Sustainable Data Preservation using datorium – facilitating a scholarly Ideal of Data Sharing in the Social Sciences

Monika Linne
GESIS - Leibniz Institute for the Social Sciences
Data Archive for the Social Sciences
Unter Sachsenhausen 6-8
50667 Cologne, Germany
+49 221 47694 - 452
monika.linne@gesis.org

ABSTRACT
This paper introduces datorium - a digital data preservation project at the Data Archive of GESIS-Leibniz Institute for the Social Sciences. datorium is a new data repository service for the research community. It functions as a web-based data sharing repository providing a user-friendly tool for researchers making data accessible for the purpose of re-use by other scholars. Sharing, managing, documenting and publishing data, structured metadata and publications will be carried out autonomously by researchers. Data and related information will be available free of charge. All uploaded research data and documentation will be peer-reviewed and digitally preserved by the GESIS Data Archive.

GESIS promotes data sharing as a scholarly ideal and facilitates cooperation between researchers. By developing datorium the Data Archive aims to collect and provide research data with a wide thematic scope for academic re-use. A further intention is to ensure long-term preservation of archived data and metadata as well as providing wide-ranging dissemination possibilities for scholars in order to increase the visibility and availability of their research projects. By providing access to their research data scholars can support new research or secondary analysis and beyond that they profit from increased citations of their work, thereby improving their professional reputation.

Categories and Subject Descriptors
H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *data sharing, web-based service.*

General Terms
Management, Documentation, Standardization.

Keywords
datorium, Social Science Research Data, Research Data Management, Data Repository, Digital Preservation, Data Sharing, Data Archiving.

1. INTRODUCTION
Organizational modification of structural frameworks is demanded thanks to accelerating and fundamental changes within academia and the potential provided by professional information management [1]. By developing a digital data sharing repository GESIS responds to this changing data landscape, where researchers call for flexible ways of distributing and re-using research data. For this reason GESIS is expanding its range of services by offering datorium: a digital data dissemination tool that allows for prompt publishing and sharing of research data with other scholars. In addition, datorium can operate as a working environment that can jointly be used by a research group in order to work together on the documentation of a research project and the publication of related findings.

One goal is expanding the variety of research data types that comes along with a wider thematic collection for data preserved at the Data Archive. With datorium the culture of data sharing, supported and promoted by the Data Archive over the past 50 years, will be pushed forward and facilitated by the re-use of archived data.

A priority of the GESIS Data Archive is to ensure data and metadata provided by the archive is of high quality. Accordingly, all material in datorium is peer-reviewed against defined quality criteria before it can be shared and made available to other scholars.

2. BACKGROUND
Since "*data sharing is essential for all verifications and all secondary analyses*" [2, p.9] producing metadata and sharing research data with other scholars ought to be taken for granted. However, transparency and accessibility of research data is still not common in the social sciences. Archiving and publishing research data in the social sciences is still the exception rather than the rule. This is especially the case for smaller research projects where data is merely a basis for publications not an output in itself. Beyond that, data sets are usually not professionally archived or available to the research community [3, 4, 5].

For this reason datorium focuses on standardized documentation and digital preservation of smaller projects to enhance the

discoverability, accessibility and the reuse of their data, opposed to bigger, national or international survey programs that are well documented, published, and intensively re-used by other scholars. In small research projects the budget often does not cover the cost of archiving and publishing data and metadata. This is a severe problem, as the value of research data is usually not exhausted after the initial research findings have been published [5].

Potential re-use of the data should be considered as part of every research project in order to facilitate further data analysis, such as secondary analysis, reanalysis, replication analysis, verification of research findings [3] or meta-analysis. However, findings of any empirical analysis can only be evaluated if full documentation of the research processes is provided [6]. King (1995) rightly points out, that "*the replication standard holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author*" [6, p.444].

