information). Green's proposed standard is hence similar to the DataCite Metadata Schema (see DataCite Metadata Working Group, 2014). However, Green's mandatory information exceeds what is necessary for a proper data citation; furthermore, information on publisher or distributor cannot be added. Nevertheless, OECD's single platform solution OECD iLibrary (http://www.oecd-ilibrary.org/) for offering publications, (related) research data, as well as a citable DOI can be named a best practice example not only for libraries.

Institutional change

Starting with a 2011 international workshop on developing data attribution and citation practices and standards (Uhlir, 2012), a number of initiatives on data citation have emerged (e.g. CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013). They deal with data citation practices, the supporting infrastructure and challenges to the implementation of good data citation practices. Leading data archives in the social sciences have also introduced guiding principles for implementing citation of their data (Inter-university Consortium for Political and Social Research (ICPSR), GESIS–Leibniz-Institute for the Social Sciences, UK Data Archive, Australian National Data Service (ANDS), etc.). Recent guides include, among others, Corti, Van den Eynden, Bishop and Woollard (2014) with examples from UK Data Archive or ZBW, GESIS and RatSWD (2015) for social science and economic data citation. Corti et al. (2014) recommend:

> A citation for a data collection should include enough information so that the exact version of the data being cited can be located, but does not include information on the sponsor or copyright ownership. Any acknowledgement [...] should not be a replacement for a proper citation. (p. 207)

The UK Data Archive citation corresponds to a large extent to the data citation guidelines of Dodd (1979), thereby being quite comprehensive. To facilitate data citation, unique identification, as well as localization UK Data Archive, ANDS, and other data archives use persistent identifiers like DOIs.

Scientific journals like the *American Sociological Review* or the *American Economic Review* have started to adopt and promote data citation and archiving policies, a shift that has been well received by the scientific community (Smit, 2011). But further support is needed by the publishers. "Some journals have implemented data archives for their respective journals, but data availability is often not enforced. Also an overall infrastructure for publication-related research data is currently not yet available at specialized data centres" (Vlaeminck, 2013,

p. 17). It seems that publishers in general may not see the need to implement data archives for their journals. Anita De Waard (2012) sums it up, "Overall, commercial publishers are not interested in owning or charging for research data or running repositories. There might be exceptions, but in general, this is the case" (p. 158). Publishers have a critical role when it comes to data citation and need to provide guidelines for copy editors (Hole, 2014). A promising approach of one of the publishing houses seems to be the service *elsevierconnect* (http://www.elsevier.com/connect/building-a-global-infrastructure-for-research-data) provided by Elsevier. Besides general guidelines for data citation, Elsevier supports data sharing on various levels and is developing tools for storing, discovering, and reusing research data.

Data policies of e.g. research funders, research organizations, and universities can also address the topic of research data citation. Unfortunately, there are as yet few mentions of data citation, and requirements are missing completely (see Figure 1, following page). This can be explained by the not yet established research data culture. However, we recommend to add an obligation to cite research data within data policies comparable to the obligation to cite publications. Libraries can support the establishment of a research data citation culture within their library catalogues and with citation workshops.

Research data citation with persistent identifiers

Few studies and projects have addressed challenges and best practices of research data citation with persistent identifiers like DOIs, as yet. Natasha Simons (2012) is an exception. She comments on problems like versioning and granularity, but offers no solution. These problems are reinforced by experiences of ANDS and Griffith University (Simons, Visser & Searle, 2013). Micah Altman and Mercè Crosas (2013) conclude that challenges of provenance (chain of ownership, history of transformation), identity (equivalence, versioning and granularity), and attribution (citizen science, many contributors) for data citation remain. First approaches for solution are named, e.g. regarding dynamic datasets. The citation of dynamic datasets is an important aspect, which Huber et al. (2015) also address. They suggest different ways for circumvention.

