Data Service Infrastructure for the Social Sciences and Humanities

EC FP7

Grant Agreement Number: 283646

Deliverable Report

Deliverable: D4.2
Deliverable Name: 4.2 “Report about Preservation Service Offers”
Deadline: 31 December 2012
Nature: R

Responsible: NSD, UiB.

Work Package Leader: NSD

Contributing Partners and Editors: Vigdis Kvalheim (NSD), Dag Kiberg (NSD), Eirik Vestrheim (NSD), Trond Kvamme (NSD), Koenraad De Smedt (UiB), Anje Müller Gjesdal (UiB), Mike Priddy (DANS), Astrid Recker (GESIS), Przemek Lenkiewicz (MPI), Paul Trilsbeek (MPI), Sally Widdop (CITY), Eric Balster (CenterData), Bartholomäus Wloka (OEAW).

Executive summary

Following the “Description of Work”, Annex 1 to the grant agreement for the project "Data Service Infrastructure for the Social Sciences and Humanities (DASISH)" this task 4.2 report is very much related to task 4.1 report in so far as it will describe and analyse a selection of existing institutional and academic deposit services within the Social Sciences and Humanities (SSH) based on a common framework and set of information categories that seeks to synthesize the knowledge contained within the different deposit services. By analyzing available publications and documentation; policies and practices; and by accessing and delineating activity by activity based on consistent terminology for those activities, this report provides a set of best practices and guidelines for key activities within SSH data deposit services. These guidelines can contribute to the establishment of a trust relationship between the researchers and the data service centres. The outcome of this report will, together with the State of the Art Report (D4.1), feed into D4.3 and D4.4.

CONTENT

Executive summary 2

1 Objective and Methodology 5

2 Administrative Context 11

3 Functions and Processes - Data Curation 15

4 Recommendations 17

5 UK Data Archive 22

5.1 Organizational Framework 22

5.2 Technological Environment 29

5.3 Data Curation 32

6 GESIS Data Archive 47

6.1 Organizational Framework 47

6.2 Technological Environment 54

6.3 Data Curation 58

7 Norwegian Social Science Data Services 71

7.1 Organizational Framework 71

7.2 Technological Environment 75

7.3 Data Curation 75

8 DANS 85

8.1 Organizational Framework 85

8.2 Technological Environment 90

8.3 Data Curation 95

9 The Language Archive 103

9.1 Organizational Framework 103

9.2 Technological Environment 106

9.3 Data Curation 114

10 European Social Survey – Data Archiving 120

10.1 Organizational Framework 120

10.2 Technological Environment 123

10.3 Data Curation 127

11 Survey of Health, Ageing and Retirement in Europe – Data Archiving...................... 137

11.1 Organizational Framework 137

11.2 Technological environment 141

11.3 Data curation 145

12 Commercial Deposit Services.. 153

12.1 Summary 153

12.2 Dropbox 153

12.3 Figshare 158

12.4 Flickr 160

12.5 Youtube 165

12.6 Conclusion 167

1 Objective and Methodology

This task lays the foundation of the work to be carried out by task 3. The objective according to the description of work is to:

- *"To work out an overview of the offered services and their characteristics*
- *To develop guidelines and recommendations for such services that will create the necessary trust relationship between the researchers and the data service centres".*

Source: DASHISH Annex 1 – Description of work, version 2011 – 08 - 15

Europe has a number of facilities for storage of data, some of which are designed to meet the specific needs of the research community, while others are general public services. Due to the number of repositories available it would have been difficult if not impossible to give a comprehensive description of all deposit services, even if only those affiliated to the five research infrastructures represented in the DASISH project were included. Consequently, it was decided as more appropriate and manageable to make a qualified selection among relevant deposit services. Hence this report should not be viewed as a complete review of the whole European deposit landscape. Instead we have selected 11 deposit services in order to make a suitable knowledge-base for presenting recommendations for the operation of such services.

The selection of deposit services was carried out through a 'purposeful', two-step sampling procedure. Firstly, we provided a broad list based on qualified suggestions from task participants within the different communities. Secondly, we filtered out a selection of services that have reached a certain level of maturity and which we considered as 'qualified' representative services of their designated communities. The selected deposit services were distributed among task partners for further information gathering, assessment and analyses.

Generally, when gathering sets of information for comparison the challenge is to make sure that the information collected is comparable. Hence, we emphasised developing an analytic framework or model containing a controlled set of information categories and consistent terminology to facilitate our work. Based on a common set of information categories our goal was to synthesize the knowledge contained within the different deposit services, through their practices, policies and available documentation.

While D4.1 started from existing reference frameworks and audit checklists, these are not a definitive reflection of best practice in the sense of what's actually being implemented in the repositories. Hence in D4.2, by comparing selected services in a clear fashion, that is, by delineating activity by activity, providing consistent terminology for those activities, and relating them to the professionally approved reference models discussed in D4.1, one can come closer to a more definite version of best practices and guidelines for certain activities within the SSH data repositories.

The fundamental base for the selection process was to cover all areas represented within DASISH. However, our experience is that deposit services offered by institutions, i.e. the institutional archiving-landscape, are more developed within the Social Sciences than the Humanities. This is probably a result of the relatively longer history of centralized deposit services for the Social Services than for the Humanities. The Social Sciences in the US and Europe were among the first disciplines to establish specialized archives for digital data.

Based on input from all involved partners we were selected the following collection of repositories:

- Social Sciences Data Archives
  - UK Data Archive (UKDA)
  - Norwegian Social Science Data Services (NSD)
  - GESIS Data Archive (GESIS)
- Humanities
  - Data Archiving and Networked Services (DANS)
  - The Language Archive (TLA)
- Survey Organizations
  - European Social Survey (ESS)
  - Survey of Health, Ageing and Retirement in Europe (SHARE)
- General Web Services
  - Figshare
  - Flickr
  - Dropbox
  - Youtube

UK Data Archive (UKDA) is the primary repository for social science research data in the UK since 1967. As a national data collection service the Archive, originally called Data Bank, was created by the Social Science Research Council, now the Economic and Social Research Council (ESRC), to bring together "social survey research materials for storage, retrieval and secondary analysis of the information in them". From 1 October 2012, the University of Essex is providing co-ordination for the new UK Data Service which will integrate the Economic and Social Data Service (ESDS), the Census Programme, the Secure Data Service and other elements of the data service infrastructure currently provided by the Economic and Social Research Council, including the UK Data Archive. UKDA is located in Essex, UK. UKDA is a member of CESSDA.

Norwegian Social Science Data Services (NSD) was established in 1971 and was from the start institutionally affiliated to the Research Council of Norway. On 1 January 2003, NSD became a limited liability company owned by the Norwegian Ministry of Education and Research. As a national research infrastructure NSD's main objective is to promote and facilitate empirical research. This is achieved by collecting, processing, archiving, maintaining and disseminating data to research communities. NSD is located in Bergen, Norway. NSD is a member of CESSDA.

GESIS Data Archive (GESIS) was originally founded in 1960 as Central Archive for Empirical Social Research (Zentralarchiv für empirische Sozialforschung, ZA), Europe's first data archive in the Social Sciences. In 1986 it became a member of the newly founded GESIS, a collaboration of three independent Social Sciences infrastructure institutions. Since 2007, the Data Archive is one of five scientific departments of GESIS. GESIS offers services tailored to the needs of the Social Sciences community and aligned with relevant thematic and structural developments in the field. These services are informed by the results of GESIS's research as well as its close co-operation with universities and other partners. GESIS is located in Cologne and Manheim, Germany.

Data Archiving and Networked Services (DANS) is an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO). DANS promotes sustained access to digital research data. For this purpose, DANS encourages researchers to archive and reuse data in a sustained manner, e.g. through the online archiving system EASY. In addition, the institute provides training and advice, and performs research into sustained access to digital information. DANS is located in Amsterdam, the Netherlands. DANS is a member of CESSDA but is here representing all SSH domain.

UKDA, NSD, GESIS, and DANS are members of Council of European Social Sciences Data Archives (CESSDA)[1]. CESSDA is an umbrella organisation for social science data archives across Europe. Since the 1970s the members have worked together to improve access to data for researchers and students. Preparations are underway to move CESSDA into a new organisation known as CESSDA European Research Infrastructure Consortium (CESSDA ERIC).

As the name of CESSDA indicates, the main focuses of the member's deposit services are on data for the Social Sciences. These are archives that to a certain extent store data from other research domains, too. DANS does, however, provide advanced services for data access and preservation for Social Sciences as well as the Humanities. In this report, DANS is therefor included as representing Digital Research Infrastructure for the Arts and Humanities (DARIAH).

[1] CESSDA is now in the process of becoming an ERIC. The name will then be changed from *Council* of European Social Sciences Data Archives to *Consortium* of European Social Sciences Data Archives. Members of CESSDA will most likely become Service Providers for CESSDA ERIC.

Although there are differences, the CESSDA members share many of the same characteristics regarding their objective and purpose. Basically they are all working towards supporting empirical research and providing the research communities with services that makes data available for empirical research and evidence based knowledge. All CESSDA services examined have as one of their task to be the primary repository for social science research data. Some of them, as for instance DANS, provide services also for disciplines outside the Social Sciences. The DANS strategic plan for the period 2011-2015 includes, as an example, a further development to a discipline- independent data-organisation.

The Language Archive (TLA) is established with joined forces of The Max Planck Society (MPG), the Berlin-Brandenburg Academy of Sciences (BBAW) and the Royal Netherlands Academy of Sciences (KNAW) at the Max Planck Institute for Psycholinguistics in Nijmegen, Netherlands (MPI-PL). Although the type of data content is somewhat different the primary goal and policy of TLA is basically similar to the CESSDA institutions. That is to store and preserve digital language resources, to give access to researchers and other interested users and to develop and integrate new technologies advancing language research. TLA is located at the Max Planck Institute for Psycholinguistics in Nijmegen, Netherlands.

European Social Survey (ESS) is an academically-driven large-scale cross-national social survey, and represents one of the two social surveys that are on the ESFRI European Roadmap for Research Infrastructure and are represented in DASISH. The other one is Survey of Health, Ageing and Retirement in Europe (SHARE, see below). ESS was set up in 2001 as a time series survey for monitoring change in social values in Europe. ESS conducts a survey round every second year.

ESS as such, does not have any data deposit service itself. NSD has been the official data archive for the ESS since 2002. NSD is responsible for overseeing the deposit, adjustment, archiving and dissemination of survey data and documentation for each round of the ESS as well as providing resources to enhance analysis for data users. ESS is located in City University, London, UK. ESS Data Archive is located at NSD.

Survey of Health, Ageing and Retirement in Europe (SHARE) is a multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks of more than 55,000 individuals from 20 European countries aged 50 or over.

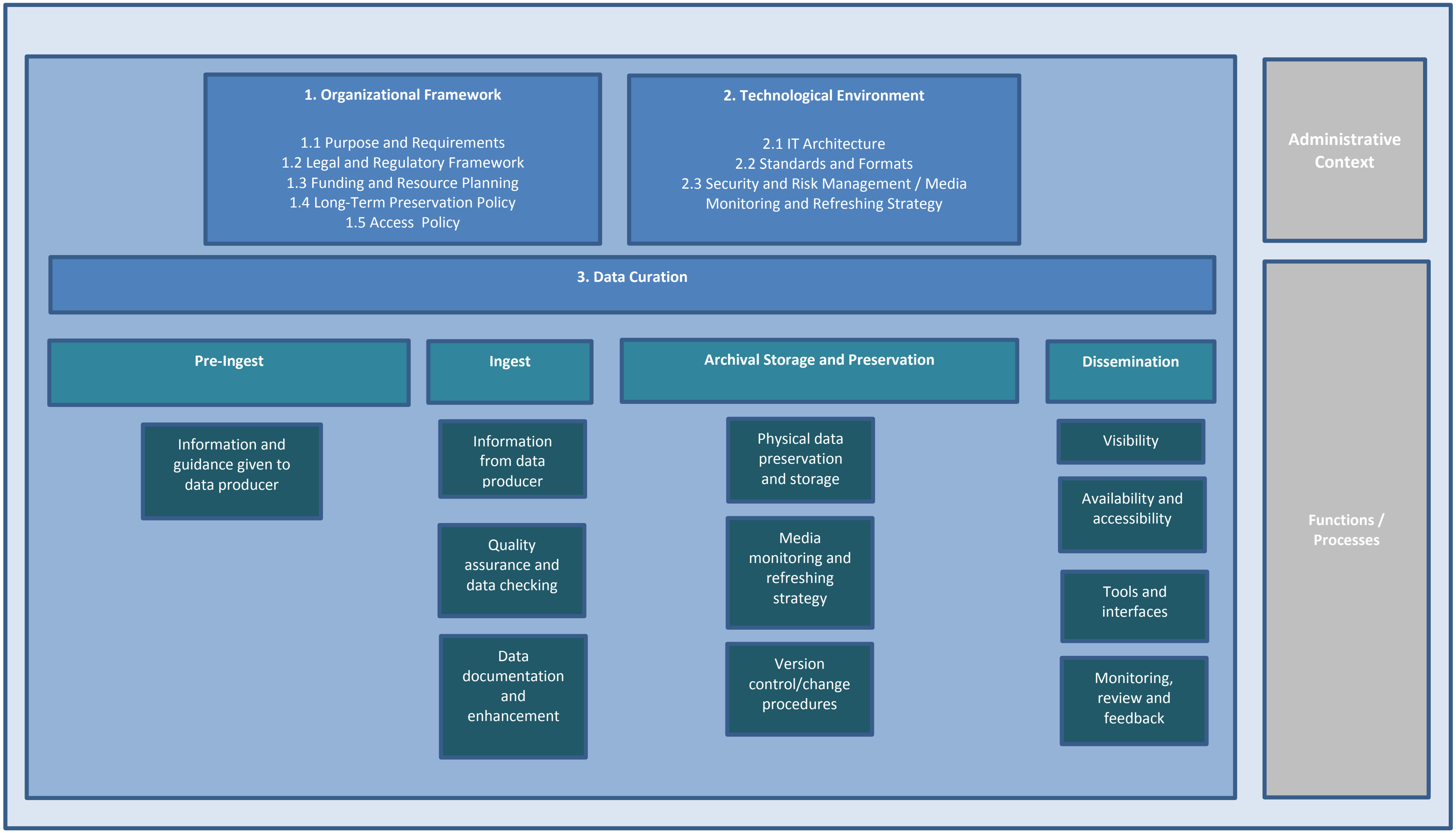
As for ESS, the data are available to the entire research community free of charge. The data is deposited at two data archives. Applications can either be submitted through the SHARE website (http://www.SHARE-project.org) or the GESIS-Data Archive in Cologne, Germany. The administration of the data archive at the SHARE website is managed by CentERdata.

In addition to officially recognized deposit services set up to serve academic research, many researchers frequently use various commercial (and in some cases non-commercial) upload services for uploading, storing and sharing their research data. A selection of such services has been examined; **Dropbox, Figshare, Flickr and Youtube**. These are services that are not designed for research purposes but, due to their usability and availability, are used by researchers.