In Germany replication and evaluation of research findings from the social sciences is difficult because most publications provide neither data sets nor the syntax used for analysis [7]. This is down to insufficient data management practices in the research routine [8]. Vardigan et al. (2008) specify the problem: *"because good documentation is paramount to effective data use, data archives have long encouraged data producers to document their data thoroughly, starting at the very beginning of a research project and in effect creating an audit trail of all variable transformations that take place over the life of the project. In reality, there is little incentive for data producers to follow these guidelines and documentation is often hastily assembled just before deposit into an archive. Furthermore, documentation is most often produced with word processing software and then rendered into PDF, making reuse difficult"* [9, pp.108-109].

Taking these issues into account, the necessity of developing a data-sharing repository including standardized data documentation is obvious. As Vardigan et al. put it, *"for a secondary analyst to understand a given dataset, he or she must have access to good documentation"* [9, p.108]. Providing this is an essential part of the GESIS Data Archive efforts. datorium will promote this ideal by offering an accessible and user-friendly repository tool.

3. THE PROJECT DATORIUM

Consequently, increasing and intense discussion about Open Access Publication and an attitude shift towards data sharing has emerged. Therefore systems and infrastructures have to be built in order to meet these requirements [1]. In Germany GESIS, in cooperation with other research institutions, have established digital research data initiatives, e.g., the Social Science Open Access Repository (SSOAR), the Social Science Portal SOWIPORT, the social science Literature Information System SOLIS and the social science Research Information System SOFIS.

However, a gap has existed until now for a user-friendly repository which can be managed autonomously by researchers combining upload of data sets with corresponding standardized documentation and metadata. By joining the range of digital research initiatives datorium is closing this gap in the portfolio of GESIS services and the German academic landscape.

As a member of the Leibniz association GESIS is jointly financed by the German federal government and states and it pursues exclusively non-profit objectives. Therefore usage of datorium will be free of charge to data providers and data users. Users of datorium will neither be charged for the upload/download of research data or the review carried out by the Data Archive. This lowers the barrier to unfunded research projects making their data available to the broader academic audience, as well increasing its visibility to an interested public and ensuring its long-term preservation.

Usually access to data in repositories or subject specific data centers is limited to a tight thematic or defined institutional user group. datorium will be thematically open for research data from the wide field of the social sciences and does not restrict the service to institutional members. Thus another barrier can be lowered, allowing researchers to easily publish their research data and related findings.

3.1 The metadata scheme

The metadata schema of datorium is a defined list of structured metadata that gives a standardized description of a research dataset to assist the user. Along with the aforementioned reasons for the necessity of data sharing this ensures an easy way to cite and trace research data. The datorium metadata schema is a simplified version of GESIS's data catalogue (DBK) [10], to which metadata schema from da|ra[1] and DataCite[2] have essentially contributed in the development.

datorium's metadata schema uses mandatory core elements where the user must provide a description of a data set. Additionally the user can choose further optional metadata elements for specification. The metadata schema of datorium is compatible with Data Documentation Initiative (DDI 2) codebook standards and metadata schema from da|ra and DataCite. In addition datorium adopts the metadata standard of the Dublin Core Metadata Initiative[3] that provides core metadata vocabularies in support of interoperable solutions for discovering and managing resources.

By meeting international metadata standards datorium addresses the rising demand for a standardized research data management in the social sciences and serves as a helpful tool for researchers to document their research data in order for it to be understood and re-used by other researchers.

---

[1] da|ra is the registration agency for social science and economic data jointly run by GESIS and the German National Library of Economics, Leibniz Information Centre for Economics (ZBW). This infrastructure lays the foundation for long-term, persistent identification, storage, localization and reliable citation of research data [11].

[2] DataCite is an international consortium founded in London in 2009 comprised of sixteen members from ten different countries, to pursue the common goal of supporting the acceptance of research data as independent citable scientific objects through worldwide uniform standards. On the basis of the DOI-system research data is registered with DOI names to enable comprehensive linking of scientific work with the underlying research data [12].

[3] The Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description [13]. It is a relevant metadata standard that is commonly used for the description of research data [14].

3.2 Flexible authorization possibilities

Data re-use leads to a decrease in redundant, repeated, data collection and enables more research with more data in less time at less cost.