SERVICE DESCRIPTION

da|ra operates as an allocation agency for DataCite (http://www.datacite.org), an international consortium pursuing the goal of supporting the acceptance of research data as independent citable scientific objects (see Brase, 2010; Brase & Farquhar, 2011). A DOI allocation agency is a member within the DataCite initiative, which allocates DOI names to data objects provided by the so called publication agents (data archives, repositories,

Data Policy	Recommendations regarding research data citation
Organisation for Economic Co-operation and Development (2007)	Research institutions and professional associations should develop appropriate practices with respect to the citations of data and the recording of citations in indexes, as these are important indicators of data quality. (p. 20)
European Science Foundation (2011)	All primary and secondary data should be stored in secure and accessible form, documented and archived for a substantial period. It should be placed at the disposal of colleagues. The freedom of researchers to work with and talk to others should be guaranteed. (p. 6) Intellectual contributions of others should be acknowledged and correctly cited. (p. 7)
Deutsche Forschungsgemeinschaft (2013)	Primary data as the basis for publications shall be securely stored for ten years in a durable form in the institution of their origin. (p. 74)
National Science Foundation (2014b)	Plans for data management and sharing of the products of research, including preservation, documentation, and sharing of data, samples, physical collections, curriculum materials and other related research and education products should be described in the Special Information and Supplementary Documentation section of the proposal […]. (p. II-10)
Universität Bielefeld (2013)	Across the entire data lifecycle—from data collection to publication—research data should be handled and documented diligently and according to appropriate subject-specific standards. (p. 1)
Georg-August-Universität Göttingen (2014)	Management of research data includes their planning, collection, processing, and preservation. It ensures the access to, and the reuse, reproducibility, and quality assurance of all research data underpinning research results. (p. 1) Research data management is generally the responsibility of the person leading a project and the researcher who is acting in an individual capacity. A particular responsibility is the adherence to good practices of research as well as standards in their subject area. (p. 1)
Humboldt-Universität zu Berlin (2014)	All HU researchers are encouraged to process research data resulting from their research activities according to the conventions and standards of their respective scientific community. They should document the complete research lifecycle including tools and procedures that they used. (p. 1)
Universität Heidelberg (2014)	Die Verantwortlichkeit für den Lebenszyklus von Forschungsdaten, insbesondere die Sicherstellung und Bereitstellung der Forschungsdaten zur langfristigen Archivierung liegt primär beim Projektverantwortlichen (PI). [no English translation available**] (p. 1)

** Can be translated as "The project leader is primarily responsible for the life-cycle of research data, in particular for securing and availability of research data for long-term archiving."

Figure 1. Research data policies and citation of research data.

libraries, etc.). In turn DataCite is a member of the International DOI Foundation (IDF) and one of nine registration agencies (i.e. also CrossRef) which have the authority to allocate DOI names. It is the only agency that focuses on identifiers for research data.

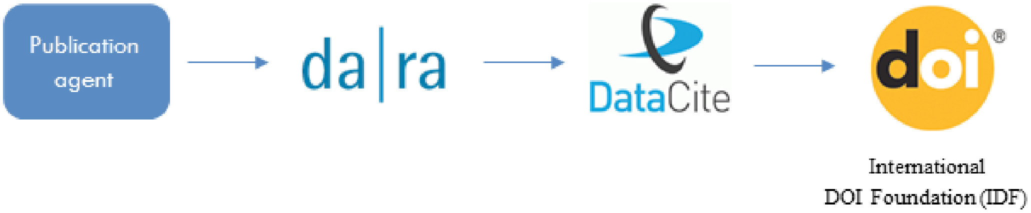


Figure 2. Relations da|ra – DataCite.

In line with the ideals of good scientific practice, there is a demand for open access to existing primary data so as to not only have the final research results but also to be able to reconstruct the entire research process. GESIS and ZBW therefore offer a registration service for social and economic research data. By registration, updated and structured metadata is assigned to the digital object using the DOI name. This infrastructure lays the foundation for long-term, persistent identification, storage, localization, and reliable citation of research data in Germany and beyond.

The DOI registration service is available to various data providers in the social sciences and economics, among them data archives, research data centers, libraries, and data repositories.[2] A growing number of university libraries are assigning DOI names to their data holdings from different research communities. A case in point is the newly established HeiDATA Dataverse Network (https://heidata.uni-heidelberg.de/dvn/) hosted by the University Library Heidelberg. This institutional repository is managed by the Competence Centre for Research Data, a joint institution of the University Library and the Computing Centre and is a client of da|ra with currently 45 DOIs registered for datasets and more than 10,000 DOIs for texts. HeiDATA serves as a central infrastructure for the data producing researchers at the University of Heidelberg. It supports the researchers in the process of generating metadata for their data as well as enables them to link the data to the related publications available in the libraries' open access repository.

