The International Journal of Digital Curation
Volume 8, Issue 2 | 2013

The GESIS Data Archive for the Social Sciences: A Widely Recognised Data Archive on its Way

Natascha Schumann and Reiner Mauer,
Data Archive,
GESIS Leibniz Institute for the Social Sciences

Abstract

This paper describes initial experiences in evaluating an established data archive with a long-standing commitment to preservation and dissemination of social science research data against recently formulated standards for trustworthy digital archives. As stakeholders need to be sure that the data they produce, use or fund is treated according to common standards, the GESIS Data Archive decided to start a process of audit and certification within the European Framework of Certification and Audit, starting with the Data Seal of Approval (DSA). This paper gives an overview of workflows within the archive and illustrates some of the steps necessary to obtain the DSA as well as to optimize some of its services. Finally, a short appraisal of the method of the DSA is made.

International Journal of Digital Curation (2013), *8*(2), 215–222. http://dx.doi.org/10.2218/ijdc.v8i2.285

 doi:10.2218/ijdc.v8i2.285

Introduction

Founded in 1960 as the Central Archive for Empirical Social Research in Cologne, the GESIS Data Archive for the Social Sciences is one of the pioneers in the field of long term preservation of research data. Nowadays, the archive is part of GESIS – Leibniz-Institute for the Social Sciences[1], which is an infrastructure institution for the social sciences in Germany providing services at all stages of the research data lifecycle.

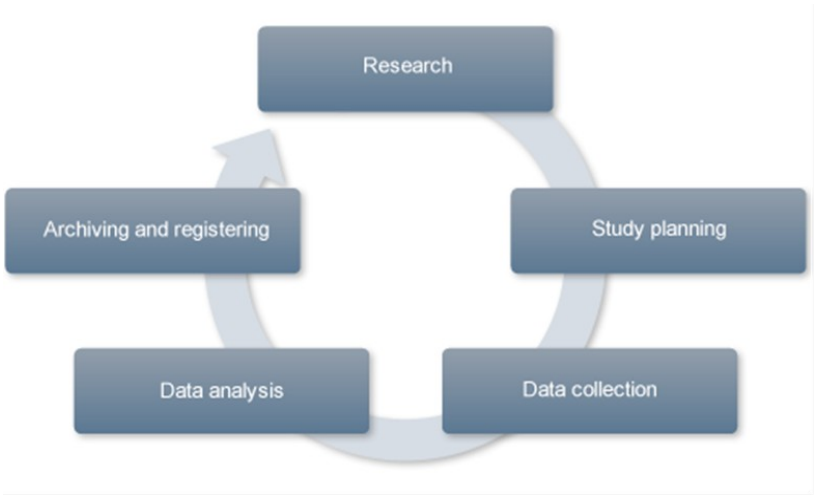

Figure 1. The research data cycle.

The primary focus of the Data Archive is to provide an excellent data service for national and international comparative surveys from the fields of social and political science research. Its holdings comprise of more than 5000 studies on numerous social science topics; all prepared, documented and made available for re-use. Archival and curation tasks regarding acquisition, ingest, data processing and documentation, preservation and provision of access are carried out on the basis of clearly defined processes.

The GESIS Data Archive is well established within the social science community on both a national and an international level. In the past, most of the activities of the archive concentrated on domain-specific aspects, such as understanding and curating social science research data. As a consequence, the archive has strong expertise in curating empirical social research data and is well integrated into a network of European and international social science archives. On the other hand, until recently the archive had not perceived itself as part of the digital preservation community, which is obviously due to the fact that the digital preservation community only evolved during the last ten to fifteen years.

This has changed fundamentally in recent years and nowadays GESIS is involved in preservation initiatives dealing with topics like persistent identifiers, metadata standards and linked data, with the aim of becoming a more active player in the digital

[1]GEIS – Leibniz Institute for the Social Sciences: http://www.gesis.org/en/institute/

preservation community. Another important development is the fact that the organisation of work within the archive has significantly changed during the last decade. There is a much higher level of specialization. Tasks have been differentiated further, and some have been expanded to separate services, like the registration and maintenance of persistent identifiers.