Based on input from all involved partners we were able to identify and group together a general set of attributes and activities in the data repository infrastructure. On the **administrative level**, we identified a set of organizational and key supporting processes areas, including strategies, policies, finances, etc., and a set of attributes concerning the technological environment. Secondly, we identified a set of **processes**, or key practices and procedures involved in the data repository data curation activities.

Information on each deposit service was gathered through open sources on the services' websites and/or from staff members at each institution. Both in the information collection and compilation phase as well as the reporting are structured in three main segments corresponding to the categories identified in T4.1 and further in specific sub-sections as shown in figure 1.

Figure 1: Data Repository Attributes – A Model for Trust


1. Organizational Framework
1.1 Purpose and Requirements
1.2 Legal and Regulatory Framework
1.3 Funding and Resource Planning
1.4 Long-Term Preservation Policy
1.5 Access Policy
2. Technological Environment
2.1 IT Architecture
2.2 Standards and Formats
2.3 Security and Risk Management / Media Monitoring and Refreshing Strategy
Administrative Context
3. Data Curation
Pre-Ingest
Ingest
Archival Storage and Preservation
Dissemination
Information and guidance given to data producer
Information from data producer
Quality assurance and data checking
Data documentation and enhancement
Physical data preservation and storage
Media monitoring and refreshing strategy
Version control/change procedures
Visibility
Availability and accessibility
Tools and interfaces
Monitoring, review and feedback
Functions / Processes

2 Administrative Context

To ensure that the information can be understood and used by the consumers of the future information system, the information must not only to be preserved but maintained actively through carefully planned and managed procedures. One of the most important tasks in this context is to ensure that not only the data sent into the future remains its integrity and authenticity, but that the integrity and authenticity is also guaranteed for the whole preservation system and its collections (Schott, et.al.).

The fundamental requirement of deposit services for scientific use is that the services can comply with requirements for safety and longevity. Services must hence operate in a perspective that spans decades. At the same time research communities as well as the general community needs to be assured that data delivered today can be retrieved and used tomorrow - and the day after tomorrow. In order to meet such requirements, a number of conditions need to be fulfilled. Firstly, the service must be funded in a sustainable and suitable manner and from sources that are likely to maintain funding over time. Furthermore, the service must have objectives and policies that support the maintaining of the repository and meet the needs of the research objectives. The services must be operated in a predictable and transparent manner and the operation must be perceived as such by the users of the services.

Reviewing various deposit services within the SSH area, we have focused on how and to what extent the selected deposit services meet expectations and qualities that are significant in this respect. The question is basically; how can we be guaranteed that the services offered have characteristics that make us believe that data deposited today keeps its value in the future?

At the end of the answer of course is that we can never be absolutely certain. We have to rely on the services organizational model, funding scheme and procedures implemented. Some of the services examined have been up and running for more than 40 years while others is established in more recent times. Some base their operation on public funding from recognized public funding agencies, while others are more projects based or relies on commercial income. Some are designed to be archival institutions for research data only; some are commercial upload services, while others do not have a deposit service or archive by their own.

UKDA, NSD, GESIS, and DANS are members of Council of European Social Sciences Data Archives (CESSDA)[2]. CESSDA is an umbrella organisation for social science data archives across Europe. Since the 1970s the members have worked together to improve access to data for researchers

[2] CESSDA is now in the process of becoming an ERIC. The name will then be changed from *Council* of European Social Sciences Data Archives to *Consortium* of European Social Sciences Data Archives.

and students across Europe. Preparations are well underway to move CESSDA into a CESSDA European Research Infrastructure Consortium (CESSDA ERIC).

As the name CESSDA indicates the main focus of the member services are on data for the Social Sciences. However, these archives are to a lesser or greater extent archiving data from other research domains, too.

During the selection process it became obvious to us that the institutional and academic deposit services offered to scholars across Europe are far more developed within the Social Sciences than the Humanities. Through the ESFRI-process we do now, however, see emerging infrastructures for digital data services also within the Humanities.

DANS does provide services for data access and preservation for Social Sciences as well as the Humanities. In this report, DANS is included to represent Digital Research Infrastructure for the Arts and Humanities (DARIAH). Similar multi-discipline services are also found at NSD and UKDA.

The Language Archive which is established with joined forces of The Max Planck Society (MPG), and the Royal Netherlands Academy of Sciences (KNAW), is partly funded by The Max Planck Society (MPG), The Berlin-Brandenburg Academy of Sciences (BBAW) and the Royal Netherlands Academy of Sciences (KNAW). Also TLA is funded by a mix of basic funding and external projects. The three major funding bodies of the TLA (BBAW, KNAW and MPG) secure the basic operation of the facility by providing funds for employing about 7 core members. By now, TLA has more than 20 staff members, most of them being funded by external research projects.

Although there are differences, CESSDA members share much of the same characteristics regarding funding, objectives, etc. Basically they are all working towards supporting empirical research. All CESSDA services examined are supposed to be primary repository for social science research data. Some of them, as for instance DANS, provide services also for other disciplines than the Social Sciences. The DANS strategic plan for the period 2011-2015 includes a further development to a discipline-independent data-organisation.

At DANS two-thirds of all revenues are structural grants by the Royal Netherlands Academy of Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO). Financial plans by KNAW/NWO are guaranteed for every 4 years. DANS has to deliver annual reports and every 5 years a roadmap about the future strategy.

Similar arrangements are found at NSD, which has around one-fifth of all revenues from the Research Council along with long and semi-long funding from various ministries and the European Union. As NSD is owned by the Ministry of Research and Education it should provide a solid basis for NSD’s business planning both in the short and long term.

The UKDA is funded by the Economic and Social Research Council (ESRC), the Joint Information Systems Committee (JISC) and the University of Essex and has thus a trusted and acknowledged funding basis. GESIS is funded jointly by the federal government and the federal states according to Article 91b[3] of the German Federal Constitution. GESIS generates additional revenues by offering value-added services to the Social Sciences community. Funding in accordance with Article 91b is assigned in seven-year cycles.

The services collection policy does largely reflect the scope and objective of the services. The policy varies from system where;

- data projects have an obligation to deposit data;
- the repository is actively contacting relevant projects, surveys, or programs;
- the repository passively addressing potential depositors;
- a combination of all of the above.

DANS is an example of a repository that applies a combination of different concepts. Via a system for self-archiving the data producers deposit their research data with DANS themselves using the deposit service in the EASY archive. Such an arrangement is also implemented at for instance TLA.

In the Netherlands certain data projects have an obligation to deposit data at DANS due to the requirements from their subsidiary, for example, research funded by the Netherlands Organisation for Scientific Research (NWO). Also Dutch archaeology adheres to a national regulation, which rules that all digital documentation from archaeological research projects needs to be deposited for long-term preservation. Similar contractual deposit rules are found in Norway and the United Kingdom. In Norway NSD on mandate from the Research Council of Norway (RCN) archives data from research projects from the Social Sciences, Humanities, Medical and Health Research and the Natural Sciences whereas in the UK the UK Data Archive receive data funded by the Economic and Social Research Council (ESRC) and data which fall under the remit of the relevant National Archives' Operational Selection Policy (OSP30)[4].

The GESIS Data Archive's acquisition policy is basically thematic focused. The objective is to build a broad collection of data of general relevance to Social Sciences research. Thus, while there are certain thematic areas in which the Data Archive has built particularly comprehensive holdings (see Table 2), the collection policy does not per se limit the thematic focus of data to certain areas.

[3] Education programmes and promotion of research

[4] http://www.nationalarchives.gov.uk/documents/information-management/osp30.pdf

As stated above ESS as well as SHARE is not primarily a data service and repository but at data creator, and consequently in a different situation than the other facilities included. ESS conduct a comprehensive data collection through their survey rounds in up to 34 countries. Data from the interviews, together with paradata such as timing variables, call record data and information about neighbourhood characteristics that are collected at the time of interview are transferred to ESS Data Archive at NSD for depositing, further processing and dissemination. The description of ESS will thus primarily focus on the work of the ESS Data Archive at NSD.

Both the research community and the society as such need trusted repositories. A trusted digital repository is one whose mission is to provide reliable, long-term access to managed digital resources. In order to fulfil such mission the repository needs to have a long-term and secure funding to ensure that data is maintained over time and to achieve the necessary confidence in the research community. Among the services described herein, the facilities specifically designed to meet the needs of the research community have a financial situation which provide the basis for trust.

To be a trusted repository requires having applicable repository strategies. We have thus looked at the services' implemented policies. All SSH archives examined have a policy of active preservation. For UKDA the policy initiates "measures which are necessary for the protection of its collections, and to meet, or extend, nationally and internationally agreed standards for the preservation of digital materials". Similar formulations are also found at for instance at NSD that has implemented a set of strategies involving a focus on hardware, software and human knowledge, and where the goal is to preserve data and be able to make them available in perpetuity. GESIS voices its commitment to preserve and provide access to Social Sciences research data in its by-laws and mission statement. Another example is DANS which ensure that archived data can still be found, accessed and used in the future.

Trusted long-term deposit services need to strive against fulfilling international well recognized standards. *Data Seal of Approval*[5] (DSA) that was originally developed at DANS is a standard that several repositories have obtained. As part of the construction of CESSDA ERIC, CESSDA will conduct a conformity assessment project, involving use of DSA, starting February 2013. DSA is described in more detail in D4.1[6] along with a number of other relevant standards or reference models.

DANS, UK Data Archive and TLA already hold a DSA each. GESIS Data Archive has currently entered a certification process, in which obtaining the Data Seal of Approval will be the first

[5] http://datasealofapproval.org

[6] 4.1 Roadmap for Preservation and Curation in the SSH

step. Also other repositories not included in this report have obtained DSA, for instance The Archaeology Data Service (ADS)[7].

None of the commercial services included in this report can be regarded or recommended as trusted long-term repositories and are thus no replacements for data archives designed and designated for research infrastructural purposes. The main concern is that there are no guarantees of longevity and continuity in terms of use. It is also difficult to be confident that data is stored in a way that makes it possible to retrieve and read them after long term storage.

Dropbox for example, claims to offer a very secure storage service. Nevertheless several research institution in the US advice against using it for storing data of a sensitive or critical nature, and also raise concerns as to whether IPR is properly protected using cloud-based services. FigShare cannot be regarded as a repository for long-term preservation as long as it "reserves the right, at its sole discretion, to modify or replace any of the Terms of Use, or change, suspend, or discontinue the service (including without limitation, the availability of any feature, database, or content) at any time [...] Company may also impose limits on certain features and services or restrict your access to parts or all of the Service without notice or liability."

In Flickr the rights to (data) materials is tied to on sole user. The material will thus be cancelled and materials can be permanently deleted after the user's death, which is contrary to the overall purpose of the ESFRI process; i.e. building robust and sustainable institutional solutions for data management and access. YouTube offers no guarantee that the data stays available at any time in the future. There exists no written assurance from Google as to the long term preservation of the content. On the contrary; "YouTube reserves the right to discontinue any aspect of the Service at any time."

3 Functions and Processes - Data Curation

Data curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle[8]. In the model (see figure 1, page 10) created for this report the digital curation lifecycle comprises of four steps, namely Pre-Ingest; Ingest; Archival Storage and Preservation; and Dissemination.

[7] A discipline-based service providers within the Arts and Humanities Data Service (AHDS). ADS is located at University of York, UK.

[8] http://www.dcc.ac.uk/digital-curation/what-digital-curation

Pre-Ingest

- As a broader term this can be understood as the conceiving and planning of the creation of digital objects, i.e. the evaluation of digital objects and the selection of those requiring long-term curation and preservation; and the adhere to documented guidance, policies and legal requirements.

- It may also be understood as the pre-accessioning phase or the preliminary phase. In the PAIMAS standard (CCSDS 2004[9]) it is defined as: "...the initial contacts between the Producer and the Archive and any resulting feasibility studies, preliminary definition of the scope of the project, a draft of the SIP definition [see below] and finally a draft Submission Agreement".

- It is closely related to the Formal Definition Phase (CCSDS 2004), which "…includes completing the SIP design with precise definitions of the digital objects to be delivered, completing the Submission Agreement with precise contractual transfer conditions such as restrictions on access and establishing the delivery schedule".

In this report we limit the pre-ingest phase to *information and guidance given to the data producer*.

Ingest

- The ingest function in our model is in compliance with the Ingest Functional Entity defined in the OAIS model. That is, the functional entity that contains the services and functions that accepts information (data) from the data producer and ensures the storage and understandability of that information. In the OAIS model the information received from the data producer is referred to as the Submission Information Package (SIP), i.e. the information delivered by the producer to the repository.

- For information gathering purposes we have subdivided this functionality into three sub-functions, namely *Information from data producer, Quality assurance and data checking,* and *Data documentation and enhancement*.

Archival Storage and Preservation

- Broadly speaking this covers all actions that ensure the long-term preservation and retention of the ingested digital objects, and that keeps the data in a secure manner as outlined by relevant standards. In the OAIS model this is referred to as the Archival Information Package (AIP), i.e. the information and associated descriptive data (metadata) which is preserved within the repository.

[9] CCSDS 651.0-M-1. (2004). Producer-Archive Interface Methodology Abstract Standard (Magenta Book. Issue 1. May 2004). Also available as ISO 20652:2006.

- For information gathering purposes we have subdivided this functionality into three sub-functions, namely *Physical data preservation and storage*, *Media monitoring and refreshing strategy*, and *Version control/change procedures*.

Dissemination

- This covers all activities that ensure that data are accessible to designated users for first time use and reuse. Some material may be publicly available, whilst other data may have access restrictions. It is closely related to the OAIS functional entity named Dissemination Information Package (DIP), i.e. information derived from the AIPs and sent by the archive/repository to the data consumer.

- For information gathering purposes we have subdivided this functionality into four sub-functions, namely *Visibility, Availability and accessibility; Tools and interfaces;* and *Monitoring, review and feedback*.

4 Recommendations

In this report selected existing institutional and academic deposit services are described and analysed based on the recommendations and guidelines for proper data management presented and discussed in the state of the art report (D4.1). Based on this and the needs of the participating communities, the following set of recommendations will work as a point of departure for a broader review and recommendations for improvement and promotion of service providers for SSH researchers to be conducted in WP4.3.

Note that although we have divided the main functionalities into sub-functions for information gathering purposes, the recommendations we provide in this chapter are mostly at a higher level. In WP4.3 they will be operationalized into more concrete requirements and questions to be answered by possible institutional and academic deposit services.

Specialised support services

Specialised deposit services should replace "private" or project based deposit solutions in order to support and enhance long-term preservation and open access to research data.

The purpose of deposit services for scientific use is that the services can comply with requirements for safety, longevity and continued access. Services must hence operate in a perspective that spans decades. At the same time research communities as well as the general community needs to be assured that data delivered today can be retrieved and used tomorrow - and the day after tomorrow. To achieve this objective requires facilities which are well institutionally embedded and can demonstrate a high degree of permanence.