[2] See list of publication agents at http://www.da-ra.de/en/about-us/our-users/

Criteria for research data registration

Four criteria are especially important for research data registration. First, one must have the *authority* to register the research data. If one does not own the data, the creator or the data owner must grant permission to register the data. Second, there has to be a guarantee for data *persistence*. A long-term commitment to maintain the data in a usable and accessible form needs to be made. At least ten years of access have to be guaranteed, which comply with the rules of good scientific practice in Germany (Deutsche Forschungsgemeinschaft, 2013, p. 74).[3] Particularly after the end of a research project, this has to be ensured and clarified. Therefore, institutional solutions are preferable to personal commitments. Third, research data should be *accessible* to external users. Not all research data can or need to be published and registered. Copyright issues or the sensitivity of the data could impede publication, and it may be necessary to impose embargoes or access restrictions. Nevertheless, research data restricted to specified user groups can also be registered. In such cases, information about how to obtain and use the data should be made clear at the point of access (i.e. landing page). The publication of research data aims at making research data citable as well as re-usable for secondary research. Therefore, the registration of research data is not suitable for data that is completely unavailable to users outside of an institution. The decision for or against research data publication and registration lies with the researcher or research group. Fourth, research data aimed for publication and registration should have *citation potential*. It has to be of likely interest to other researchers/users and may be cited in future works.

Furthermore, the five Joint Declaration of Data Citation Principles (JDDCP) should be taken into account (Data Citation Synthesis Group, 2014). These principles "cover purpose, function and attributes of citations" of research data (Data Citation Synthesis Group, 2014). In addition to the previously mentioned criteria *access* and *importance*, research data should be cited that represents *evidence* and gives *credit* and *attribution* to the researchers in a sufficient manner and with *unique identification* like a DOI or another persistent identifier (see Tonkin, 2008).

Data, metadata, citation

The da|ra Metadata Schema (Helbig et al., 2014b) for the registration of resources with da|ra

[3] The recommendations in US data management policies vary between three to seven years of documentation and archiving. The National Science Foundation (NSF) only recommends the sharing of scientific data "within a reasonable time", but it does not name a specific time span for long-time archiving (National Science Foundation, 2014a). The European Science Foundation (ESF) expects that "original scientific or scholarly research data should be documented and archived for a substantial period (at least 5 years, and preferably 10 years)" (European Science Foundation, 2011, p. 13).

is based on the DataCite Metadata Schema (DataCite Metadata Working Group, 2014). However, it has certain value added metadata properties to meet the special demands of social science and economic research data. These properties encompass, among others, the mandatory property "General Resource Type" (see below), *Journal of Economic Literature* (JEL)[4] and GESIS[5] classifications, as well as discipline-specific thesauri like e.g. *Thesaurus for the Social Sciences*[6] and *Standard Thesaurus for Economics* (STW).[7]

There are six mandatory metadata properties in the current da|ra Metadata Schema (doi:10.4232/10.mdsdoc.3.1) that have to be provided for the registration of a DOI name via da|ra (see Figure 3).

Property	**Definition**
General Resource Type	The general type of a resource (collection, dataset, text, video, image, audio or interactive resource).
Title	The title of a resource.
Creator	The name of the principal investigator or author. May be a corporate/institutional or a personal name.
URL	Each DOI name has a URL to which it resolves (landing page).
Publication Date	The publication date of the resource submitted by the publication agent; format sub-properties date, monthyear or year are possible.
Availability (controlled)	Conditions governing access to primary resource.

Figure 3. Mandatory da|ra metadata properties.

In addition, optional and recommended properties can be added. For all metadata properties the respective names, definitions, attributes, conditions, cardinality (maximum occurrence), and value domains are defined. Some properties comply with International Organization for Standardization (ISO) norms. These norms determine, for example, which

[4] JEL Classification was developed for the *Journal of Economic Literature* and classifies articles, books and other economic literature (American Economic Association, n.d.).

[5] GESIS Classification for the Social Sciences is offered for classifying and retrieval of social science literature and research projects (GESIS, 2013).

[6] *Thesaurus for the Social Sciences* (TheSoz) is a list of keywords for the social sciences (GESIS, 2014).