In certain cases legitimate concerns over data privacy laws, commercial, or national security exist which prevent uncontrolled re-use and prevent data misuse [1]. Therefore datorium offers depositors full control of data and metadata they supply. Depositors have the option to choose a defined category of data access. This means depositors decide who is authorized to access their data. If users wish to access a data set with restricted accessibility, they can request it by clicking an "Apply-Button". An e-mail is automatically sent to the depositor containing information about the person asking for access. The depositor can login to datorium to permit or to deny access to this user. The GESIS Data Archive receives copies of access-requests for the purpose of documentation. However, this also enables the Data Archive to monitor reactions to requests. If the depositor does not provide an answer to a request within 10 to 20 days, the GESIS Data Archive will contact the depositor to investigate possible problems.

Depositors also have the option of data and metadata being preserved only. Here depositors' benefit from the possibility of having data archived by the GESIS Data Archive without providing wider access. If the depositor later decides to publish data, this is easily done.

In detail the access categories are:

a) Free Access: unrestricted download of research data for all registered users, without having to contact data depositors and request access permission.

b) Restricted Access: users have to apply for permission to download the data by contacting the depositor. The depositor manages data access autonomously.

c) No Access: preservation of data and metadata only without publishing. Publishing at a later time is possible.

3.3 Data review

Data and documentation uploaded to datorium is subject to a review process. This is carried out manually, by a 'curator', working at the Data Archive. The review contains technical controls for file formats, data readability, and is checked for viruses. Furthermore, integrity of data and documentation, completeness, data quality, intellectual property and legal aspects are clarified and verified. The curator carries out additional controls for data consistency such as wild codes, missing values, question routing, and weighting factors, etc. A data set is not published until it fulfills review criteria, assuring high quality data is provided.

In the case of rejection, the GESIS curator contacts the depositor and requests correction of the critical content or additional information needed for publishing. Minimum requirements for publishing are specification of the project's title, principle investigator(s), publication year of the data set, and availability status (access category).

As part of the review process all published research datasets and documentation receive a DOI[4] (Digital Object Identifier) in order to:

a) establish easier access to research data on the Internet

b) increase acceptance of research data as legitimate, citable contributions to the scholarly record

c) *support data archiving that permits results to be verified and re-purposed for future study.* [15]

This DOI is published in conjunction with an automatically generated citation that consists of: [primary investigator] ([Year of Current Version]): [Title]. [Data Collector]. GESIS Datenarchiv, Köln. [Study number] Datenfile Version [Number of Version], [DOI].

3.4 Scholarly Collaboration with datorium

Regardless of academic discipline collaborative research is increasingly common. Most social science fields are heading towards cooperative research endeavors. Particularly in sociology the tendency for scholars to work together in the search for systematic knowledge and the understanding of social phenomena is growing [17].

Taking this growing trend into account, multiple users from different locations, regardless of geographic boundaries, can use datorium as a virtual working environment. Researchers are given the option to invite collaborators into a group. This facility helps documentation of research data. Network or project members can communicate through datorium to discuss working progress and track the latest documentation procedures other colleagues of the research group have worked. Subsequent work on documentation can be organized so the burden for each research member is reduced. With datorium interaction and collaboration between researchers is facilitated and supported. Synergies that may emerge from collaboration between scholars who use datorium as a virtual working environment may lead to fruitful social networking options and further research outputs.

4. DATA PRESERVATION

In the first phase of datorium, data and documentation will not be preserved in datorium itself but in the archival storage system of the GESIS Data Archive, where long-term preservation and access to digital objects is provided through file format migration. The Data Archive of GESIS keeps data 'alive' by *"keeping data safe, comprehensible, and secure from physical damage or technological obsolescence so it is available for re-use or repurposing in contemporary or historical research"*[18]. In order to prevent data loss the data archive frequently replaces storage media and checks log files for hardware or software

[4] *A DOI is an acronym for 'digital object identifier', meaning a 'digital identifier of an object'. A DOI name is an identifier (not a location) of an entity on digital networks. It provides a system for persistent and actionable identification and interoperable exchange of managed information on digital networks. A DOI name can be assigned to any entity – physical, digital or abstract – primarily for sharing with an interested user community or managing as intellectual property. The DOI system is designed for interoperability; that is to use, or work with, existing identifier and metadata schemas. DOI names may also be expressed as URLs (URIs)* [16].

errors. Additional scans with checksums or hash functions are carried out to verify bit streams of archived data and documentation remain unchanged [18].