Purpose and Requirements

Deposit services should have a clear mandate, operational status and responsibilities as a data-archive and communicate these internally and externally, to all repository stakeholders.

These key stakeholders include the producer, the rights holder, the repository and the consumer.

As discussed in the D4.1 report, under the trust maturity level 1, some repositories may be in a position where purpose, scope and objectives may only exist implicitly or may not be stated at all. This may apply both to emerging repositories and experienced repositories that have dispersed into several heterogeneous tasks and processes and wants to reestablish itself. Hence the repositories should explicitly define the source of the repository's mandate (e.g. whether the repository receives a significant proportion of its material from a legally mandated source; defining the maturity level, etc.)

Scope and objectives

The repository should have a mission statement that reflects its goals and its commitment to the preservation of, long term retention of, management of and access to digital information.

The statement should have a clear focus on the end purpose of the archiving process which is to serve the consumers or "designated communities" of current and future users; hence it should include an explicit definition of its designated community, the designated communities' associated knowledge base(s) and the type of material in the repository. These definitions should be appropriately accessible.

The statement could also clarify which of the three acquisition strategies (push, pull, self-creation) is intended to account for the significant portion of the total material in the repository.

Repository policies (Collection policy, Storage policy, Access policy, Criteria for evaluating data)

The repository should have written policies or other documents that specify the type of information it will accept, preserve, retain, manage, and provide access to.

A clear set of policies or guidance documents can reduce long term cost by defining the aim and direction of collections and services for more efficient decision making. The policies should be accessible and understandable for all relevant stakeholders, and should have a clear end-purpose perspective.

The long-term perspective should be explicitly defined and should consider the broader economic setting in which preservation decisions are made.

Ethical and Legal Framework

The repository should identify and comply with roles, rights and obligations concerning use of various types of data.

For example the intellectual property rights agreement should cover permissions needed for content and associated software, for future migration/emulation of content to new formats for the purposes of preservation, and for permissions in respect of copyright protection mechanisms.

More generally the drafted agreements should cover statutory permissions and legal deposit obligations in respect of electronic materials; grant and contractual obligations in respect of electronic materials; conditions, rights and appropriate interests of the designated community; confidential information and protection of the confidentiality of individuals and institutions; and protecting the integrity and reputation of data creators or other stakeholders.

Funding and Resource Planning

The repository should create and maintain business and financial plans and have a clearly defined funding model.

The repository should monitor and analyze financial risks, benefits, investments and expenditures and carry out promotional activities suitable to the repository's needs. Business models and financial plans should be reviewed regularly. The repository should have and maintain a succession plan and contingency plans for financial cutbacks or emergencies (e.g. if the repository ceases to exist).

Concerning human resources, the repository should support training and development activities.

Pre-Ingest

The repository should facilitate solutions that gather necessary information, establish contact with data producers, and provides the data producers with all necessary information.

The ‘necessary information’ includes in particular information about the repository services, the minimal requirements that have to be fulfilled for the repository to accept data for archiving, as well as licensing options and the deposit agreement.

The service solutions should make it possible for the contact to be initiated either by the repository or the data producer (or other relevant stakeholders).

Ingest

The repository should process and validate the received information from the data producer and initiate appropriate follow-up action with the data producer if necessary.

Ingest functions include receiving information (data packages), from the producer, performing quality assurance on the received information, and generating an Archival Information Package (AIP) which complies with the repository`s formatting and documentation standards (CCSDS 2012[10]). The AIP should include administrative, descriptive, structural and technical metadata.

Archival Storage and Preservation

The repository should perform data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats to ensure trustworthy long-term preservation and retention of data.

This recommendation underlines that the repository should take actions to ensure that the information stored in the data-archive remains accessible to, and understandable by, the Designated Community over the long term.-

The archival storage and preservation should take into consideration the full curation lifecycle of digital materials. This includes plans for management and administration of all curation lifecycle actions, and monitoring developments in standards, formats, hardware, software and storage technologies, user communities and reservation requirements. There should always be a focus on the end-users` needs.

[10] CCSDS 650.0-M-2. (2012). Reference Model for an Open Archival Information System (Magenta Book, June 2012). Also available as ISO 14721:2012.

Dissemination

The repository should provide tools and interfaces that makes the archived data easy accessible, by using unique persistent identifiers for each available data package.

Dissemination, or the access activities function, should provide one or more interfaces to the information holdings of the repository. This interface will normally be via computer network, but might also be implemented in the form of a walk-in facilities, etc.

Giving the users a way to request new features could also help the repository staying up-to-date. The repository should also monitor the availability of data and checksums.

5 UK Data Archive[11]

5.1 Organizational Framework

5.1.1 Purpose and Requirements

5.1.1.1 Scope and objectives[12]

UK Data Archive (UKDA) has been the primary repository for social science research data in the UK since 1967. As a national data collection service the Archive, originally called Data Bank, was created by the Social Science Research Council, now the Economic and Social Research Council (ESRC), to bring together "social survey research materials for storage, retrieval and secondary analysis of the information in them".

For over three decades, preservation of these collections has been a core function of this enterprise. The ESRC Research Data Policy emphasises the importance and requirement of depositing ESRC data with the Archive, and The National Archives' Acquisition and Disposition Strategy regulates the disposition of datasets created by government departments.

In 2005 the UK Data Archive was designated a Place of Deposit[13] by The National Archives. This status has meant that the Archive has had to modify a number of its procedures to ensure that its previous emphasis on usability with reliability, and levels of integrity has been replaced with a much stronger emphasis on authenticity, integrity and reliability, while not ignoring usability.

From October 1st 2012, the University of Essex is providing co-ordination for the new UK Data Service which will integrate the Economic and Social Data Service (ESDS), the Census Programme, the Secure Data Service and other elements of the data service infrastructure currently provided by the Economic and Social Research Council, including the UK Data Archive.

The integration follows an economic evaluation of ESDS, which reveals that for every pound currently invested in data and infrastructure, the service returns £5.40 in net economic value to users and other stakeholders[14].

The UK Data Service will provide a unified point of access to the extensive range of high quality economic and social data, including valuable census data. It is designed to provide seamless

[11] Most of this text is copied directly from original sources. Additional information was provided by Hilary M. Beedham, UKDA.

[12] Preservation Policy (2011:1).

[13] The UKDA, responsible for archiving government records since 1967, became a designated Place of Deposit for public records for The National Archives on 1 January 2005. The award to the UKDA has no direct impact on the service provided to ESDS, but it strengthened the relation between UKDA and government departments and will help to ensure the continued flow of high quality government data for ESDS users (see http://www.data-archive.ac.uk/news-events/news.aspx?id=1377)

[14] http://www.esrc.ac.uk/_images/ESDS_Economic_Impact_Evaluation_tcm8-22229.pdf

access and support to meet the current and future research demands of both academic and non-academic users, and to help them maximise the impact of their work.

The UK Data Service will:

- act as a trusted national digital repository for a wide range of data providers and users
- provide a single point of access and support to a broad range of high-quality economic and social research data
- provide controlled access to sensitive and/or disclosive data through secure settings
- raise the awareness of the data held by the UK Data Service among those who are not yet using the service, especially among those in business, third sector and at all levels of government
- extend use of its data holdings to the widest possible academic, policy and practitioner communities for generating greater impact
- develop and promote common standards and agreed strategies for data preparation, processing, documentation and preservation to promote data sharing and reuse
- help the social science community to develop the skills necessary to use the data available
- work with a wide range of stakeholders in the UK and overseas, including data suppliers, data funders and users, Institutional Repositories and Doctoral Training Centres

The UK Data Service will be a distributed service led by the University of Essex in collaboration with the University of Manchester and the University of Southampton. It will incorporate the new Census Support Service led by the University of Leeds. All of the host organisations are making a significant contribution to running the new service.

The ESRC has established a UK Data Service Governing Board that will have the responsibility and authority to ensure that the service is developed, managed and maintained in a manner that maximises its benefit as a long-term world class data resource.’

It is also worth mentioning in this context that the UK Data Archive remains the Host Organisation for the UK Data Service, and it continues to undertake strategic work in the wider context of data service infrastructure. The key strategic goals for the UK Data Archive are presently:

- Promoting best practice in data curation
- Raise standards in data management
- Raise standards in data security
- Drive archival innovation
- Advance professionalization of data service infrastructures
- Integrate these activities into the UK Data Service

5.1.1.2 Collection policy[15]

The UKDA collects data, information and other electronic resources of long-term interest and use across the range of social science and historical disciplines. They are acquired to support research and teaching activities in the UK and elsewhere. Materials of interest fall into four key areas:

- data and electronic resources for research: these are data that are suitable for informed use in a variety of research settings;
- data and electronic resources for teaching and learning: these are datasets that are accompanied by purpose-written teaching materials;
- replication data and electronic resources: these are collections of materials, (data, computer programs and instructions, and related outputs) necessary for the replication of published or unpublished research;
- data and electronic resources which have a statutory obligation to be made available to the public: these are data which fall under the remit of the relevant National Archives' Operational Selection Policy (OSP30)[16].

The UKDA seeks to identify and acquire material within the following areas:

Discipline coverage: at the broadest level, data and other electronic resources relating to society, in particular data about individuals or groups of individuals. This includes strategic social science and economic datasets, e.g. unemployment statistics and major household surveys.

Geographic coverage: data across a broad geographic coverage focusing on the UK and cross-national datasets but including material from other countries where appropriate and in particular where these provide opportunities for comparative research e.g. European data.

Temporal coverage: there are no restrictions on temporal coverage, although historical accessions are usually acquired through the History Data Service. Time series and panel data: data are sought which create or add to a time series and/or panel survey.

Thematic coverage: data which assist in the creation of a coherent body of materials relating to a particular discipline or field of enquiry e.g. health, nutrition, etc.

In addition, the UKDA seek to acquire material:

- at the specific request or recommendation of a user or group of users;
- when the data collection has been fully or partially funded by the ESRC or when data have been obtained from another source using ESRC funds;

[15] UKDAs Collections Development Policy has undergone a considerable change, but as it has not yet been completely finalised and made available. This section refers UKDAs former Collection Development Policy (see UK Data Archive Collection Development Policy (2008:2-3)).

[16] http://www.nationalarchives.gov.uk/documents/information-management/osp30.pdf

- when the data collection at least partially fulfils the subject criteria above and the costs of ingest, preservation and dissemination are paid for in advance;
- for replication purposes;
- accompanied by appropriate teaching materials.

5.1.2 Criteria for evaluating data[17]

The following criteria are used to evaluate potential datasets acquired by or submitted to the UKDA.

- Ensure they fall within the 'Scope of Collections' (see 7.1.1.2 Collection Policy).
- Assess their content, long-term value and the level of potential interest in their re-use. Factors influencing this evaluation include:
  - the geographic and/or temporal scope is significant;
  - the subject coverage of the data is broad and may be of interest across the social science and Humanities disciplines;
  - the data are not generally available in any other form e.g. paper;
  - accession into the UKDA makes the resource more accessible;
  - the dataset adds to or is made more valuable by existing holdings, in particular where it fits into an existing series;
  - the dataset fills a gap in the existing holdings;
  - there is research and/or teaching activity in the subject area covered by the data;
  - the data are such that their continued access would otherwise be threatened.
- Determine if they may be viably managed, preserved and distributed to potential secondary users. Factors influencing this evaluation include:
  - the data are of a type with which the UKDA has expertise or may easily obtain expertise or expert advice;
  - the data format can be converted to suitable dissemination and preservation formats;
  - the level and quality of documentation reach an appropriate standard to enable a secondary analyst to make informed use of the data.
- Determine if there is another archive, data centre or institutional repository more appropriate for dissemination, curation or preservation of these data. UKDA welcome datasets that are documented to UKDA standards.

There are also criteria for rejection of data. These are for guidance only and the UKDA may be prepared to accept data that fall into one or more of these categories after discussion and agreement. The criteria for rejection are as follows:

- there are problems with the sample size, the data are very localised or are of peripheral interest for social scientists;

[17] UK Data Archive Collection Development Policy (2008:3-4).

- the documentation is insufficient to enable understanding and further analysis of the data;
- the data would be more effectively dealt with by another organisation or institution;
- there are insurmountable legal obstacles e.g. rights management issues have not been, or are unlikely to be satisfactorily resolved before ingest;
- full use of the data would not be possible without infringing legislation, e.g. Data Protection Act;
- ethical issues remain unresolved before ingest;
- the depositor wishes to impose unnecessarily stringent access conditions or wishes to place an indefinite embargo on use. Unless this is unavoidable and the data are judged to be of long-term value, the UKDA may be willing to accept material for preservation-only subject to payment of the appropriate charges.

5.1.3 Legal and Regulatory Framework[18]

The legal and regulatory frameworks for the management of the data collections accessioned by the UK Data Archive are complex. The University of Essex is the legal entity under which the Archive functions. The Archive is a centre within the University of Essex, and has no legal status.

The relationship between the depositor of a data collection and the Archive is based on:

- a legally-binding deposit agreement and licence (known as the Licence Agreement) which confirms the rights and obligations of both parties and offers an opportunity for depositors to specify the conditions under which access may be given to third parties;
- an assertion of copyright and intellectual property rights to ensure that the data creator/depositor has cleared all necessary permissions;
- where necessary, negotiations for licence agreements with third parties to enable the Archive explicitly to distribute the material to particular user communities.

The Archive will not ingest materials that have unclear ownership or unresolved rights issues.

In preserving its collections, the Archive follows:

- Copyright, Design and Patents Act, 1988 and amendments to this Act;
- Data Protection Act, 1998;
- Freedom of Information Act, 2000;
- EU Copyright Directive, 2001;
- Environmental Information Regulations, 2004;
- English or UK law for commercial agreements and contract law;
- Current best practice.

[18] Preservation Policy (2011:5-6).

In terms of national standards for the management of information security, the Archive is registered to BS ISO/IEC 27001: 2005 - Information technology -- Security techniques -- Information security management systems – Requirements and follows the Cross Government Actions: Mandatory Minimum Measures.

The UK Data Archive also expects depositors to undertake an expedited review of ethical issues relating to datasets which they may wish to deposit with the Archive. Depositors are expected to ensure that there is no potential for risk of harm to any participants in making data available to third parties. The ESRC's Research Ethics Framework (REF) provides guidance to both the Archive and to its researchers.

5.1.4 Funding and Resource Planning[19]

The UK Data Archive is funded by the Economic and Social Research Council (ESRC), the Joint Information Systems Committee (JISC) and the University of Essex.

The UK Data Archive is committed to supporting continued funding for all of the operations relating to preservation management. Resource management for preservation of digital resources includes:

- technical infrastructure, including equipment purchases, maintenance and upgrades, software/hardware obsolescence monitoring, network connectivity, etc.;
- financial plan, including strategy and methods for financing the digital preservation programmes and commitment to long-term funding;
- staffing infrastructure, including recruitment, induction and ongoing staff training.

The Archive has established a rolling planning scheme for lifetimes of computer equipment and storage media to facilitate forward planning for the necessary upgrades.