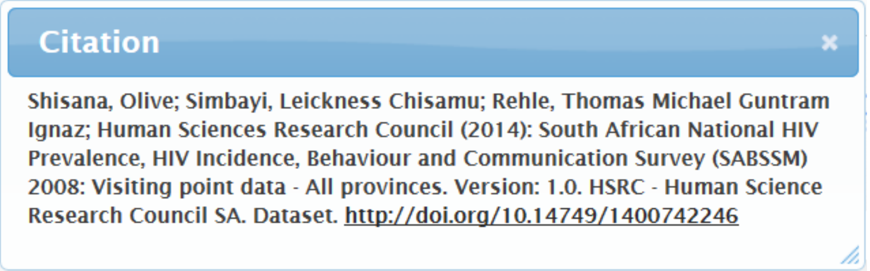
[7] *STW Thesaurus for Economics* provides keywords and standardized vocabulary on economic subjects (ZBW, 2015).

code for a language or geographic coverage has to be applied (e.g. ISO 639-2, ISO 3166-2/3). Identifiers like Virtual International Authority File (VIAF, http://viaf.org/), Open Researcher and Contributor ID (ORCID, http://orcid.org/), or German Gemeinsame Normdatei (GND, English: Universal Authority File) are used for the unique identification of persons and institutions. As mentioned before, controlled vocabularies such as thesauri and classifications are applicable. These vocabularies are complemented by da|ra controlled terms. Once registered, the mandatory metadata should not be changed. Therefore, it is essential that this metadata is free of errors.

On the basis of da|ra Metadata Schema (doi:10.4232/10.mdsdoc.3.1), da|ra recommends the following citation style for research data, which is also applied in its own information system: *Creator (Publication date): Title. Version. Publication agent. General resource type. Identifier*

Property	**Definition**
Creator	The creators, i.e. primary researchers or authors are displayed in the order in which they were added. Only the name is used for people, not their affiliation. If only an institution is specified as the creator, that is used. The data about the creators can include up to five names; beyond that they are abbreviated with 'et al.'.
Publication date	The publication date is the year in which the resource was published in digital form by the publication agent.
Title	Here, the main title of the resource and any additional title must be indicated. Additional titles include: Other titles: All subtitles must be stated in the citation. Main and subtitles are separated with ".". Alternative and translated as well as original titles are not quoted in the citation. Collective title: A collective title, if existent, is also specified in the citation.
Version	The version represents the version number of the resource.
Publication agent	The publication agent is the name of the institution or of the data centre which published the resource.
General resource type	The general resource type describes the general type of the resource. The resource types include collection, dataset, text, video, image, audio and interactive resource.
Identifier	The identifier is the Digital Object Identifier (DOI) or any other (persistent) identifier available. It is displayed in the citation as a URL link including the necessary resolver and without "doi:".

Figure 4. Definition of da|ra citation elements.



Shisana, Olive; Simbayi, Leickness Chisamu; Rehle, Thomas Michael Guntram Ignaz; Human Sciences Research Council (2014): South African National HIV Prevalence, HIV Incidence, Behaviour and Communication Survey (SABSSM) 2008: Visiting point data - All provinces. Version: 1.0. HSRC - Human Science Research Council SA. Dataset. http://doi.org/10.14749/1400742246

Figure 5. Example of da|ra recommended citation.

The citation style is complementary to the recommendations of DataCite (DataCite Metadata Working Group, 2014, p. 7; Starr & Gastl, 2011) as well as FORCE11 (https://www.force11.org/node/4771).

Pitfalls and issues of research data registration

Some areas of research data registration are problematic or there are still no best practice solutions available. These areas include versioning, granularity, assurance of metadata quality, handling of dynamic datasets, and the use of missing values within research data citation. Versioning has three important aspects (Helbig & Hausstein, 2014a, p. 14). First, a DOI name should not be changed (International DOI Foundation, 2015). Though not a requirement of DataCite, da|ra strongly advises to avoid a change in the registered object, too. Once registered, the object has to remain as it is to avoid misquotation. Second, each change must be saved as a new version and a new DOI name must be assigned. Revisions, updates, or expansions always require a new version (Dodd, 1979; Simons, 2012). Third, the responsibility for versioning rests with the publisher or publication agent. This is also contractually guaranteed. The second aspect can be especially challenging for publishers of research data. Corrections of typos or other types of errors require a new version, which not everybody is willing or able to prepare and register. At the beginning, it is useful to define milestone versions of the resource and to store those separately. Changes in these milestone versions should be documented and saved as a new version. Links to prior versions or the use of related identifiers are recommended (Altman & King, 2007; Starr & Gastl, 2011). There are various ways to formulate versioning. However, unity and a systematic approach are essential.

Our recommendations regarding versioning are (three-digit version number X.X.X):

- Increase of the first digit if new data is added (e.g. waves, samples, etc.)