In the second phase, storage of digital objects takes place in the datorium system. Initially, bit stream preservation is used. It is envisioned to later replace this with format migration, depending on the quality and volume of uploaded research data.

5. OBJECTIVES REACHED SO FAR

At the beginning of 2012 the concept and requirement specifications for the repository were generated. The next step was to select a software tool that conformed to the needs of the new repository. After evaluating several open source tools, the decision was made in favor of DSpace[5], since it met most specified datorium requirements.

datorium is presently in its first testing phase (https://datorium.gesis.org, currently registering authorized users only). Interested and authorized researchers are given the opportunity to enter data from their research projects on datorium and autonomously generate metadata. Besides providing initial depositors with an easy way of archiving and distributing their research data, these depositors help develop datorium by providing feedback and suggestions to the GESIS Data Archive.

In order to provide datorium as quickly as possible to the social science community, at present datorium only serves as a tool for the upload of research data, documentation and generating metadata. Currently datorium does not perform as an online platform for publication of research data. For this reason publication of research data is taking place through GESIS's standard distribution system. To ensure high data quality, data and documentation uploaded in the repository must go through the review processes and data preparation procedures carried out by the curator of the Data Archive and described in chapter 3.3.[6] After this immediate review research descriptions are published according to their content through the appropriate retrieval system (e.g., GESIS Data Catalogue, ZACAT-GESIS Online Study Catalogue, Online database HISTAT, Extended Variable Overview, CESSDA Catalogue).

6. WHAT FOLLOWS NEXT

Additional service components will be implemented as further progress is made to the end of 2013. After the rollout of this next stage data producers can upload data sets and publish them through the datorium platform. Here research data will no longer have to be published via the standard retrieval systems of GESIS. For secondary users of data a common retrieval interface will be built that gives an integrated access to both the holdings of datorium and the standard distribution systems of GESIS. At this point the foundation of most of datorium´s targeted aims will have been set – expanding the scope of data types archived at GESIS, to use datorium as a collaborative virtual working environment, and provide user-friendly and fast retrieval via the datorium publication platform.

Beyond this, data producers can apply for data preparation and *added value archiving* carried out by the Data Archive. A*dded value archiving* is set-up for special datasets. For instance, added value data documentation provides extensive (partly multilingual) standardized descriptions of question texts and answer categories, codes and classifications, or interviewer instructions. In addition supplementary contextual information, like comparable questions, codebooks, variable reports or technical reports, is also added. Moreover this elaborate data preparation contains data cleaning, standardization, harmonization, integration/accumulation and enhancement by additional context data.

In the case of added value archiving, storage of research material takes place in the standard GESIS archive and publication is carried out through standard GESIS retrieval systems. This service is also free of charge.

7. ESSENTIAL BENEFITS OF DATORIUM

The benefits of using datorium are various. Some have been mentioned, but it is important to reiterate them, since their advantages should be viewed from different user perspectives:

a) *Benefits for data depositors:*

Data depositors can publish their research data, documentation, and findings free of charge. datorium allows self-funded and small research projects to benefit through increased visibility of their work, with potential citation and an associated enhanced professional reputation. Furthermore, since datorium is a virtual working environment it is possible to conduct collaborative research. Academic partners in multi-partner cooperative projects will be able to produce documentation for research data together in authorized working groups.

Data providers have full control over their data, because by choosing a defined category of data access they autonomously manage access to their data.

Data is published after a quick review to guarantee data and documentation quality. This allows depositors to receive rapid feedback from the research community.

b) *Benefits for data users:*

datorium gives data users free and fast access to the latest research data, which might provide helpful suggestions and inspire new research. Data can be re-used for secondary research, supporting data repurposing. Financially unsupported secondary research can be conducted at low cost, since data collection efforts can be reduced to a minimum. This means data sharing and data documentation permits research findings to be verified and data re-purposed for future research.

As the Data Archive reviews submissions users can be sure they are dealing with high-quality data, without copyright issues or other complicated legal aspects (e.g., observance of data protection).