As stated in the Archive's Strategic Plan, 2010-15, the preservation of data and documentation to ensure they remain usable over time is a core activity of the Archive. The Archive, therefore, makes every effort to remain up to date with any relevant technological advances to ensure continued access to its collections.

5.1.5 Long-Term Preservation Policy[20]

The UK Data Archive exists to support high quality research, learning and teaching in the Social Sciences and Humanities by acquiring, developing and managing data and related digital resources, and by promoting and disseminating these resources as widely and effectively as possible. To ensure the continued use of these resources the Archive follows a policy of active preservation with the aim of ensuring the authenticity, reliability and logical integrity of all

[19] Preservation Policy (2011:13).

[20] Preservation Policy (2011:3).

resources entrusted to its care while providing formats suitable for research, teaching or learning, in perpetuity.

The formulation and biennial revision of a preservation policy for the UK Data Archive are essential steps in fulfilling its strategic aims and responsibilities: it gives strategic direction both to initiate any measures which are necessary for the protection of its collections, and to meet, or extend, nationally and internationally agreed standards for the preservation of digital materials.

A preservation policy helps the Archive meet legislative and accountability requirements and its user communities' expectations. The Archive ensures that it is at the leading edge of technical advances by taking a strategic approach to long-term digital preservation, and by monitoring hardware and software developments and migrate its collections accordingly.

The Archive also aims to continually improve all aspects of the preservation-related workflow by embedding an awareness of quality in all processes.

5.1.6 Access Policy[21]

Under certain circumstances, sensitive and confidential data can be safeguarded by regulating use of or restricting access to such data, while at the same time enabling data sharing for research and educational purposes.

Data held at data centres and archives are not generally in the public domain. Their use is restricted to specific purposes after user registration. Users sign an End User Licence which has contractual force in law, in which they agree to certain conditions, such as not to disseminate any identifying or confidential information on individuals, households or organisations; and not to use the data to attempt to obtain information relating specifically to an identifiable individual.

Thus users can use data for research purposes, but cannot publish or use them in a way that would disclose people's or organisations' identities.

Controlling access to data should never be seen as the only way to protect confidentiality. Obtaining appropriate informed consent and anonymising data enable most data to be shared.

For confidential data, the Archive, in discussion with the data owner, may impose additional access controls which can be:

- needing specific authorisation from the data owner to access data

[21] See http://data-archive.ac.uk/create-manage/consent-ethics/access-control

- placing confidential data under embargo for a given period of time until confidentiality is no longer pertinent
- providing access to approved researchers only
- providing secure access to data by enabling remote analysis of confidential data but excluding the ability to download data

5.2 Technological Environment

5.2.1 IT Architecture

The Archive's collections relies on an IT infrastructure that is fit for purpose and is continually monitored and periodically reviewed to ensure timely upgrades in both hardware and software.

In order to ensure resilience and provide an adequate level of redundancy, the preservation system consists of on-site, near-site and off-site storage. For the same reasons, mirror versions of on-site systems are provided. Furthermore, to reduce risk further different operating systems will be installed across the systems.

Adequate storage capacity for all holdings is maintained. Additional unlimited capacity from external media is provided at all times.

The Archive provides necessary secure networking and communications equipment, providing adequate connectivity, the ability to restrict access to valid Mac addresses and a facility to segment the network for switched separated firewall connectivity.

All servers in the Archive are protected by power surge protection systems.

Disaster recovery procedures are in place.

5.2.2 Standards and Formats

Metadata[22]

UKDA's metadata for online data catalogues or discovery portals are structured to international standards or schemes such as Dublin Core, ISO 19115 for geographic information, Data Documentation Initiative (DDI), Metadata Encoding and Transmission Standard (METS) and General International Standard Archival Description (ISAD(G)).

The DDI is an international XML-based descriptive metadata standard for social science data used by most social science data archives in the world.

[22] See http://www.data-archive.ac.uk/create-manage/document/metadata.

The Archive use DDI to structure their catalogue records. The use of standardised records in eXtensible Mark-up Language (XML) brings key data documentation together into a single document, creating rich and structured content about the data.

The metadata record can be viewed with web browsers, can be used for extract and analysis engines and can enable field-specific searching. Disparate catalogues can be shared and interactive browsing tools can be applied. In addition, metadata can be harvested for data sharing through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

Formats[23]

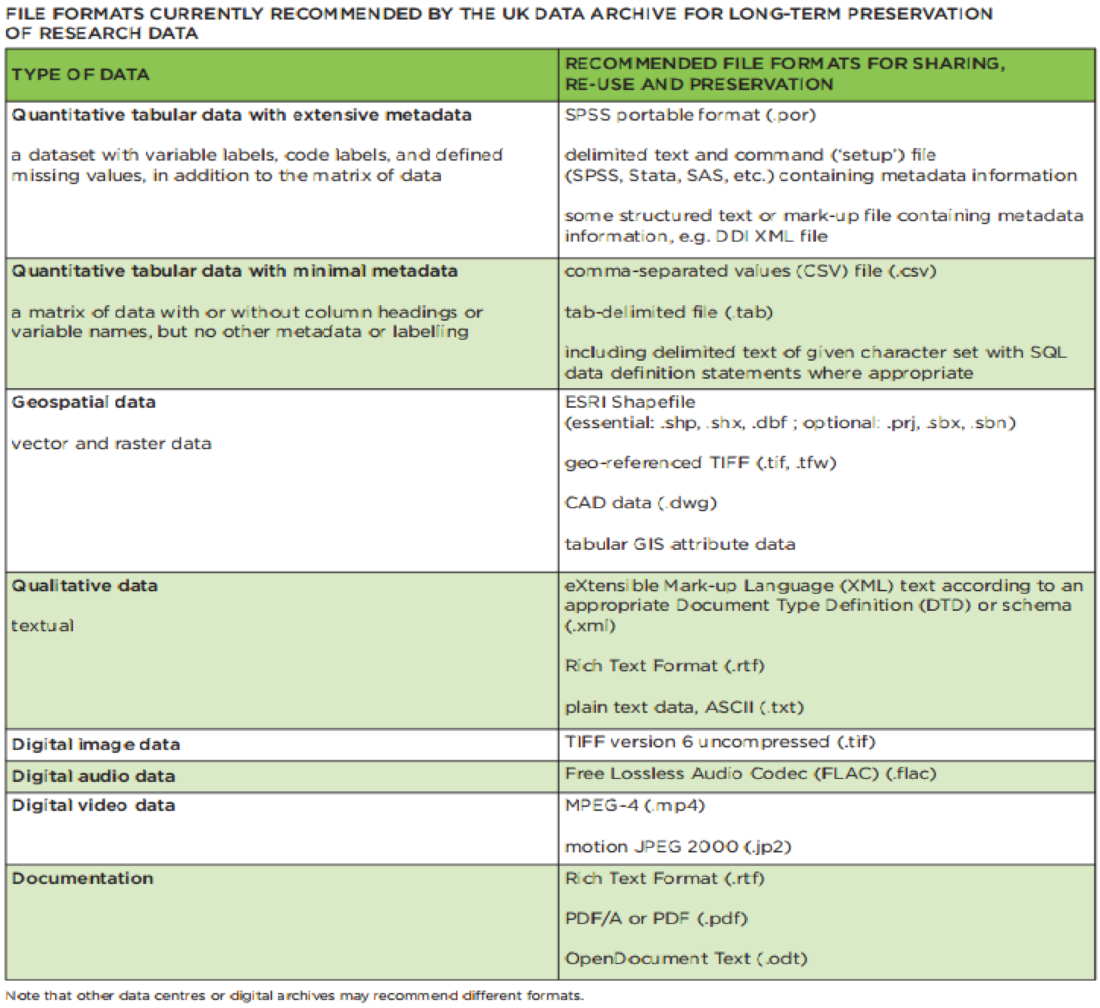
A strategy is employed to deal with the obsolescence of file formats. Appropriate information-rich preservation formats have been identified and are used in conjunction with formal documentation procedures. These formats are chosen with specific reference to the 'data types' under consideration. The Archive follows international best practice in its choice of preservation formats and data migration procedures (see table page 8).

For most electronic information it is generally possible to eliminate software dependence by sacrificing structure, but the end products of these transformations are not authentic versions of the original. In these cases the authenticity needs to be re-established through the documentation of the actions taken and validation that the substantive content has not been altered.

Thus the primary goal of the Archive's preservation policy is to ensure the long-term accessibility of electronic information while ensuring the highest level of authenticity of any formats disseminated. In effect this means that all the inherent qualities of the electronic information upon which their authenticity depends are preserved.

Defining, timing, testing and implementing migration pathways are the responsibilities of the 'Digital Preservation and Systems section'. When new formats are created from data files either through migration into new file formats or through creating new file formats for dissemination, the old files are retained.

[23] Preservation Policy (2011:11-12) and Managing and Sharing Data (2011:12)



FILE FORMATS CURRENTLY RECOMMENDED BY THE UK DATA ARCHIVE FOR LONG-TERM PRESERVATION OF RESEARCH DATA

TYPE OF DATA	RECOMMENDED FILE FORMATS FOR SHARING, RE-USE AND PRESERVATION
Quantitative tabular data with extensive metadata a dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information some structured text or mark-up file containing metadata information, e.g. DDI XML file
Quantitative tabular data with minimal metadata a matrix of data with or without column headings or variable names, but no other metadata or labelling	comma-separated values (CSV) file (.csv) tab-delimited file (.tab) including delimited text of given character set with SQL data definition statements where appropriate
Geospatial data vector and raster data	ESRI Shapefile (essential: .shp, .shx, .dbf ; optional: .prj, .sbx, .sbn) geo-referenced TIFF (.tif, .tfw) CAD data (.dwg) tabular GIS attribute data
Qualitative data textual	eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) or schema (.xml) Rich Text Format (.rtf) plain text data, ASCII (.txt)
Digital image data	TIFF version 6 uncompressed (.tif)
Digital audio data	Free Lossless Audio Codec (FLAC) (.flac)
Digital video data	MPEG-4 (.mp4) motion JPEG 2000 (.jp2)
Documentation	Rich Text Format (.rtf) PDF/A or PDF (.pdf) OpenDocument Text (.odt)

Note that other data centres or digital archives may recommend different formats.

5.2.3 Security and Risk Management/Media Monitoring and Refreshing Strategy[24]

The UK Data Archive is committed to taking all necessary precautions to ensure the physical safety and security of all data collections that it preserves:

- fire prevention and protection system;
- physical intruder prevention and detection systems;
- environmental control systems.

The repository rooms are equipped with multiple key entries and a security-protected swipe-card system linked to an on-site alarm system and to the University Security Office. The swipe-card system is maintained by the Systems and Preservation Manager, and access is restricted to

[24] Preservation Policy (2011:12-13).

three key members of staff – the Head of Digital Preservation and Systems, the Systems and Preservation Manager and the Systems and Preservation Administrator.

The repository rooms are located outside of the secure working area of the Archive.

The SSRC building in which the Archive is housed, is locked between 7.30 p.m. and 7.30 a.m. and all weekend, and is regularly patrolled by University of Essex security staff.

All machine room computer systems are locked by a logon password system to prevent unauthorised access in the case of a security breach of the room.

The Archive's suite of information and premises security are documented in our Statement of Applicability for ISO/IEC 27001: 2005. The Archive was recommended for registration to this standard in June 2010. These are detailed in the Archive's Information Security Policy, Information Security Management Policy and Premises Security Procedures.

The UK Data Archive operates a media monitoring procedure as part of its AMASS® preservation system. This allows it to check for potential future problems of wear and tear on media and act before the problems become severe.

Digital Linear Tapes (DLTs) that are used for preservation are re-tensioned every six months and each full tape in the system is also copied every year onto a new tape. This is scheduled annually and a log of all actions is kept for checking. The Systems and Preservation Administrator is responsible for performing the media refreshing procedure.

Idle tape media are automatically ejected from the DLT drives and placed in the carousel at set regular intervals to prevent excessive wear of both tapes and the drive.

The CD-R/DVD-R media are checked on a scheduled basis, every two years. If any media have either recoverable or non-recoverable errors then they are regenerated from the on-site mirror preservation server. A log is kept of all refreshment results and all storage media are provided with a date stamp indicating the time they were written and the next renewal date. CD-Rs are used within three months of purchase to ensure a short time period between when they are acquired and when they are written.

5.3 Data Curation[25]

The Archive follows the broad guidance given in the OAIS reference model.[26] The primary value to the UK Data Archive of the OAIS reference model is that it provides a framework of best practice on which its activities can be based. When, in 2005, the Archive assessed its

[25] Preservation Policy (2011:7-8).

[26] http://public.ccsds.org/publications/archive/650x0m2.pdf

conformance with the OAIS model, the main divergence between model and practice was the strict separation of Archival Information Packages (AIPs) from Dissemination Information Packages (DIPs), and there were a number of activities within the data management function relating to monitoring and management which were not appropriate to the stated objectives of the Archive.

5.3.1 Pre-Ingest Function[27]

Successful application of the pre-ingest process helps ensure that data collections are submitted at a standard which requires a lower level of processing at the ingest stage and potentially greater levels of usability through the provision of adequate documentation. Proactive pre-ingest activity also allows for various issues which might impact on preservation activities (relating to consent, confidentiality, ethics, legal issues and data formats) to be considered and addressed before data collection starts.

5.3.1.1 Information and guidance given to data producer[28]

UKDA provides a thorough guide designed to help researchers and data managers, across a wide range of research disciplines and research environments, produce highest quality research data with the greatest potential for long-term use.

Particular attention is paid to:

- whether data need to be anonymised
- whether you have permission to share research data - this may depend on consent and confidentiality agreements with participants
- whether copyright permission needs to be sought with regard to data ownership
- ensuring that data files are well documented and described to enable informed secondary use of the data
- whether data are in a condition and format suitable for sharing

5.3.2 Ingest Function[29]

Ingest is the first functional component of the OAIS reference model. It includes the receipt of information from a producer and the validation that the information supplied is uncorrupted and complete. This process also identifies the specific properties of the information which is to be preserved; it authenticates that the information is what it purports to be. The supplied version is known within the Archive as the ‘original’ version and this is retained for preservation in its original format and stored in the appropriate directory on the preservation system. This

[27] Preservation Policy (2011:8).
[28] See Managing and Sharing Data (2011)
[29] Preservation Policy (2011:8).

supplied version has a close correspondence to the Submission Information Package (SIP) in OAIS parlance.

The ingest function also transforms all elements of the deposited files into a valid preservation format for the specified data type. Files for preservation are copied to a different machine, as the ingest and preservation directory structures are created.

The Archive currently believes that the construction of a DIP during the ingest process (rather than automatically from an AIP on demand) has considerable benefits for the preservation process. This allows the Archive to reduce errors in co-operation with the producer and ensure understandability of the data. It is known that the production of multiple DIPs which are based on different software packages may lead to a loss of integrity in the underlying data. Hence the Archive deprecates the use of an existing DIP to create a new DIP. However, the Archive could carry out these tasks without generating a DIP during processing, and will investigate this in the coming year.