- Change of the second digit if corrections are made which influence the analysis (e.g. change of values of respondents)
- If the documentation is changed or amended (typing error or more detailed text added, etc.), only the third digit will be increased

This three-digit versioning concept is used, for example, by GESIS data archive for their data catalogue DBK (https://dbk.gesis.org/dbksearch/). In contrast, the UK Data Archive works with both editions and versions, thereby complicating citation. ICPSR as well as UK Data Archive apply one-digit versioning that impedes an assessment with one look. Rather, a study of the documentation is necessary to trace changes.

Granularity describes the degree of aggregation of the resource to be registered. Depending on discipline or resource, different levels of granularity can be applicable. DOI names can be assigned at any desired degree of precision and granularity that a publisher deems to be appropriate (Simons, 2012). For the decision on granularity, we recommend consideration of the following points (see also Helbig et al., 2014b):

1. Current citation and research practices among the user community (Citation)
2. Needs of various stakeholders (Use of data)
3. Complexity and extent of research data (Type of resource)
4. Maintenance of metadata (Sustainability)

"Functional granularity" should be applied; the deepest granular level that makes sense in terms of cost and value to the publisher should be chosen (Rust & Bide, 2000, p. 10). A balance has to be found between metadata creation and management demands and the value of data citation (CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013).

Metadata quality is of great importance because it ensures that research data can be found and used for secondary analysis as well as cited and indexed in terms of good scientific practice. Criteria for the quality of metadata are accuracy, completeness, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility (Bruce & Hillmann, 2004). Metadata should therefore be as thoroughly documented as possible. Particularly recommended metadata properties should be specified, if applicable (see Helbig et al., 2014b). Within the social sciences this applies, for example, to thesauri and controlled vocabularies specific to the discipline.

The unambiguous and persistent identification of *dynamic datasets* like open time series data or databases can become a great challenge (Green, 2009, p. 13). Quite a few records change on a regular basis or are revised within a periodical time span. This can lead to a high frequency of data publication and registration processes. There are three ways to meet this challenge. First, one can use consequent versioning of and linking between datasets associated with high quality metadata that describes changes and characteristics of the particular editions (Green, 2009, p. 14; Huber et al., 2015, p. 9). The second method involves the registration of a parent data project using the citation to state the access time point (see example (1) and (2) illustrated in Figure 6). Third, one can define milestone versions and register snapshots of the dynamic dataset to achieve persistent identification (Huber et al., 2015, p. 10). Fourth, specific time slices can be defined (see example (3) illustrated in Figure 6).

(1) Database	Hoffmann, Elke; German Centre for Gerontology, DZA (2013): Statistical Information System GeroStat. Version: 1. DZA – German Centre for Gerontology. Collection. http://doi.org/10.5156/GEROSTAT [German Ageing Survey (DEAS) - 1996, 2002, 2008; Date of retrieval: 04.05.2015]
(2) Website	IANUS - Research Data Centre for Archaeology & Classical Studies (2014): IT recommendations for sustainable digital data handling in the classical studies. Version: 1. IANUS - RDC Archaeology & Classical Studies. Interactive Resource. http://doi.org/10.13149/000.111000-a [Date of retrieval: 04.05.2015]
(3) Video	Deutsches Zentrum für Luft- und Raumfahrt – DLR (2011): Crash tests for the car of the future. DLR inaugurates dynamic component testing facility, video segment. http://doi.org/10.5446/12780#t=00:20,00:27

Figure 6. Citation examples for dynamic datasets.

Missing information is to be avoided when publishing research data. However, research data can, for example, be very old, and the exact publication date can be unknown or creatorship can be unclear. Especially archaeological and historical research data are affected. DataCite offers a list of standard values for unknown information in order to enable the registration process (DataCite Metadata Working Group, 2014, p. 38). Missing values still remain an issue for the registration that needs to be discussed and that one needs to be aware of.

Lessons learned and conclusion

The reactions of the publication agents to our best practice recommendations are quite diverse. Versioning is adopted by publication agents in terms of two digits (X.X) at most. An obstacle is clearly that new versions should require a new landing page, which is not desired or cannot be maintained by the publication agents. Granularity can occasionally be more a decision of convenience than a sound one. Circumvention of deeper DOI granularity can include the use of fingerprints,[8] for instance. These do not require additional metadata or a landing page. We try to give advice on granularity, but lastly it is a decision of the researchers or research group. Yet most decide what is best for the potential users. The da|ra Metadata Schema was especially well received by the social sciences and economics community. The outcome is publication agents beyond Germany that prefer the da|ra service to their local or other registration agencies.