---

[5] DSpace is freely available as open source software for academic, non-profit, and commercial organizations building open digital repositories. It preserves and enables easy and open access to all types of digital content including text, images, moving images, mpegs and data sets [19].

[6] This is the reason that metadata covered by datorium follows the metadata schema of the GESIS Data Catalogue DBK [10] and the da|ra Registration Agency for persistent identifiers [11].

c) *Benefits for the academic community:*

datorium facilitates cooperation between scholars and therefore supports synergistic interactions. By publishing data and findings via datorium immediate discussion within the academic community is possible, which might lead to further research based on published data. Uploaded data and related metadata will be preserved long-term, either by format migration or bit stream preservation (depending on the storage location).

Overall datorium has potential to facilitate increased secondary analysis and data re-use.

d) *Benefits for survey respondents:*

Surveys often feel like a burden to respondents. This might be due to reasons of time or for the survey containing sensitive topics that might be hard to deal with. Data re-use eases the pressure on respondents in both cases. Especially data collection from vulnerable groups *"who may be at risk from repeated data gathering intrusions into their lives*" [20, p. 7] can be reduced by data re-use [20]. If researchers re-use data as much as possible they help counteract the effect that respondents become tired or even bothered by taking part in a study.

8. CONCLUSIONS AND OUTLOOK

datorium is suitable for social scientists to efficiently document their research data with associated metadata in a standardized way. Data depositors can digitally preserve their data and share it with the research community. Publishing data and research findings with datorium ensures a visibility to the academic community.

Opening the repository to non-institutional users underlines the originality of datorium's approach, especially so with the focus on small research projects from primary investigators who do not necessarily belong to an institutional organization or are self-funded. A Peer-review carried out by the GESIS Data Archive ensures high data quality. Above this datorium can function as a working environment by allowing multiple partners to jointly create research descriptions within groups.

Because datorium uses a standardized metadata schema interoperable, for instance, with Dublin Core metadata standards it is possible to easily "*weave native Dublin Core Elements into DDI documents*" [20, p.56]. Using DDI enables efficient, accurate use of datasets through standardized documentation. This "*facilitates data access and discovery, improves overall quality, ensures long-term preservation of the information, fosters evidence-based policy making, and supports the establishment of results-based monitoring*" [9, p.108].

By using Dublin Core and DataCite metadata standards datorium meets the conditions for well-organized resource description. By providing datorium to the social science community GESIS promotes the scholarly ideal of data sharing and facilitates long-term digital preservation. Since research data documentation in datorium receives a digital object identifier (DOI), accessibility and traceability of the associated research data is highly reliable.

Implementation is carried out in two phases to provide datorium to the social science community as soon as possible. Since the end of the first phase in April 2013 the GESIS Data Archive provides interested scholars and some of the researchers, who currently have their data documented and digitally preserved by GESIS, access to datorium. In this first phase publication of research data is taking place through GESIS's standard distribution systems, as described in chapter 4. At the end of 2013 datorium will be openly accessible to registered users and research data will be published over the datorium platform (see chapter 5).

9. ACKNOWLEDGMENTS

I would like to thank my colleagues from the Data Archive at GESIS for the inspirational discussions about digital preservation and data sharing. Special thanks go to Wolfgang Zenk-Möltgen, Reiner Mauer, Andias Wira-Alam, Natascha Schumann, Stefan Müller and Laurence Horton for providing their time in numerous meetings as well as their helpful inputs and advices for the realization of datorium.

10. REFERENCES

[1] Winkler-Nees, S. (2012). Stand der Diskussion und Aktivitäten. 2.1 National. In Neuroth, M., Strathmann, S. Oßwald, A., Scheffel, R., Klump, J., Ludwig, J. (Ed.), Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme, pp.23-40. Göttingen: Universitätsverlag.

[2] Fienberg, S.E., Martin E.M. and Straf, M.L. (1985): Sharing Research Data: Report of the Committee on National Statistics. Washington, DC: National Academy Press. [Online: http://www.nap.edu/openbook.php?isbn=030903499X, Accessed 22 April 2013].