The ingest function also includes the creation of descriptive metadata for a variety of purposes and the production of multiple DIPs for usability.

As well as an unbroken audit trail of actions to ensure the authenticity and integrity of any data collection, the ingest process includes an element of depositor accountability whereby depositors are informed of all actions undertaken within the Archive before the data collection is released to a wider user community.

The UK Data Archive will not preserve depositor-submitted media or non-digital documentation in their original format. These will either be returned or destroyed securely.

5.3.2.1 Information and documentation from data producer[30]

Through *self-archiving*, the data producer describes their own dataset by

- completing the data collection deposit form.[31]
- prepare data and documentation according to best practice guidance on how to manage and share data

Good data documentation includes information on:

- the context of data collection: project history, aim, objectives and hypotheses

[30] Managing and Sharing Data (2011:9).

[31] http://www.esds.ac.uk/aandp/create/DataCollectionDepositForm.doc

- data collection methods: sampling, data collection process, instruments used, hardware and software used, scale and resolution, temporal and geographic coverage and secondary data sources used
- dataset structure of data files, study cases, relationships between files
- data validation, checking, proofing, cleaning and quality assurance procedures carried out
- changes made to data over time since their original creation and identification of different versions of data files
- information on access and use conditions or data confidentiality

At the data-level, documentation may include:

- names, labels and descriptions for variables, records and their values
- explanation or definition of codes and classification schemes used
- definitions of specialist terminology or acronyms used
- codes of, and reasons for, missing values
- derived data created after collection, with code, algorithm or command file
- weighting and grossing variables created
- data listing of annotations for cases, individuals or items

The data files are submitted in any of the following ways:

- via the University of Essex ZendTo Service, addressing the deposit to email account "acquisitions@essex.ac.uk" and noting study title or depositor surname in the dropbox description
- by CD/DVD/memory stick
- via secure electronic transmission - contact acquisitions@esds.ac.uk

The acquisitions team will confirm receipt of all materials associated with the data collection. After administrative checks, the data collection will be prepared for release.

5.3.2.2 Quality assurance and data checking[32]

UKDA uses different levels of quality control depending on how much 'additional value' is to be added to the data.

Quality control of data is an integral part of all research and takes place at various stages: during data collection, data entry or digitisation, and data checking. It is important to assign clear roles

[32] See Managing and Sharing Data (2011:14) and http://www.data-archive.ac.uk/curate/archive-quality for a detailed outline. The main validation and content checks for data and documentation are listed here. Further details may be found in the UK Data Archive Data Processing Standards document: http://www.data-archive.ac.uk/media/54782/ukda079-ds-dataprocessingstandards.pdf

and responsibilities for data quality assurance at all stages of research and to develop suitable procedures before data gathering starts.

During data collection, researchers must ensure that the data recorded reflect the actual facts, responses, observations and events.

Quality control measures during data collection may include:

- calibration of instruments to check the precision, bias and/or scale of measurement
- taking multiple measurements, observations or samples
- checking the truth of the record with an expert
- using standardised methods and protocols for capturing observations, alongside recording forms with clear instructions
- computer-assisted interview software to: standardise interviews, verify response consistency, route and customise questions so that only appropriate questions are asked, confirm responses against previous answers where appropriate and detect inadmissible responses

The quality of data collection methods used strongly influences data quality and documenting in detail how data are collected provides evidence of such quality.

When data are digitised, transcribed, entered in a database or spreadsheet, or coded, quality is ensured and error avoided by using standardised and consistent procedures with clear instructions. These may include:

- setting up validation rules or input masks in data entry software
- using data entry screens
- using controlled vocabularies, code lists and choice lists to minimise manual data entry
- detailed labelling of variable and record names to avoid confusion
- designing a purpose-built database structure to organise data and data files

During data checking, data are edited, cleaned, verified, cross-checked and validated.

Checking typically involves both automated and manual procedures. These may include:

- double-checking coding of observations or responses and out-of-range values
- checking data completeness
- verifying random samples of the digital data against the original data
- double entry of data
- statistical analyses such as frequencies, means, ranges or clustering to detect errors and anomalous values
- peer review

Researchers can add significant value to their data by including additional variables or parameters that widen the possible applications. Including standard parameters or generic derived variables in data files may substantially increase the potential re-use value of data and provide new avenues for research. For example, geo-referencing data may allow other researchers to add value to data more easily and apply the data in geographical information systems. Equally, sharing field notes from an interviewing project can help enrich the research context.

5.3.2.3 Data documentation and enhancement[33]

Documentation required for resource discovery and resource use is collected, created and held in the UKDA as metadata attached to each archived study (data collection). The study read file, the user guide, Depositor Licence and the study description or catalogue entry are the main components of such metadata. The study description is based on the DDI metadata standard for data documentation.

The study description is also mapped to the Dublin Core metadata standard, and the UKDA's online catalogue is Z39.50 compliant and compatible with the Open Archives Initiative (OAI) protocol for metadata harvesting. The catalogue is produced and maintained employing XML based on the DDI Data Type Definition (DTD). Resource discovery in the catalogue is further enhanced by the use of the Humanities And Social Science Electronic Thesaurus (HASSET) that was created and is maintained by the UKDA. The Information Development team interprets national and international standards for local implementation in its resource discovery metadata.

For studies processed in SPSS format (the vast majority) processing procedures consist of generating and archiving a UKDA data dictionary, a rich text format document created by UKDA software, giving more detail than the data dictionary generated by SPSS. A file logging any unavoidable inconsistencies between the data in SPSS and STATA formats is also generated and archived for all applicable studies. This is supplied to anyone ordering the data in STATA format, who can then locate the lost information in the data dictionary. For the small number of studies processed in other formats, equivalent information is recorded (e.g., MS Access data documenter output).

Data are always preserved in a preservation format in addition to the proprietary formats used for dissemination. Preservation formats consist of tagged or delimited text of a given character set (ASCII or UNICODE) or eXtensible Markup Language (XML). Where necessary, data definition statements (in SPSS, STATA, SAS, SQL or Visual Basic command languages as appropriate) are

[33] See Assessment of UKDA and TNA Compliance with OAIS and METS Standards (2004)

also preserved, to preserve the full information of the dataset (variable formats, variable labels, code labels, missing value definitions, etc.).

5.3.3 Archival Storage and Preservation[34]

The purpose of archival storage is to ensure that what is passed to it from the ingest process remains identical and accessible. In the Archive this function receives AIPs and DIPs from the ingest function and adds them to the permanent storage facility, oversees the management of this storage, including media refreshment and monitoring. This function is also responsible for ensuring that AIPs can be retrieved.

Due to a combination of factors relating to information security, access conditions and usability, the Archive has elected to prepare DIPs as part of the ingest process. These DIPs are accessible to users through a 'Provide Data Function' and they are also stored alongside the AIPs in the preservation system. Thus for any data collection there are always at least three versions residing on each of the different preservation systems: the original SIP, the ingested AIP and multiple DIPs. When a new version of a DIP is created it must be created from the AIP (or possibly the SIP if it is understood to be a de facto subset of the AIP) but not from an earlier recension of a DIP.

5.3.3.1 Physical data preservation and storage[35]

In order to best safeguard long-term preservation, the Archive follows a policy of multiple copy resilience. Five versions of the complete preservation system are held: main near-line copy (on the main preservation server) and a shadow copy (on main preservation server). Both are held on the main area on the Hierarchical Storage Management (HSM) system and are presently accessed only by the dedicated preservation user. The storage media used for this copy are SDLT and disc cache area. The access online copy (on the mirror preservation server) is held in a RAID 5 disc system and copies are generated for user access and dissemination. There are also a near-site online copy kept on a RAID 5 disc system on a server located in another building within the University of Essex, and an off-site online copy. Finally a disc-based offline copy exists, which is held in either DVD-R or CD-R copy.

The Archive follows best practice in the storage and housing of magnetic and optical media. In particular, for environmental conditions for storage media (BS 4783, ISO/IEC22051, BS ISO 18921:2002 and BS ISO 18925:2002) and for the storage of archival materials (BS 5454).

[34] Preservation Policy (2011:8).

[35] Preservation Policy (2011:8-9).

5.3.3.2 Preservation strategy[36]

The UK Data Archive has implemented a preservation strategy based upon open and available file formats, data migration and media refreshment. Preservation decisions at the Archive must always be made within the context of its Collections Development Policy[37], balancing the constraints of cost, scholarly and historical value, and user accessibility alongside the requirements of levels of authenticity and legal admissibility. Hence, different ingest processes may be required for material with different levels of quality and significance.

Every study within the Archive follows a consistent directory structure for storage, and this is enforced by automated checks when files are copied onto the preservation system. This has many benefits, such as the ability to locate set types of information and also to allow automated tasks (e.g. migration of file formats) to be run without the need for complicated locator scripts. In addition to this structure, file label details are kept in an in-house system to provide extra information about a file in addition to its filename. Further, file extensions are always standardised, with a single extension allowable for each type of file.

The complete chain of custody of all data collections is documented through metadata. All actions are explicit, complete, correct and current. However, only the 'original' version is an integral copy of the version deposited with the Archive. The preservation and dissemination versions are considered to be authentic and there is an audit trail of all alterations in the preservation and dissemination versions which relates back to the original deposited version.

UKDA's preservation policy is also designed to promote preservation as an integral part of the management of the Archive's collections and to ensure best use of resources by providing a framework for managing the preservation procedures. The specific aims of the preservation policy are to:

- provide authentic, reliable instances of data collections to the designated user community;
- be a trusted repository within the generally accepted scope of the term;
- maintain the integrity and quality of the data collections;
- ensure that digital resources are managed throughout their lifecycle in the medium that is most appropriate for the task they perform;
- ensure that all data collections are protected;
- ensure that the relevant level of information security is applied to each data collection;
- instil good practice in active preservation management;
- improve the speed and efficiency with which information is preserved and retrieved;

[36] Preservation Policy (2011:11-12).

[37] http://www.esds.ac.uk/news/publications/UKDACollectionsDevPolicy.pdf

- develop and maintain systems of low-cost storage, with appropriate location and with regular review;
- optimise the use of the Archive's space for storage purposes.

The concept of preservation level has recently been introduced to the Archive. The majority of data currently archived at the Archive do not rely on presentation characteristics to ensure they are understandable.

5.3.3.3 Version control/change procedures[38]

Ensuring that any alteration to the preserved version of any part of a data collection is accurately documented is integral to the authenticity of any data collection. The Archive distinguishes between two forms of alteration post ingest.:

- new version (Definition: when there is a change to the preserved metadata);
- new edition (Definition: when there is change to data or documentation).

When there is a new version of a data collection, the relevant descriptive and structural metadata must be revised and the old file retained.

When there is a new edition of a data collection, all descriptive and structural metadata must be recreated, and the old file and the previous AIP and DIPs retained within the preservation system and identified as not for issue.

5.3.4 Dissemination

5.3.4.1 Visibility[39]

UKDA provides access to over 5,000 digital data collections for research and teaching purposes covering an extensive range of key economic and social data, both quantitative and qualitative, and spanning many disciplines and themes.

Using their online Data Catalogue you can search and browse the data collections. For each collection they provide a detailed catalogue record and links to contextual information, such as user guides, questionnaires, technical reports and so on.

5.3.4.2 Availability and accessibility[40]

The service provides access to over 5,000 computer-readable datasets for research, learning and teaching purposes for many different disciplines.

[38] Preservation Policy (2011:10).

[39] http://www.esds.ac.uk/findingData/findintro.asp

[40] http://www.esds.ac.uk/findingData/aboutCat.asp?print=1 and http://www.jisc.ac.uk/media/documents/programmes/preservation/oaismets.pdf

Types of data available:

quantitative

- microdata are the coded numerical responses to surveys with a separate record for each individual respondent
- macrodata are aggregate figures, for example country-level economic indicators
- data formats include SPSS, Stata and tab-delimited formats

qualitative

- data include in-depth interviews, diaries, anthropological field notes and the complete answers to survey questions
- data formats include Excel, Word and Rich Text Format (RTF)

multimedia

- a small number of datasets may include image files, such as photographs, and audio clips

non-digital material

- paper media could include photographs, reports, questionnaires and transcriptions
- analogue audio or audio-visual recordings

The catalogue also contains historical data from the History Data Service and UK census data available via Census.ac.uk.

Note that the large majority of data are fully anonymised, unless otherwise specified in the relevant online catalogue records, and are therefore not suitable for genealogical users or family historians.Using their online Data Catalogue you can search and browse our data collections. For each collection they provide, at no cost, a detailed catalogue record and links to contextual information, such as user guides, questionnaires, technical reports and so on.

Their search tools enable you to search by variables (for survey data), major studies, latest releases or use the UK Data Archive's HASSET thesaurus. They can also help identify data located in other archives in the UK and from other data archives around the world.

A link to access the data can be found in each catalogue record. You will be required to register to access the data, using federated access management (shibboleth) authentication[41]. To

[41] http://shibboleth.net/

analyse data, users usually need to have access to appropriate data analysis software, although a range of data are available to explore, tabulate and chart online.

Access to the Data Catalogue, including online documentation such as questionnaires, does not require registration. However, to download any data you must register with ESDS, agree to an End User Licence (EUL) and provide details of your intended use. Access restrictions may apply to some users/usages and details can be found in the relevant catalogue record.

The UKDA distributes and provides access to its data collections via:

- HTTP download;
- online access;
- guest FTP;
- CD-R and DVD-R;
- other media by special request (e.g., DAT, Zip disc).

The UKDA's HTTP-based download service provides a quick and reliable means of gaining access to the most heavily used collections held at the archive. The archive also provides online access to data that have been enhanced and published in the Nesstar system. Nesstar provides the capability for data discovery, browsing, subsetting, visualisation and downloading via the Internet. The system is based on the DDI metadata standard.

Users of the archive must be registered to order data but can browse the catalogue without being registered.

All users can:

- view various levels of the metadata (bibliographic record, abstract, documentation, user guide, variable list, data dictionary);
- download or browse the documentation;
- access frequency counts for selected data (via Explore Online/Nesstar links).

Registered users can:

- browse, analyse or download the data for a growing number of the UKDA's most popular data series;
- order the data and documentation in a variety of software formats and media.

For further information on access to UKDA data collections see the Using data, Ordering/Downloading data and Nesstar documents on the UKDA web site.

5.3.4.3 Tools and interfaces[42]

ESDS has produced a series of user guides for helping get started with ESDS's online data browsing tools:

ESDS Government:

ESDS Government has an interactive web-based geographical information system, CommonGIS, with a data exploration and visualisation interface for England and Wales Vital Statistics on fertility and mortality for 2001.

ESDS International macro data:

Access to international macro-economic time series data is available via Beyond 20/20 on the ESDS International website. The Beyond 20/20 Web Data Server (WDS) can be used to display, subset and download data, runs on a standard web browser and is accessibility compliant. Available to users in UK HE/FE only.