NEXT STEPS

A unique identifier like a DOI is a prerequisite for citing research data in a suitable way and a first step forward to giving "credit where credit is due" (Nature, 2009, p. 825). Obviously, a DOI alone is not enough to promote data sharing. At least in the social sciences, what is needed is a cultural change in publication and citation behavior. This "cultural change" must come from the scientific community. As infrastructure providers we can only support this process. As Tenopir et al. (2011) rightly point out, "Building a sound infrastructure for data sharing, preservation, and use is a challenge, but is in some ways easier than changing a culture" (p. 20). First libraries in Germany and beyond act as strategic multipliers and offer training as well as information about research data management and data citation. The teaching of information literacy is one of the core competencies of libraries and should be extended to research data.

The future challenges for research data publishers and infrastructure providers is evidently the support of *interoperability* between different systems. Amongst others, the Research Data Alliance (RDA) group "Data Description Registry Interoperability" (DDRI) was formed to tackle the problem of interoperability. da|ra is a member of this group and works together with the Data Preservation Alliance for the Social Sciences (Data-PASS) on the exchange of metadata based on the Data Documentation Initiative (DDI) schema (see Data Documentation Initiative, 2014). Libraries can foster interoperability by actively taking part in the discussion and adding their valuable points of view regarding metadata exchange formats. University libraries especially will presumably include research data in their catalogues in the future and should engage more in refining metadata schemata and spreading standards for data citation.

Research data repositories and research data centers are emerging with increasing frequency. More and more universities, departments, libraries, and other institutions want to offer their

[8] Electronic fingerprints or hash values serve as identifiers for electronic resources. Common methods for fingerprinting are MD5 (Message Digest Algorithm 5) and SHA-1 (Secure Hash Algorithm).

produced or stored research data via their own repositories or services. This heterogeneity of services is, on the one hand, a huge benefit for researchers as more and more research data becomes available. On the other hand, the researcher has no central catalogue for searching for apt research data. Building a direct point of access to metadata of social science and economic research data is therefore the aim of da|raSearchNet, the follow-up project to da|ra. da|raSearchNet will enable users to easily search up-to-date references of data holdings on an international basis.

ACKNOWLEDGEMENTS

The authors wish to thank Derek Kruse (ZBW) for proofreading the manuscript.

REFERENCES

Altman, M., & Crosas, M. (2013). The evolution of data aitation: From principles to implementation. *IASSIST Quarterly*, *37*(1-4), 62-70. Retrieved from http://www.iassistdata.org/downloads/iqvol371_4_altman.pdf

Altman, M., & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine*, *13*(3/4). http://dx.doi.org/10.1045/march2007-altman

American Economic Association. (n.d.). J*EL Classification System*. Retrieved from https://www.aeaweb.org/econlit/jelCodes.php

Brase, J. (2010). Datacite: A global registration agency for research data. *Working Paper Series of German Data Forum* (RatSWD) No. 149. Retrieved from http://hdl.handle.net/10419/43624

Brase, J., & Farquhar, A. (2011). Access to research data. *D-Lib Magazine*, *17*(1/2). http://dx.doi.org/10.1045/january2011-brase

Brase, J., Sens, I., & Lautenschlager, M. (2015). The tenth anniversary of assigning DOI names to scientific data and a five year history of DataCite. *D-Lib Magazine*, *21*(1/2). http://dx.doi.org/10.1045/january2015-brase

Bruce, T. R. & Hillmann, D. I. (2004). The continuum of metadata quality: defining, expressing, exploiting. In D. I. Hillmann & E. L. Westbrooks (Eds.), *Metadata in practice* (pp. 238-256). Chicago, IL: American Library Association.

CODATA-ICSTI Task Group on Data Citation Standards and Practices. (2013). Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal*, *12*, CIDR1-CIDR75. http://dx.doi.org/10.2481/dsj.OSOM13-043

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). *Managing and sharing research data: A guide to good practice*. Los Angeles, CA: SAGE.