[3] Kühne, M. and Meusel, D. (2007): Data Sharing. Unveröffentlichtes Manuskript: Dresden.

[4] Nelson, B. (2009): Data sharing: empty archives. Nature. International weekly Journal of Science 461, pp.160-163. [Online: http://www.nature.com/news/2009/090909/full/461160a.html, Accessed 22 April 2013].

[5] Weichselgartner, E., Günther, A. and Dehnhard, I. (2011). Archivierung von Forschungsdaten. In S. Büttner, H.-C. Hobohm and L. Müller (Hrsg.), Handbuch Forschungsdatenmanagement, pp. 191-202. Bad Honnef: Bock + Herchen Verlag.

[6] King, G. (1995): King, Gary. 1995. Replication, Replication. PS: Political Science and Politics 28: 443–499. [Online: http://gking.harvard.edu/files/abs/replication-abs.shtml, Accessed 1 April 2013].

[7] Schnell, R. (2002): Anmerkungen zur Publikation "Möglichkeiten und Probleme des Einsatzes postalischer Befragungen" von Karl-Heinz Reuband in der KZfSS 2001, 2, S.307-333. Kölner Zeitschrift für Soziologie und Sozialpsychologie 54, 2002, pp.147-157.

[8] Meier, F. (2003): Qualitätsgesichertes Datenmanagement für die Sozialforschung. ZA Information/ Zentralarchiv für Empirische Sozialforschung 52: pp.58-71. [Online: http://nbn-resolving.de/urn:nbn:de:0168-ssoar-199006, Accessed 10 April 2013].

[9] Vardigan, M., Heus, P. and Thomas, W. (2008): Data Documentation Initiative: Toward a Standard for the Social Sciences. The International Journal of Digital Curation. Issue 1, Volume 3, pp.107-113. [Online:

http://ijdc.net/index.php/ijdc/article/view/66/45, Accessed 12 April 2013].

[10] Zenk-Möltgen, W. and Habbel, N. (2012): Der GESIS Datenbestandskatalog und sein Metadatenschema. Version 1.8. GESIS Technical Report 2012-1.

[11] da|ra - registration agency for social science and economic data (2013). [Online: http://www.da-ra.de/en/about-us/, Accessed 15 April 2013].

[12] DataCite (2013). [Online: http://www.da-ra.de/en/about-us/data-cite/, Accessed 12 April 2013].

[13] Dublin Core Metadata Initiative (2013). [Online: http://dublincore.org/metadata-basics/, Accessed 16 April 2013].

[14] Rice, R. (2008): Applying DC to Institutional Data Repositories. Proceedings of the International Conference on Dublin Core and Metadata Applications. p.212. [online]http://dcpapers.dublincore.org/pubs/article/view/945/941 (Accessed 19 January 2013).

[15] http://www.datacite.org/whatisdatacite, Accessed 12 April 2013.

[16] DOI Handbook (2012): The DOI System concept. [Online: http://www.doi.org/doi_handbook/1_Introduction.html#1.6.1 Accessed 18 March 2013].

[17] Babchuk, N., Keith, B. and Peters, G. (1999): Collaboration in Sociology and other Scientific Disciplines: A Comparative Trend Analysis of Scholarship in the Social, Physical and Mathematical Sciences. The American Sociologist 30:5-21.

[18] http://www.gesis.org/en/archive-and-data-management-training-and-information-centre/datenarchivierung/preservation/, Accessed 12 June 2013.

[19] http://www.dspace.org/introducing, Accessed 22 April 2013.

[20] Law, Margaret (2005) “Reduce, Reuse, Recycle: Issues in the Secondary Use of Research Data”. IASSIST Quarterly (Spring). [Online: http://www.iassistdata.org/downloads/iqvol291law.pdf, Accessed 18 June 2013]

[21] Wira-Alam, A., Dimitrov, D. and Zenk-Möltgen, W. (2012): Extending Basic Dublin Core Elements for an Open Research Data Archive. Project Report. In: Proceedings of the International Conference on Dublin Core and Metadata Applications, 2012, S. 56-61 [Online: http://dcpapers.dublincore.org/pubs/article/viewFile/3664/1887, Accessed 18 March 2013].