ESDS International micro data:

A number of cross-national surveys can be browsed and analysed online, including:

- European Social Survey
- European Values Study
- World Values Surveys
- Latinobarómetro
- Eurobarometer Survey Series, Candidate Countries Eurobarometers, Central and Eastern Eurobarometers and the International Social Survey Programme via ZACAT - GESIS Online Study Catalogue

ESDS Nesstar Catalogue:

Use the Explore online links/icons via the Data Catalogue or Major studies pages to access the ESDS Nesstar Catalogue to view variable frequencies, conduct simple online tabulations, produce graphs and subsets of data. Users of the Mac Operating System are advised to use the following web browsers: Firefox, Mozilla or Safari.

ESDS Qualidata Online:

Search, browse and download qualitative data. Data collections can be explored in many ways, including interview summaries for each individual. A number of interview transcripts have also been fully digitised and it is possible to carry out keyword searches on these transcripts.

[42] http://www.esds.ac.uk/orderingdata/exploreOnline.asp?print=1

Census data:

To access data made available under the ESRC Census Programme use the Access via Census Dissemination Unit, Access via Cathie Marsh Centre for Census and Survey Research and Access via Census Geography Data Unit (UKBORDERS) links from the Data Catalogue results web pages or the links at Census.ac.uk. Available to users at UK higher or further education institutions only.

5.3.4.4 Monitoring, review and feedback[43]

Submissions to the UKDA are reviewed and approved by the Acquisitions Review Committee (ARC) which meets fortnightly and reviews all new acquisitions or submissions that have been received in the period since the last meeting. Membership of ARC comprises representatives from both the Outreach and Training, and Data and Support Services sections. The committee is chaired by the Associate Director, Outreach and Training. When the need arises, expertise is sought from the relevant Advisory Board or relevant subject experts. The ARC co-ordinates and develops all management policies and issues that relate to the Acquisitions functions of the UKDA. The Acquisitions Review Process has its own Processing Guide and a work plan is created for each ‘accepted’ submission. This plan specifies:

- that all files will be preserved in their original format;
- that all files will be converted to the appropriate preservation format, if necessary;
- the additional data formats or versions in which the data and documentation will be made available;
- the composition of the user guide for each resource;
- the level of validation, cataloguing and indexing, and additional documentation that needs to be created, if any.

At the submission stage, there is no formally signed Submission Agreement although depositors do supply a data submission form. Instead, on acceptance by the ARC, a legal agreement, the Licence form, is drawn up, by which the data are deposited formally whilst the depositor retains legal ownership of the dataset. This is an instance where the UKDA differs from the OAIS reference model. The licence gives the UKDA the right to process the data for preservation and dissemination whilst also keeping the original deposited dataset intact. Data transfer and submission are both handled by the acquisitions staff within the Outreach and Training Section. This section also works closely with data depositors to provide guidance and advice with regards to data creation and deposit. ee also the workflow diagrams at Ingest entity above.

[43] http://www.jisc.ac.uk/media/documents/programmes/preservation/oaismets.pdf

The UKDA runs a series of service specific customer help desks which accept queries by email, post, telephone and fax and which provide guidance to customers on finding, accessing and using data, as well as user support, registration support and user training. The help desks also manage the receipt and throughput of orders.

Sources

Documents

Reference Model For An Open Archival Information System (OAIS) (2006)

Preservation Policy (2011)

Managing and Sharing Data – UK Data Archive (2011)

UK DATA Archive Collections Development Policy (2008)

Operational Selection Policy OSP30 (2006)

Assessment of UKDA and TNA Compliance with OAIS and METS Standards (2004)

E-sources

Access Control:

http://data-archive.ac.uk/create-manage/consent-ethics/access-control

Metadata:

http://www.data-archive.ac.uk/create-manage/document/metadata

Data Deposit Form:

http://www.esds.ac.uk/aandp/create/DataCollectionDepositForm.doc

Finding Data Overview:

http://www.esds.ac.uk/findingData/findintro.asp

About the Data Catalogue:

http://www.esds.ac.uk/findingData/aboutCat.asp?print=1

Online Data Analysis:

http://www.esds.ac.uk/orderingdata/exploreOnline.asp?print=1

ESDS Economic Impact Evaluation

http://www.esrc.ac.uk/_images/ESDS_Economic_Impact_Evaluation_tcm8-22229.pdf

Shibboleth:

http://shibboleth.net/

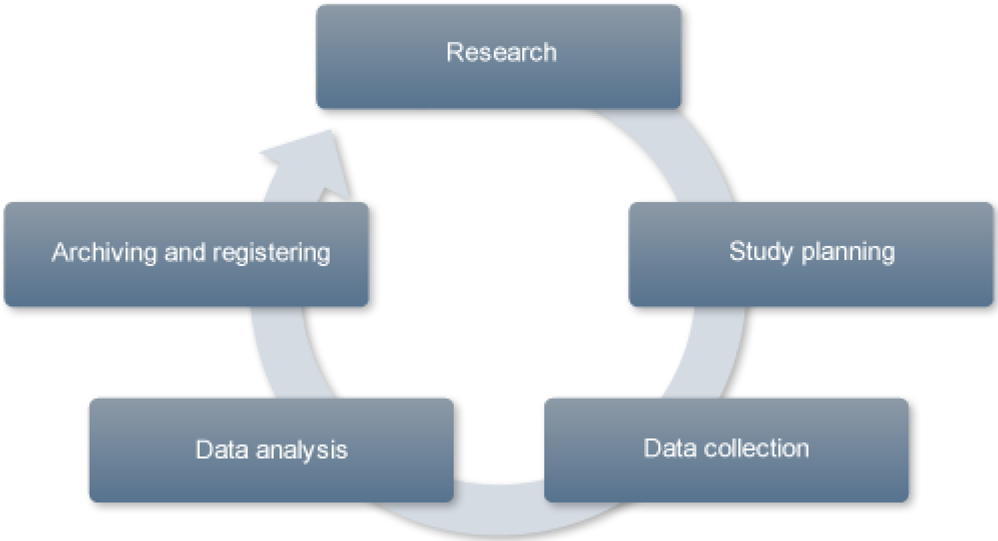
6 GESIS Data Archive[44]

6.1 Organizational Framework

The Data Archive was originally founded in 1960 as Central Archive for Empirical Social Research (Zentralarchiv für empirische Sozialforschung, ZA), Europe's first data archive in the Social Sciences. In 1986 it became a member of the newly founded Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen (GESIS), a collaboration of three independent Social Sciences infrastructure institutions. Since 2007, the Data Archive is one of five scientific departments of GESIS – Leibniz-Institute for the Social Sciences, Germany's biggest research-based Social Sciences infrastructure institution. GESIS on the one hand offers **services** tailored to the needs of the Social Sciences community and aligned with relevant thematic and structural developments in the field. These services are informed by the results of GESIS's research as well as its close co-operation with universities and other partners.

Services offered by GESIS are organized according to the research data cycle (see fig. 1). See Table 1 for examples of services offered in each segment.

FIGURE 1: RESEARCH DATA CYCLE



[44] This segment was provided by Dr. Astrid Recker, GESIS.

Table 1: Examples for services offered

Research: searching for data and information	E.g. online data catalogue, SSOAR - open access repository, sowiport - Social Sciences information portal, Research Data Centers, Data Catalog (DBK), etc.
Study planning and design	E.g. consultation on sample design, consultation on online surveys, general project support with regard to survey instruments, data collection.
Data collection and preparation	E.g. cognitive pre-testing-lab, advice and support on telephone sampling, occupation coding, text and content analysis.
Data analysis	E.g. provision of data from the Data Archive for analyses; consultation and support for data analysis.
Archiving and registration of research data	Services of the Data Archive and da|ra, the DOI registration agency.

See http://www.gesis.org/en/services/ for more detailed information.

On the other hand, GESIS conducts **research** in three large subject areas

- Survey Methodology,
- Social Structure, Attitude and Behavior in Modern Societies,
- Applied Computer and Information Science.

As a member of the Leibniz Association, the umbrella organization to currently 87 research institutions which “conduct research and provide infrastructure for science and research and perform research-based services – liaison, consultation, transfer – for the public, policy-makers, academia and business” (http://www.research-in-germany.de/leibniz), GESIS is part of a strong network of publicly funded research institutions. In addition, GESIS has long-lasting and strong ties with universities.

GESIS actively promotes stronger networking of the German research infrastructure with institutions in Europe and worldwide. Accordingly, the Data Archive is a member of CESSDA (Council of European Social Science Data Archives, http://www.cessda.org) and contributes to international data service networks (International Federation of Data Organizations for the Social Sciences, IFDO, http://ec2-50-17-181-92.compute-1.amazonaws.com/). It also acts as the national representative of the Inter-University Consortium for Political and Social Research (ICPSR, http://www.icpsr.umich.edu/) in Germany.

6.1.1 Purpose and Requirements

6.1.1.1 Scope and objectives

The guiding principle of all activities of the Data Archive is to make possible and support reproducible and intersubjectively verifiable empirical research as well as the re-use of research data for novel research questions and methods, or for the purpose of ex-post comparative or historical research.

Accordingly, key activities center around providing data services for national and international comparative surveys from the fields of social and political science research and making the resulting high-quality data accessible for re-use. The surveys are archived and processed according to internationally recognized standards – including metadata and classification standards such as DDI or ISCED as well as accepted standard procedures of data management and processing (see Jensen 2012, appendices A2-A4 for an overview of relevant standards) – and made accessible to the scientifically interested public.

The Data Archive currently provides access to over 5,100 national and international studies (equaling nearly 600,000 digital objects), including those listed in Table 2.

Table 2: Selected holdings of GESIS

Survey Programs, (esp. intercultural comparative research, national survey programs)	• ALLBUS • National Election Studies (GLES, Landtagswahlen, Politbarometer etc.) • International Social Survey Programme – ISSP • Eurobarometer • European Values Study – EVS • International Election Studies (CSES, PIREDEU)
Historical Studies	Heterogeneous holdings ranging from • studies based on historical text documents (e.g. church registers, written documents of finance departments, court documents) and official statistics, • to collective biographies (e.g. studies on parliamentarians' biographies), • or studies with quantitative time series data (e.g. the historical statistics of the official statistical sources for the German Empire and its member states).
Comprehensive holdings on the topics of	• Political attitudes and behavior • Values • Health • Youth • Media use
Official Microdata (access provided by GESIS)	• European Microdata • Microcensus • Income and Expenditure Survey • GDR Microdata

Reference studies and surveys for which there is a high demand are combined into 'study collections' accompanied by value-added services (e.g. harmonization and standardization of variables in order to facilitate comparisons across time or regional units) – frequently in co-operation with the primary researchers or experts. In addition, four Research Data Centers have been established in cooperation with the department "Social Monitoring and Social Change" for programs where GESIS is involved in data collection or is responsible for data preparation, archiving and distribution: for the ALLBUS, for selected International Survey Programs, for the extensive series of studies in electoral research (Election Studies) as well as for microdata from official statistics (German Microdata Lab). All RDCs meet the criteria of the German Council for Social and Economic Data (RatSWD; cf. http://www.ratswd.de/eng/dat/fdz.html).

Further key activities of the Data Archive include:

- implementation of a secure data center to enable and promote access to more sensitive or confidential data in accordance with legal regulations;
- implementation of Datorium (see Table 3), a data sharing platform (based on DSpace) as a low-threshold opportunity for individual researchers and projects to archive and share their data in compliance with funders' requirements (see Wira-Alam, Dimitrov, and Zenk-Möltgen 2012);
- da|ra (http://www.da-ra.de/en/), the DOI registration service for social science and economic data in Germany, offered in cooperation with DataCite, the international initiative to establish easier access to digital research data, and ZBW - Leibniz Information Centre for Economics;
- Archive and Data Management Training Center (http://www.gesis.org/en/admtc), providing training and consultancy in research data management and digital preservation of research data;
- development of software for data processing, archiving, and analysis as well as for the online-search of data and variables (in cooperation with the GESIS department "Knowledge Technologies for the Social Sciences");
- involvement in the formation of international standards (e.g. DDI, standards for data citation);
- participation in national and international research and development projects.

TABLE 3: DATORIUM (AVAILABLE FROM EARLY 2013)

Objective	• to further promote and facilitate a culture of data sharing, enable individual researchers and small projects to comply with funders' requirements and to easily deposit and share their data • to significantly increase the archiving and sharing of small to medium sized studies in the

	Social Sciences
Services	• depositors are provided with an easy-to-use upload form allowing for the assignment of simple metadata and determining access/licensing conditions based on a simple scheme • review of submitted data by the Data Archive according to archive standards; subsequent publication in Datorium • DOI registration • tiered preservation services (based on bitstream preservation; option to upgrade to full digital preservation including format migration)
Target audience	• individual researchers • small to medium sized projects and studies

6.1.1.2 Collection policy

According to the GESIS by-laws, GESIS's service and data offers are to enable and support research of social developments in national and international comparative and historical perspective. According to the Data Archive's acquisition policy, data are therefore not collected with a tight-knit thematic focus, but rather with the objective of building a broad collection of data of general relevance to Social Sciences research. Thus, while there are certain thematic areas in which the Data Archive has built particularly comprehensive holdings (see Table 2), the collection policy does not per se limit the thematic focus of data to certain areas. No national or legal regulations exist as of yet making the submission of data to the Archive mandatory.

6.1.1.3 Criteria for evaluating data

The Data Archive acquires data both actively (i.e. by contacting relevant projects, surveys, or programs) and passively (i.e. by addressing potential depositors through the website or publicity materials). Data are evaluated by means of criteria meant to determine their relevance and their suitability for archiving by GESIS. In particular, the following aspects are considered:

- Can the data be used to answer questions relevant to the Social Sciences?
- Are data well prepared and well documented?
- Is it quantitative data relevant to a broad circle of potential users?
- Is the study method or the content unique?
- Can the data be used for international or historical comparative research?

6.1.2 Legal and Regulatory Framework

With regard to privacy issues, the handling and processing of data is governed by the Federal Data Protection Act (Bundesdatenschutzgesetz, BDSG) in combination with the Data Protection Acts of the German federal states ("Länder"). Questions of intellectual property rights are regulated by the German Act on Copyright and Related Rights (Gesetz über Urheberrecht und verwandte Schutzrechte, UrhG) (see Centre for Intellectual Property Law (CIER), 2011; Jensen, 2012; and Büttner, Hobohm, and Müller, 2011 for discussions of the legal status of research data in Germany).

To enable the Data Archive to preserve and offer data for re-use, data producers sign a deposit agreement when submitting data for archiving. According to the deposit agreement, the Data Archive may

- archive all data and documentation and process them further for the purpose of long-term preservation and re-use. For this purpose, the archive may use the technical means, formats and methods it deems most suitable;
- digitize all texts belonging to the study (if they are not in digital format already), and make them available (along with the data) via the archive webpages according to the archive access policy and categories (see below);
- publish metadata for the study.