Data Citation Synthesis Group. (2014). J*oint Declaration of Data Citation Principles*. M. Martone (Ed.). San Diego, CA: FORCE11. Retrieved from https://www.force11.org/datacitation

DataCite Metadata Working Group. (2014). *DataCite metadata schema for the publication and citation of research data v3.1*. http://dx.doi.org/10.5438/0010

Data Documentation Initiative. (2014). *DDI-Lifecycle 3.2*. Retrieved from http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/#3.2schema

Deutsche Forschungsgemeinschaft. (2013). *Proposals for safeguarding good scientific practice: Recommendations of the Commission on Professional Self Regulation in Science. Memorandum*. WILEY-VCH: Weinheim. Retrieved from http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf

De Waard, A. (2012). Linking data to publications: Towards the execution of papers. In P. Uhlir (Ed.), *For attribution - Developing data attribution and citation practices and standards: Summary of an international workshop* (pp.157-159). Washington, DC: The National Academies Press. Retrieved from https://download.nap.edu/catalog.php?record_id=13564

Dodd, S. A. (1979). Bibliographic references for numeric social science data files: Suggested guidelines. *Journal of the American Society for Information Science*, *30*(2), 77-82. http://dx.doi.org/10.1002/asi.4630300203

European Science Foundation. (2011). *The European Code of Conduct for Research Integrity*. Strasbourg: Ireg. Retrieved from http://www.esf.org/fileadmin/Public_documents/Publications/Code_Conduct_ResearchIntegrity.pdf

Fecher, B., Friesike, S., Hebing, M., Linek, S., & Sauermann, A. (2015). A reputation economy: Results from an empirical survey on academic data sharing. *Working Paper Series of German Data Forum* (RatSWD) No. 246. Retrieved from http://www.ratswd.de/dl/RatSWD_WP_246.pdf

Georg-August-Universität Göttingen. (2014). *Research data policy of the Georg-August University Goettingen (incl. UMG)*. Retrieved from http://www.uni-goettingen.de/en/488918.html

German Data Forum. (2010). Recommendations for expanding the research infrastructure for the social, economic, and behavioral sciences. *Working Paper Series of German Data Forum* (RatSWD) No. 150. Retrieved from http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_150.pdf

GESIS. (2013). *Classification for the social sciences*. July 2013 version. Retrieved from http://www.gesis.org/fileadmin/upload/dienstleistung/tools_standards/KassifikationSozialwissenschaften_Stand_Juli_2013_dt_en__2_.pdf

GESIS. (2014). *Thesaurus for the social sciences*. Retrieved from http://www.gesis.org/en/services/research/thesauri-und-klassifikationen/social-science-thesaurus/

Green, T. (2009). We need publishing standards for datasets and data tables. *OECD Publishing White Papers*, OECD Publishing. http://dx.doi.org/10.1787/787355886123

Hamermesh, D. S. (2013). Six decades of top economics publishing: Who and how? *Journal of Economic Literature*, *51*(1), 162-172. http://dx.doi.org/10.1257/jel.51.1.162

Helbig, K. & Hausstein, B. (2014a). *Best practice guide for the registration of resources with da|ra*. Version: 1.0. GESIS - Technical Reports 2014/18. http://dx.doi.org/10.4232/10.bpg.1.0

Helbig, K., Hausstein, B., Koch, U., Meichsner, J., & Kempf, A. O. (2014b). *da|ra Metadata Schema*. Version: 3.1. GESIS - Technical Reports 2014/17. http://dx.doi.org/10.4232/10.mdsdoc.3.1

Hole, B. (2014). *Data citation: A critical role for publishers* [SlideShare slides]. Retrieved from http://de.slideshare.net/brianhole/data-citation-a-critical-role-for-publishers

Huber, R., Asmi, A., Buck, J., Lucas, J. M. de, Diepenbroek, M., Michelini, A., & participants of the joint COOPEUS/ENVRI/EUDAT PID workshop. (2015). *Data citation and digital identifiers for time series data. Environmental research infrastructures*. http://dx.doi.org/10.6084/m9.figshare.1285728

Humboldt-Universität zu Berlin. (2014). *Humboldt-Universität zu Berlin Research Data Management Policy*. Retrieved from https://www.cms.hu-berlin.de/en/ueberblick-en/projekte-en/dataman-en/info/policy

International DOI Foundation. (2015). DOI prefix. In *DOI-Handbook* (Chapter 2.2.2). http://dx.doi.org/10.1000/182

International Organization for Standardization. (2010). I*SO 639-2:1998: Codes for the representation of names of languages. Part 2: Alpha-3 code*. Retrieved from http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=4767

International Organization for Standardization. (2013a). I*SO 3166-2:2013: Codes for the representation of names of countries and their subdivisions. Part 2: Country subdivision code*. Retrieved from http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=63546