One consequence of these changes is that processes have to be more and more standardized and documented.

Workflows at the GESIS Data Archive

The archive's workflows are quite complex and they are not only focussed on preservation issues but to a large extent also on curation activities. Consequently, many of the procedures are intellectual ones, where staff members need to have a strong understanding of the data, including the underlying scientific concepts, and how to handle it. The workflow is oriented towards the OAIS reference model but starts with a pre-ingest phase followed by ingest, data management and archival storage stages and finally, access services. The ingest process at the GESIS Data Archive – which as a social sciences archive puts a strong emphasis on extensive quality control, data processing and enhancement – cannot be adequately captured by OAIS in this form and detail. Some of the activities performed during ingest at the Data Archive are placed in different functional entities in the OAIS model.

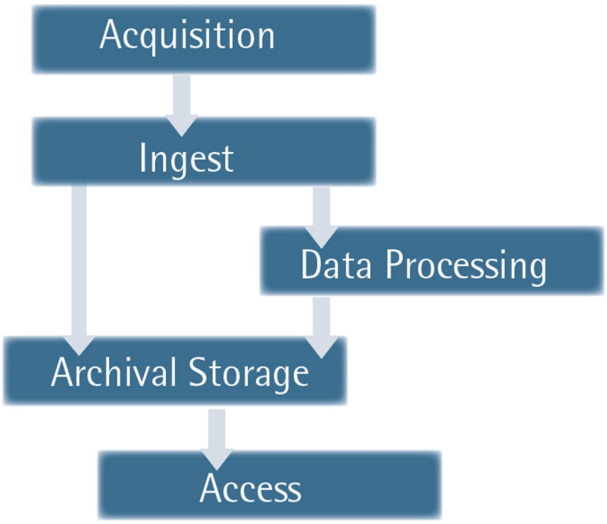

Figure 2. The workflow of the data archive.

Before any data are transferred into the archive a lot of acquisition activities take place. It has to be clarified if a study matches the scope of the archive, i.e. content falls into the key areas of monitoring society and social change in Germany, international comparative survey research, or is in any other regard of scientific relevance. In this pre-ingest phase data creators are advised on how they can best prepare for the actual ingest into the archive. The pre-ingest phase is completed by the signing of an archiving contract, which defines mutual rights and obligations.

The first activity within the ingest process is to check whether all delivered material is complete, correct and in a suitable technical condition (e.g. readable, virus free, etc). Further checks concerning plausibility, consistency, data weighting and data protection are carried out. An important task in the ingest process is the generation of different kinds of metadata, comprising of descriptive metadata for cataloguing, and administrative and technical metadata for preservation issues.

Data and supplementary material are deposited in a file based archival system. The complete submission information package (SIP), consisting of one or more data sets and further supplementary documents, such as questionnaires, codebooks and method reports, is transferred unmodified to archival storage and is stored there as part of the archival information package (AIP). Copies of the original objects of the SIP are then transferred to formats suitable for preservation purposes and added to the AIP. All objects intended for dissemination are converted to different distribution formats (DIP). Beyond mere conversion activities, data sets are further processed by the archive (e.g. introduction of ID and version variables, error correction). All changes are documented in a way that allows for going back to the original version at any time. Each study is assigned a Digital Object Identifier (DOI), a permanent, persistent identifier used for citing and linking electronic resources. GESIS is not only assigning DOIs to its own holdings, but as a member of the Data Cite Consortium and maintainer of the allocation agency da|ra[2] GESIS also offers DOI services to a wider community.

The studies are made available via different services from GESIS. The Data Catalogue[3] is the central access point to the holdings of the archive and comprises of study descriptions for all archived studies, mainly including micro data from survey research and aggregate time series data. Further portals allow for access to special holdings and/or offer particular additional services. For example the ZACAT[4] Online Study Catalogue mainly contains data from international comparative surveys, and besides extensive documentation offers – partly multilingual – (simple) online analysis and visualisation features. Metadata from this portal are also searchable through a pan-European platform, giving access to the holdings of different social science data archives.