For the stated purposes, the archive receives all necessary rights of (non-exclusive) use as laid down in German copyright law (especially §§16 and 19 UrhG). Thus the Data Archive receives permission from the data producers to carry out long-term preservation action, e.g. migration to a different file format, as well as making several copies of the data and their documentation for backup and distribution.

The following standard licensing conditions are available to data producers:

- Category 0: Data and documents are released for everybody.
- Category A: Data and documents are released for academic research and teaching.
- Category B: Data and documents are released for academic research and teaching if the results won't be published. If any publications or any further work on the results is planned, permission must be obtained from the Data Depositor.
- Category C: Data and documents are only released for academic research and teaching after the data depositor's written authorization. (http://www.gesis.org/en/services/data-analysis/data-archive-service/usage-regulations/#3_Access_categories)

Use of the data made accessible by the Data Archive is governed by usage regulations. These make it a requirement, for example, to inform the Data Archive about the completion of the project for which the material was used, to delete the data (and the medium carrying them) after completion of the project, and to quote all used documents according to scientific conventions and to send two specimen copies of his/her publication to the Data Archive (see http://www.gesis.org/en/services/data-analysis/data-archive-service/usage-regulations/).

6.1.3 Funding and Resource Planning

GESIS is legally registered as a non-profit association and is sponsored jointly by the federal government and the federal states according to Article 91b of the German Federal Constitution. GESIS generates additional revenues by offering value-added services to the Social Sciences community (cf. GESIS By-laws, § 15). Funding in accordance with Article 91b is assigned in seven-year cycles.

Like GESIS, the Data Archive offers its central services – including online access to and deposit of data – free-of-charge. There is a handling-charge if users request data that require customization (e.g. compilation on a CD-ROM or customization of data and documentation according to user needs; cf. http://www.gesis.org/en/services/data-analysis/data-archive-service/charges/).

6.1.4 Long-Term Preservation Policy

According to GESIS's by-laws, among the association's primary objectives is the "archiving, documentation, and long-term preservation of Social Sciences data, including the indexing of data as well as the high-quality enhancement of particularly relevant data to prepare them for re-use" (§ 2; my translation). Thus, GESIS voices its commitment to preserve and provide access to Social Sciences research data in its by-laws and mission statement. The Data Archive's preservation principles and practices have also been communicated in contributions to relevant publications (e.g. Neuroth et al. 2012; Büttner, Hobohm, and Müller 2011; Jensen 2012).

Seeking to demonstrate its status as a trusted digital archive, and to document and communicate its preservation principles and activities more transparently within a formal and standardized framework, the Data Archive has entered a certification process. In this process, obtaining the Data Seal of Approval will be the first step. In this context, a position for a digital preservation specialist has been created, who will – among other things – also combine existing publications and internal documentation into a published preservation policy. This policy will express the Data Archive's adherence to the following core principles:

- sufficient documentation of all processes relevant to the preservation of the archive's holdings, including changes made to data in order to guarantee their future usability (e.g. migration);
- employing monitoring processes to ensure an active preservation management;
- providing users and data depositors with sufficient information to make guidelines and procedures of the Data Archive transparent;
- interacting with and contributing to Social Sciences and digital preservation communities and initiatives on a national and international level, e.g. with regard to the development and implementation of standards;

- referencing and implementing established standards and models in the field of digital preservation (e.g. OAIS) and metadata (e.g. DDI); this includes the use of persistent identifiers (DOIs) for all archived studies.

6.1.5 Access Policy

GESIS's offers are directed primarily at researchers (both in universities and non-university research institutions) and students – in particular in empirical social research with a focus on the areas of sociology and political science as well as social science in its entirety. Other target groups include those working in related political, social and commercial social science environments. The Data Archive strongly promotes data sharing and re-use and hence seeks to make data available as openly and easily accessible as possible. However, legal regulations and respect for the needs and requirements of data depositors make it necessary to manage access according to the principles outlined in Legal and Regulatory Framework.

If not indicated differently, the Data Archive makes data and documents available for scientific analysis carried out in academic research and teaching according to the stated access categories. Institutes and individuals outside academic research and teaching can apply for provision in written form.

6.2 Technological Environment

6.2.1 IT Architecture

The IT Architecture of the GESIS Data Archive combines secure hardware with a number of different – partly open source – software applications for object management and access (see Figure 2).

On the **storage layer**, the Data Archive uses a file-based rather than a database approach. Among the advantages of this approach are the capacity to store Archival Information Packages as complete "physical" entities (i.e. packages) and easy to realize back-up procedures on file level. Thus, Submission, Archival, and Dissemination Information Packages (SIPs, AIPs, DIPs,) are saved in a hierarchical directory structure on a secure archive server. The directory organization is designed to mirror the data life cycle (e.g. by "bundling" together originally submitted data and documents, normalized and processed data/documentation for archiving, and data/documentation for access) and to express the technical and logical relationships between the archived objects. The process of depositing Information Packages to Archival Storage is governed by strict rules and conventions (e.g. for file names and directory structures) which are laid down in documentation available to the Archive staff. Dissemination Information Packages are mirrored to the access server on a daily basis.

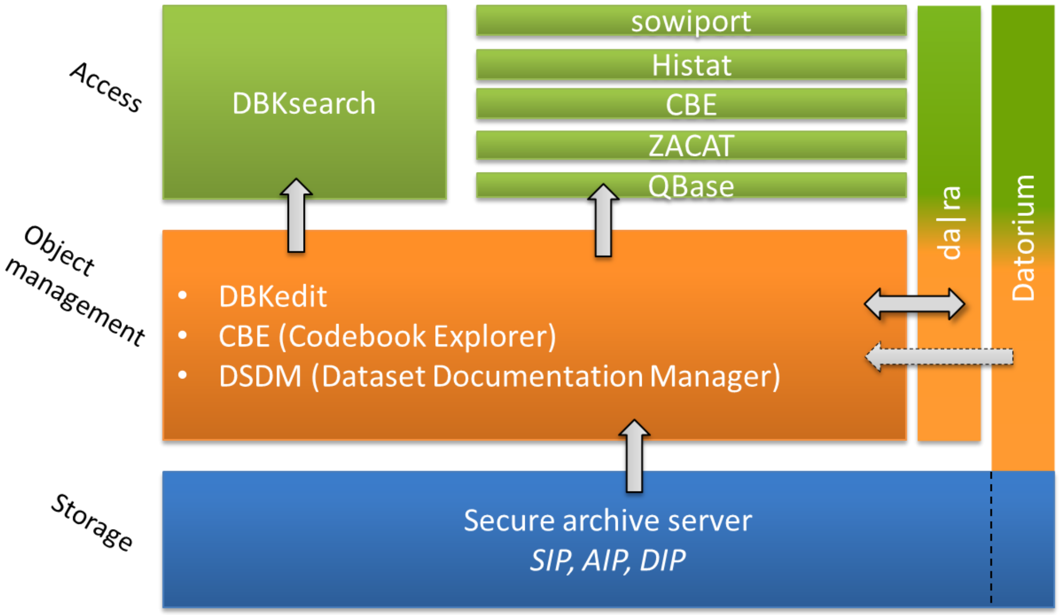
In the **object management layer**, different software is used depending on the depth with which a given study is indexed and documented (standard vs. value-added documentation, see below).

For all studies, metadata are added and administered with the software DBKedit, which is made available for re-use under a GNU General Public License (http://www.ddialliance.org/node/861). For added-value documentation, the tools Dataset Documentation Manager (DSDM; http://www.ddialliance.org/node/856) and Codebook Explorer (CBE; http://www.gesis.org/en/services/research/codebookexplorer/) are used (see Luijkx, Brislinger, and Zenk-Möltgen 2003).

In the **access layer**, users can perform metadata-guided searches for (and within) studies as well as explore and analyze data on variable level using four platforms: DBKsearch, ZACAT, Qbase, and CodebookExplorer, which is provided as a standalone offline tool (see Tools and interfaces for a more detailed description of the scope and functionality of each platform). Time series data from historical Social Sciences and economic research can be accessed through the Histat portal. Additional entry points are the search interface offered on the webpage of the DOI-registration agency da|ra (http://www.da-ra.de) and the sowiport search engine (http://www.sowiport.de).[45]

The DSpace installation Datorium is currently being integrated into the existing architecture and will be functional on all three layers in early 2013.

FIGURE 2: OVERVIEW OF IT ARCHITECTURE



[45] Note that both services contain not only metadata for gesis holdings, but also for holdings of other repositories and archives.

6.2.2 Standards and Formats

There are no principal restrictions as to which **file formats** are accepted by the Data Archive. Currently, data are mostly submitted in standard (proprietary) statistical file formats, predominantly Stata or SPSS. Relevant documentation is most commonly submitted as PDF or other text documents. If documentation or complementary material is submitted in print format – which happens occasionally – documents are scanned and saved as TIFF files.[46]

However, most data and documentation are received electronically by email or on physical storage media such as DVDs. In the future, an upload of data through an online submission form collecting basic metadata will be implemented as part of Datorium.

The **archival format** currently used is SPSS portable. Although this is not an open standard, SPSS is a de facto standard in the Social Sciences, which can be expected to be used by the Data Archive's designated community in the future. At the same time, SPSS portable allows for sufficient compatibility between different operating systems and SPSS versions to be suitable as an archival format. However, the Data Archive is aware of the advantages of open formats in the context of digital preservation and accordingly the future use of SPSS portable as an archival format is currently subject to discussion; thus, a decision may be made in the future to use a different, more suitable format (e.g. ASCII).

The standard **dissemination formats** used are SPSS system file, SPSS Portable, and Stata.

The **metadata** used to describe studies for the Data Catalog (DBK) are compliant with the DDI standard (http://www.ddialliance.org) – that is, DDI2 or DD3 compliant metadata is available for download from DBK in XML format (Zenk-Möltgen and Habbel 2012). The indexing and documentation of data in studies/study collections is carried out using national and international standards, and classifications where necessary and relevant. These include:

- ISO 3166 - Codes for the representation of names of countries and their subdivisions
- ISO 639 - Codes for the representation of names of languages
- ISCO - International Standard Classification of Occupations
- NUTS - Nomenclature of territorial units for statistics
- Description of study content
    - ZA Classification (http://www.gesis.org/unser-angebot/recherchieren/thesauri-und-klassifikationen/zentralarchiv-klassifikationsschema-za-klassifikation/#c28447)

[46] The paper original is also retained.

- CESSDA Topic Classification (http://www.gesis.org/unser-angebot/recherchieren/thesauri-und-klassifikationen/cessda-topic-classification/#c28449).
- In addition, the European Language Social Science Thesaurus (ELSST; http://www.gesis.org/unser-angebot/recherchieren/thesauri-und-klassifikationen/european-language-social-science-thesaurus-elsst/#c28451) makes the content of the ZACAT Online Catalog browsable by keyword in the CESSDA portal.

To enable persistent identification, the GESIS Data Archive assigns DOIs to all of its studies via the da|ra service.

6.2.3 Security and Risk Management / Media Monitoring and Refreshing Strategy

The security and risk management is carried out in close co-operation with GESIS's IT department, which administers the servers and takes care of backups, media monitoring and refreshing. These procedures are governed by an IT security concept which is frequently reviewed.

To protect the data, the following backup and access control procedures are in place to guarantee the (physical) safety of the digital archive holdings:

1) Physical protection measures:
   a. The computing center and server rooms are secured against unauthorized access by means of an electronic access control system.
   b. Smoke and water detectors are in place, temperatures in the computing center are monitored.
2) Redundant data storage in different locations:
   a. Frequent (up to daily) incremental and complete back-ups to onsite disk and tape libraries (tapes stored in suitable vault). In addition, frequent back-ups to offsite tape libraries.
3) Diversity of storage media (hard disk, tape) and frequent media refreshment.

The backup and storage procedures (redundant and distributed storage) in place allow for fast and complete recovery / restoral of the archive holdings in case of a disaster.

6.3 Data Curation

6.3.1 Pre-Ingest Function

6.3.1.1 Information and guidance given to data producer

Pre-Ingest communication with data producers adheres to important steps of the Preliminary and Formal Definition Phases as outlined in the PAIMAS standard (CCSDS 2004).

The Preliminary Phase is designed to gather necessary information, establish contact with data producers, and to provide the data producers with all necessary information – in particular about the archiving services offered by GESIS, the minimal requirements that have to be fulfilled for GESIS to accept data for archiving (see Information and documentation from data producer), as well as licensing options and the deposit agreement.

During the Formal Definition Phase, agreements and schedules sketched before are discussed in more detail and on a more concrete level. In this phase, subject-related, technical, and legal questions are clarified, and agreements made with regard to delivery formats, schedules, etc. This phase concludes with the signing of the deposit agreement (see Legal and Regulatory Framework).

6.3.2 Ingest Function

6.3.2.1 Information and documentation from data producer

The Data Archive requests data producers to submit all materials necessary for a secondary analysis. This includes at least

- information about the primary researcher(s) and title of the study
- the data itself, prepared for direct use in statistical software packages if possible,
- the instrument or instruments used for data collection (e.g. questionnaire),
- a methodological description of the data collection and preparation procedures.

No materials are accepted that are subject to any copyright restrictions which may interfere with the use of the data as outlined in the deposit agreement (e.g. copies of complete books).

6.3.2.2 Quality assurance and data checking

After a submission has been received, the data and all accompanying material are assessed with regard to content, structure and format:

- Completeness and consistency of the material: was all required material submitted and do data and documentation match?
- Technical control: formats, readability, malware/virus checks

After this initial completeness check and technical control, all files are saved on the archive server in their original versions and formats – this SIP is not altered and will be retained in its original form. Subsequently, the data are converted to or saved in the archival format and undergo further checks and – where necessary – corrections. These include

- control of plausibility;
- control of consistency;
- control of weightings;
- disclosure control, control of anonymity.

Any corrections carried out at this stage are documented in an SPSS syntax file which is archived in AIP along with the dataset. All significant corrections/changes of the data will be discussed with data depositors beforehand. The Data Archive's standards for data checking and correction procedures are documented in a wiki and outlined in Jensen (2012).

6.3.2.3 Data documentation and enhancement

All data received by the GESIS Data Archive is documented by the Archive staff by means of a study description drawing on a standard set of descriptive, administrative, and structural metadata (building on the documentation and material received from the data producer).

Descriptive metadata, published via the Archive's online Data Catalog (DBK) include

- **bibliographic information**, such as study number, study title, current version, date of collection, principal investigators, authoring institution, PID (DOI), topic classification;
- **information on study content**, such as abstract, topics, demographic information;
- **information on methodology**, such as geographic coverage, selection method, mode of data collection, data collector;
- **information on data and available documents**, such as information on number units and variables in the dataset, analysis system(s) used, access modalities;
- **information on errata and versions**, such as errata in current version, versions list;
- **further information**, such as comparable or related studies, related publications and study groups.

The metadata schema used by the Data Archive is compatible with DDI. The published metadata for a dataset is available for download in DDI2 and DDI3 format (see Zenk-Möltgen and Habbel 2012).

Further structural and administrative metadata is added for internal use. Among others, this provides relevant technical and provenance information. Detailed information on the metadata schema was published in Zenk-Möltgen and Habbel 2012 (in German).