International Organization for Standardization. (2013b). *ISO 3166-3:2013: Codes for the representation of names of countries and their subdivisions. Part 3: Code for formerly used names of countries*. Retrieved from http://www.iso.org/iso/catalogue_detail?csnumber=63547

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, *331*(6018), 719-721. http://dx.doi.org/10.1126/science.1197872

Ludwig, J. & Enke, H. (Eds.). (2013). *Leitfaden zum Forschungsdaten-Management: Handreichung aus dem WissGrid-Projekt*. Glückstadt: Hülsbusch. Retrieved from http://www.wissgrid.de/publikationen/Leitfaden_Data-Management-WissGrid.pdf

Mooney, H. & Newton, M. P. (2012). The anatomy of a data citation: Discovery, reuse, and credit. *Journal of Librarianship and Scholarly Communication*, *1*(1), eP1035. http://dx.doi.org/10.7710/2162-3309.1035

National Science Foundation. (2014a). Dissemination and sharing of research results. In *Proposal and award policies and procedures guide. Part II - Award and administration guide (Chapter VI)*. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/aag_print.pdf

National Science Foundation. (2014b). Proposal preparation instructions. In *Proposal and award policies and procedures guide. Part I - Grant Proposal Guide (Chapter II)*. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_print.pdf

Nature (2009). Credit where credit is due. Editorial. *Nature*, *462*(7275), 825. http://dx.doi.org/10.1038/462825a

Organisation for Economic Co-operation and Development. (2007). *OECD principles and guidelines for access to research data from public funding*. Retrieved from http://www.oecd.org/science/sci-tech/38500813.pdf

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, 1, e175. http://dx.doi.org/10.7717/peerj.175

Rust, G., & Bide, M. (2000). *The <indecs> metadata framework. Principles, model and data dictionary.* WP1a-006-2.0. Retrieved from http://www.doi.org/topics/indecs/indecs_framework_2000.pdf

Sieber, J. E., & Trumbo, B. E. (1995). (Not) giving credit where credit is due: Citation of data sets. *Science and Engineering Ethics*, *1*(1), 11-20. http://dx.doi.org/10.1007/BF02628694

Simons, N. (2012). Implementing DOIs for research data. *D-Lib Magazine*, *18*(5/6). http://dx.doi.org/10.1045/may2012-simons

Simons, N., Visser, K., & Searle, S. (2013). Growing institutional support for data citation. Results of a partnership between Griffith University and the Australian National Data Service. *D-Lib Magazine*, *19*(11/12). http://dx.doi.org/10.1045/november2013-simons

Smit, E. (2011). Abelard and Héloise: Why data and publications belong together. *D-Lib Magazine*, *17*(1/2). http://dx.doi.org/10.1045/january2011-smit

Starr, J. & Gastl, A. (2011). isCitedBy: A metadata scheme for DataCite. *D-Lib Magazine*, *17*(1/2). http://dx.doi.org/10.1045/january2011-starr

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M. & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, *6*(6), e21101. http://dx.doi.org/10.1371/journal.pone.0021101

Tonkin, E. (2008). Persistent identifiers: Considering the options. *Ariadne*, *56*. Retrieved from http://www.ariadne.ac.uk/issue56/tonkin/

Uhlir, P. F. (Ed.). (2012). *For attribution: Developing data attribution and citation practices and standards: Summary of an international workshop*. Washington, D.C: National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=13564

Universität Bielefeld. (2013). *Principles and guidelines on handling research data at Bielefeld University*. Retrieved from https://data.uni-bielefeld.de/policy

Universität Heidelberg. (2014). *Research data policy. Richtlinien für das Management von Forschungsdaten*. Retrieved from http://www.uni-heidelberg.de/universitaet/profil/researchdata/

Vlaeminck, S. (2013). Data management in scholarly journals and possible roles for libraries – Some insights from EDAWAX. *Liber Quarterly*, *23*(1), 48-79. Retrieved from http://nbn-resolving.de/urn:NBN:NL:UI:10-1-114595

ZBW. (2015). *STW Thesaurus for Economics*. Version 9.0. Retrieved from http://zbw.eu/stw/versions/latest/about.en.html

ZBW, GESIS, & RatSWD (Eds.). (2015). *Auffinden - Zitieren - Dokumentieren: Forschungsdaten in den Sozial - und Wirtschaftswissenschaften*. http://dx.doi.org/10.4232/10.fisuzida2015.2