Usage of these services is granted free of charge. However, a registration is usually required and data is subject to different access classes which range from ‘open for everyone’ to more restrictive forms where, for example, data depositors must approve usage of the data.

Changes and Challenges

Within the last few years the requirements regarding long-term preservation of, and access to, research data have fundamentally changed. Research data is increasingly seen as ‘… a valuable asset, on which science, technology, the economy and society can advance’ (High Level Expert Group on Scientific Data, 2010). However, supporting sharing and re-use of research data requires adequate long-term

[2]da|ra Registration Agency for Social and Economic Data: http://www.da-ra.de/en/home/

[3]GESIS Data Catalogue: http://www.gesis.org/en/services/research/data-catalogue/

[4]ZACAT: http://www.gesis.org/unser-angebot/recherchieren/zacat-online-study-catalogue/

preservation concepts and activities. Many new initiatives and projects dealing with research data have been initiated; more and more data and other resources have been stored in distributed repositories, and questions of interoperability, linking mechanisms and persistent identifier systems need to be addressed. With the emergence of a variety of new institutions, services and projects dealing with research data, questions of sustainability and trustworthiness are becoming a focus of attention. These questions have been addressed by different initiatives developing criteria for trusted digital repositories. Although approaches differ in detail, they share the common goal to support repositories in checking and demonstrating their capability to preserve digital collections in the long run and to keep them accessible.

Bitstream preservation is the basis of all other measures, but it is not quite sufficient in the terms of digital preservation of data. Data are endangered by obsolescence in the technical environment, as hardware, software, operating systems and formats change rapidly. To address these challenges preservation planning activities are needed. This includes the monitoring technological developments and requires migration steps when formats are starting to become obsolete.

Audit and Certification

General Aspects

The need to prove trustworthiness is not only an issue for new players but also for established ones like the GESIS Data Archive. Data users, depositors as well as funding agencies need to be sure that the data they are using, producing or funding is treated according to established standards and norms. To ensure and emphasize the concern to become a trusted digital archive which fulfils current and future requirements of accessing research data, the archive has started an internal project to preparing an for an audit and certification process. As a first step we have decided to apply for the Data Seal of Approval (DSA)[5]. The DSA is embedded in a European Framework for audit and certification[6] together with the German DIN 31644 (DIN, 2012)/nestor Seal[7], and the ISO 16363 (ISO, 2012) is required for all further certification activities.

The framework consists of three levels of certification, according to different needs and possibilities of digital archives:

1. Basic Certification,
2. Extended Certification,
3. Formal Certification.

The first level, called basic certification, can be reached by acquiring the Data Seal of Approval. The procedure for the DSA is a self-assessment. The repository has to

[5]Data Seal of Approval: http://www.datasealofapproval.org/

[6]European Framework for Audit and Certification of Digital Repositories: http://www.trusteddigitalrepository.eu/

[7]Nestor Seal: http://www.langzeitarchivierung.de/Subsites/nestor/EN/nestor-Siegel/siegel_node.html

 doi:10.2218/ijdc.v8i2.285

meet 16 criteria that reflect different roles concerning data management, the data producer, the repository and the data user.

The next level is the extended certification in addition to a successful DSA. This could be a self-assessment based on either ISO 16363 or DIN 31644.

The formal certification is the highest level of certification and requires a full external audit and certification based on ISO16363 or DIN 31644. It can be assumed that in the long-term, funding by the EU will require that research data has to be preserved in a digital repository that at least possesses a basic certification.

Among other advantages, a certification or self-assessment is an excellent instrument to become aware of the capabilities, strengths and, most importantly, weaknesses or gaps within an institution. Beyond the fact that this process can lead to quality improvements in workflows and services, it has the valuable effect of establishing a common understanding for the mission and goals of an institution amongst staff, management and other stakeholders.

An audit and certification process meets a minimum of two basic aspects for a digital repository. The first is to ensure that it fulfils the requirements of being a trusted digital repository and to display this to the public. The second is the process itself: the preparation and implementation of a self-assessment is a kind of gap analysis, indicating which procedures are already implemented and which ones are missing. This process can be used as an opportunity to address known but not yet eliminated weaknesses, e.g. the creation of a preservation policy.

Activities of the Archive

The first steps include a compilation of available documentation concerning workflows, policies, internal and external standards etc., checking these against requirements and creating missing or improving existing documentation. These steps are time intensive, even more so because many of them depend on each other. Staff members from different teams have to be involved in this process and coordination is required with regard to all changes carried out. Additionally, the DSA, as well as the other certification levels, puts a strong emphasis on transparency. Thus evaluated or newly created documents have to be publicized on the website. Accordingly, we are about to relaunch our website to provide more information about our workflows and services, as well as to provide access to our preservation policy.

In the next steps, existing workflows and processes need to be evaluated. For some procedures it might make sense to adapt or renew them in one way or another. However, this process is a balancing act between keeping a functioning system running while also adapting it to new requirements. Besides the technical or organizational aspects, one especially has to take into account that staff members are used to working within the existing setting. To achieve a sustainable effect, staff need to be convinced of the positive aspects associated with changed standards and workflows. This could be best achieved by involving them in the whole process, which is advisable anyway since utilising their expertise is crucial to a successful implementation.

A central activity in the process so far consists of formulating a preservation policy. Although all our work is based on different guidelines, best practices and other policies, until now we did not have a single document focusing on the general principles of our preservation activities. The challenge in creating such a policy is to provide enough information without running into danger of becoming too detailed. A preservation policy should be a stable document that must not be changed frequently. The level of granularity and the length of the document depend on the institution's mission and the availability of other resources with more detailed information. The process of developing a preservation policy includes and stimulates basic discussions within the organisation, which reveal different point of views and at best can lead to a common understanding of the mission of an institution.

Another motivation to start a certification process was a changing perception of the importance of transparency. Even though the archive always had well organised workflows, known and applied internally, we are faced with an increasing and legitimate external demand for comprehensible information on how we ensure integrity and accessibility of our holdings for the long term.

Assessment Methodology

The Data Seal of Approval was developed as a low threshold and lightweight alternative to a regular certification. It is aimed at, amongst others, digital repositories that may not have the resources for a time and cost intensive ISO certification, but are interested in establishing their trustworthiness. The DSA employs a self-assessment approach, which will be reviewed by a board member of the DSA. As this approach is embedded in the European Framework for Audit and Certification and declared as "basic certification" it is a very good starting point for more advanced certification steps. In comparison to a fully external certification (e.g. ISO) the DSA is a relatively cost-efficient measure, which might deliver an additional argument to convince decision maker within institutions. The 16 DSA guidelines consider many important aspects for a trusted digital repository. The required answers could differ in content, length and quality from institution to institution. This openness takes into account the diversity of existing repositories and archives, and can be seen as one of the strengths of the DSA approach. Furthermore, the DSA supports applicants by providing sufficient information and examples throughout the whole process.

The openness of the DSA approach forces institutions to reflect, interpret or develop their own understanding of their mission, roles, workflows and services. For example, they have to think about conformity to common standards, i.e. is the archive OAIS compliant and what does that mean in practice?

Conclusions

To sum up, the DSA is a very helpful method for both starting a certification process and evaluating your existing organisation of work. And for those who aim for the next level of certification within the European Framework, it is not only a precondition but also creates a good foundation for the next level. Even though we are at the very beginning of the process, we are already observing positive effects. For example, the process of reflection and review leads to a critical look at established procedures and

generates ideas for improvements in existing services, or the creation of new ones. As digital preservation is an on-going process in a changing scientific world, a digital archive would do well to evaluate its services periodically, not just to obtain a seal.

References

DIN. (2012). DIN 31644: Information and documentation - Criteria for trustworthy digital archives. Retrieved from http://www.nabd.din.de/cmd?level=tpl-art-detailansicht&committeeid=54738855&subcommitteeid=112656173&artid=147058907&bcrumblevel=2&languageid=en

High Level Expert Group on Scientific Data. (2010). Riding the wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data to the European Commission.

ISO. (2012). ISO 16363:2012: Space data and information transfer systems - Audit and certification of trustworthy digital repositories. Retrieved from http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510