Value-added documentation: While all archived data is processed and documented in the outlined form, individual datasets of particular importance and important study collections falling into the Data Archive's core areas of collection are processed (cumulated, harmonized, standardized), documented, and enhanced in much greater depth – not only on study level, but on the level of individual questions and variables – with additional tools (in particular, the Dataset Documentation Manager and the CodebookExplorer). Thus, the Data Archive offers specialized curation services in particular for the ALLBUS, the International Social Survey Programme (ISSP), Eurobarometer, European Value Study (EVS), Election Studies Germany, Comparative Study of Electoral Systems (CSES) among others. The aim of this process is to provide users with high-quality datasets, which can be explored, analyzed, and re-used in an effective and user-friendly manner.

Value-added documentation includes, for example, information on

- question and answer texts (partly multilingual),
- show cards,
- dataset coding,
- standard coding (missing values, etc.),
- derivation information for index/derived variables,
- variable use (e.g. weights),
- data appraisal information.

Further contextual information is, for example, provided with regard to topics, panels, scales, time and space integration.

6.3.3 Archival Storage and Preservation

6.3.3.1 Physical data preservation and storage

In addition to the backup procedures described in Security and Risk Management / Media Monitoring and Refreshing Strategy, the Archive has the following technological and organizational measures in place to assure that the data bitstream is securely stored and cannot be altered without authorization: Write access to the archive server is highly restricted and governed by a set of strict rules and regulations. Only two members of the Data Archive staff are authorized to add, delete or change files on the archive server. The transfer of files into the archive takes place by means of a special transfer folder in the network, from which data and documentation to be archived are picked up, checked once again for conformity with the Data Archive's preservation standards (file formats, naming conventions, etc.), and transferred onto the archive server by the two authorized staff members.

6.3.3.2 Preservation strategy

Supported by a constant monitoring of technology (storage technology and media, software and file formats) as well as a normalization of file formats during ingest, the Data Archive pursues a migration strategy to ensure long-term access to its holdings. Data and documentation are archived in well-defined, standardized file formats to ensure that efficient migration strategies can be developed when this becomes necessary. While the refreshing of storage media takes place continuously, format migrations are undertaken only if the readability / interpretability of archive holdings is endangered by technological obsolescence, if they cannot be processed and used anymore in a state-of-the-art manner, and/or if a format migration brings considerable advantages with regard to user-friendliness and the work of the archive.

During format migrations utmost care is taken not to alter the significant properties of the archival objects. The migration procedure is documented thoroughly in order not to compromise the archival objects' authenticity during the migration process. A large-scale migration project recently completed by the Archive concerned holdings originating from an IBM Mainframe used in the 1990s, which were migrated into a PC-readable format.

6.3.3.3 Version control/change procedures

The versioning of data is governed by a versioning guideline that is strictly adhered to in order to meet the requirements of DOI-assignment and the standards of trusted digital preservation. Each datasets receives a version number upon publication in the data catalog. The version number is a three-digit number (major. minor. revision). Examples for the rules according to which digits are changed are indicated in Table 4. The version number is included as a variable in the data set and added to file name in accordance with a set of naming conventions. The version history, indicating the major/minor changes made to the data, is documented in the metadata in DBK.

A distinction is made between Errata and version changes. Lists of Errata supplement an existing version to indicate known but so far not corrected errors in the data set. Corrections of Errata are then documented in the version history and result in the assignment of a new version number upon publication of the data set.

TABLE 4: EXAMPLES FOR VERSIONING RULES

Position 1: Major	**Position 2: Minor**	**Position 3: Revision**
Addition of one or several new sample(s) (usually countries) to an integrated or cumulated data set	Change of a variable, i.e. corrections or additions in the data set concerning labels, recodings, data formats, etc.	"Trivial" corrections (not relevant to the "meaning"/interpretation of the data as such), e.g. correction of obvious misspellings
Addition of one or several wave(s) to a cumulated data set		Simple revision of labels not relevant to the meaning/interpretation of the data
Addition (Deletion) of one or several variable(s) to (from) a data set		

As write access to the archive server is restricted, unauthorized change of archived data is not possible.

6.3.4 Dissemination

6.3.4.1 Visibility

The data made available by the GESIS Archive are distributed through various channels of GESIS's web offers as well as the CESSDA page. Possible entry points through the GESIS main web page and GESIS-related web pages are listed in Table 5 (see below for a more detailed description of the different catalogs and search systems).

TABLE 5: ACCESS POINTS FOR RESEARCH DATA

GESIS home page	Newly published data are announced on the index page through an RSS feed
GESIS "Services"	The "Services" section of the main homepage contains a simple search box for the Data Catalog DBK and directs users to the other available search interfaces for research data
Data Catalog (DBK)	Study descriptions for the complete archive holdings
ZACAT Online Study Catalog	Study and variable descriptions for added value studies
Codebook Explorer (CBE)	Standalone tool for offline usage of selected databases
Qbase	Search in full text of questionnaires and codebooks for selected holdings
Histat	Access-platform for time series data collected in economic and social historical research
sowiport	Social Sciences portal integrating information and metadata on scholarly literature, data, projects, etc. from domestic and international providers. The sowiport search engine also accesses the Data Catalog (DBK)
da|ra	The da|ra search engine covers metadata from the Data Catalog (DBK)
CESSDA web site	About 600 data sets archived by GESIS are searchable (along with holdings of the other CESSDA member archives) through the CESSDA catalog, which accesses GESIS's ZACAT
Datorium (from 2013)	Data sharing platform based on DSpace.

6.3.4.2 Availability and accessibility

As stated in GESIS's by-laws, among its primary tasks is "to create user-friendly and high-quality possibilities for access to all the information and data relevant to empirical social research […]"

(§2e; my translation). To facilitate access to the data, all data sets that can be shared without further consultation with the data depositor are available for download via the different search platforms. Users are required to register once and log in before downloading a data set. As the registration and download system are fully automated, the data are available free-of-charge immediately and at any time convenient to users.

In addition to using the download system for immediate download, for data in license categories B and C (see chapter 1.2), users can order data from the Archive's Data Service via a shopping cart system, by e-mail or telephone. They will receive this data (in customized form, if they wish) on a CD-ROM or DVD or via a secure download using Cryptshare. For this data service handling fees are charged (see http://www.gesis.org/en/services/data-analysis/data-archive-service/charges/).

6.3.4.3 Tools and interfaces

An overview of the main tools available to users to search, explore, and analyze studies (data sets and accompanying material) is given in Table 6 .

Table 6: Online and offline search tools

	1) **Data Catalog (DBK)**	2) **ZACAT Online Study Catalog**
Short description and scope	The DBK comprise the study descriptions for all archived studies and empirical primary data from survey research, historical social research	ZACAT (based on Nesstar WebView) allows users to search for, browse, analyze and download social science survey data for selected study collections. ZACAT includes documentation of full question and answer texts on variable level, partly in two or more languages. Available study collections: • International Social Survey Programme (ISSP) • Eurobarometer • European Values Study • ALLBUS (German General Social Survey) • Studies from Eastern Europe: Comparative Studies, Election Studies, Studies from individual countries • Election studies: German National Election studies 1949-2002, German Federal State Election studies • Childhood, adolescence and becoming an adult 1991-1997 • Politbarometer
Search/browse options	• Simple and advanced search • Boolean Operators • free text search in the following fields: Title, Study No., Date of Collection, Principal Investigator / Authoring Entity, Institution, Data Collector, Abstract, Geographic Coverage (ISO3166-Code), Geographic Coverage (ISO3166-Label), Geographic Coverage (free), Universe, Selection Method, Mode of Data Collection, Analysis System, Publications, Notes, Topic Classification, Availability Status Browsing of complete list of studies	“Searching and Browsing • Simple and Advanced search facilities • Ability to browse metadata, data and associated documentation • Ability to view metadata in a multiple languages Analytical Tools • Graphical representations of data in various forms […] • Cross-tabulations • Correlation and regression analyses • Ability to apply survey weights.”

		Quoted from http://www.nesstar.com/help/4.0/webview/getting-started/getting-to-know-nesstar-webview.html
URL/Access	http://www.gesis.org/en/services/research/data-catalogue/ Registration and log-in required for download.	http://www.gesis.org/en/services/research/zacat-online-study-catalogue/ Analysis, creating and downloading tables require registration and log-in.

	3) CodebookExplorer (standalone offline tool)	4) Qbase
Short description and scope	For selected surveys, the CodebookExplorer makes it possible to search for keywords in study or variable descriptions, compare question texts, carry out simple analyses (e.g. frequencies or crosstabs), and to display questionnaires.	Allows to search within codebooks and questionnaires for selected German and international studies. These include among others: • ALLBUS Codebooks and Questionnaires • Politbarometer Codebooks • Standard Eurobarometer – Codebooks (EN) and Bilingual Master questionnaires (FR/EN) • EVS European Values Study - Integrated Dataset - Master questionnaire 1981, 1990, 1999/2000, 2008 (EN) and Variable Report 1981, 1990, 1999/2000, 2008 (EN) • ISSP International Social Survey Programme - Codebooks and Master questionnaires (EN)
Search/browse options	"Depending on the database, different functions are available, like Study, Category, Trend or Scale View, each of them in an Explorer View. To every Explorer View there is a Description View, e. g. Study Descriptions or Category Descriptions (depending on the database). Databases can contain the respective information in original language or in English language. All fields are searchable. For every study in the ZA CodebookExplorer the questionnaire can be displayed as PDF File. In the Analysis View an overview of the results of the surveys can be produced with simple frequency tables or cross-tabulations, descriptives or comparative analyses. Graphs of them can be displayed and configured Through an assignment of variables to thematic categories, groups of variables are characterised as comparable, which facilitates to find variables according to thematic considerations. The Category View presents these	• Free text search in one or multiple study documents • Boolean Operators • Wildcards (left/right) • Tense conflation (left/right): • Further operators (see http://isysweb.gesis.org/help/command.html#CommandBasedOperatorRef)

	variables and makes it possible to edit, add, and delete available entries." Quoted from https://info1.gesis.org/CEI2/help/index.htm	
URL/Access	http://www.gesis.org/unser-angebot/recherchieren/codebookexplorer/ Order of CD-ROM with the respective databases required.	http://www.gesis.org/en/services/research/german-question-text/, http://www.gesis.org/en/services/research/english-question-text/

5) Histat

Short description and scope

Database providing access to metadata and data of about 260,000 time series data from more than 360 studies in Social Sciences and economic historical research.

Search/browse options

- Study metadata can be browsed according to general subject category, time period, and author’s name
- Free text search in the table titles and variable names of the data as well as in the data tables’ source section
- Free text search in the study descriptions
- Results can be limited to a certain time period or subject category

See Franzmann 2010 for more detailed information.

URL/Access

http://www.gesis.org/histat/en/index

Registration and login required for download.

6.3.4.4 Monitoring, review and feedback

GESIS monitors usage of its web offers by means of web and download statistics. This includes statistics documenting the use of the Data Archives' services available to registered users (download of datasets in particular). In addition, the Data Archive generates statistics on data provision and consulting activities using other communication channels (e.g. e-mail, phone).

To ensure that the services offered meet the needs of its designated communities, GESIS and its departments frequently carry out evaluations and user surveys as part of the quality management (this includes, among others, a large-scale portfolio analysis of all GESIS products and services). In order to systematically include important stakeholders in the strategic development of GESIS and its services, a scientific and a user advisory board were created.

In addition to the internal monitoring and evaluation activities, GESIS and its departments undergo frequent external audit and evaluation by the Leibniz Association.

Sources

Büttner, Stephan, Hans-Christoph Hobohm, and Lars Müller. 2011. Handbuch Forschungsdatenmanagement. Ed. Stephan Büttner, Hans-Christoph Hobohm, and Lars Müller. Bad Honnef: Bock + Herchen.

CCSDS. 2012. "Reference Model for an Open Archival Information System (OAIS). Recommended Practice." http://public.ccsds.org/publications/archive/650x0m2.pdf.

CCSDS. 2004. Producer-Archive Interface Methodology Abstract Standard. http://public.ccsds.org/publications/archive/651x0m1.pdf.

Centre for Intellectual Property Law (CIER). 2011. The Legal Status of Research Data in the Knowledge Exchange Partner Countries. http://www.knowledge-exchange.info/default.aspx?id=461.

Franzmann, Gabriele. 2010. HISTAT (Historische Statistik). Recherche- Und Downloadsystem Für Studien Mit Zeitreihen, Zur Historischen Demografie, Zur Empirischen Sozial- Und Wirtschaftsgeschichte Sowie Zur Historischen Statistik Deutschlands. Eine Kurze Beschreibung. http://www.gesis.org/fileadmin/upload/dienstleistung/daten/hist_sozialforschung/dokumente/HISTAT_Beschreibung.pdf.

Hausstein, Brigitte, Wolfgang Zenk-Möltgen, Anja Wilde, and Natalija Schleinstein. 2011. "Da|ra Metadatenschema. Version 1.0." doi:10.4232/10.mdsdoc.1.0.

Jensen, Uwe. 2012. "Leitlinien Zum Management Von Forschungsdaten. Sozialwissenschaftliche Umfragedaten." http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2012/TechnicalReport_2012-07.pdf.

Luijkx, Ruud, Evelyn Brislinger, and Wolfgang Zenk-Möltgen. 2003. "European Values Study 1999/2000 – A Third Wave: Data, Documentation and Database on CD-ROM." ZA-Information (52): 171–182. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/za_information/ZA-Info-52.pdf.

Neuroth, Heike, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump, and Jens Ludwig. 2012. Langzeitarchivierung Von Forschungsdaten. Eine Bestandsaufnahme. Ed. Heike Neuroth, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump, and Jens Ludwig. Boizenburg: Verlag Werner Hülsbusch. http://nestor.sub.uni-goettingen.de/bestandsaufnahme/nestor_lza_forschungsdaten_bestandsaufnahme.pdf.

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

Wira-Alam, Andias, Dimitar Dimitrov, and Wolfgang Zenk-Möltgen. 2012. "Extending Basic Dublin Core Elements for an Open Research Data Archive." In International Conference on Dublin Core and Metadata Applications. http://dcevents.dublincore.org/index.php/IntConf/dc-2012/paper/viewPaper/102.

Zenk-Möltgen, Wolfgang, and Norma Habbel. 2012. "Der GESIS Datenbestandskatalog und Sein Metadatenschema. Version 1.8."

7 Norwegian Social Science Data Services[47]

7.1 Organizational Framework

7.1.1 Purpose and Requirement

7.1.1.1 Scope and objectives

NSD was established in 1971. The organisation was institutionally affiliated to the Research Council of Norway. On 1 January 2003, NSD became a limited liability company owned by the Norwegian Ministry of Education and Research.

NSD's main objective is to improve opportunities and working conditions for empirical research. This is achieved by collecting, processing, facilitating, archiving, maintaining and disseminating data to research communities.

The majority of NSD's users are attached to the universities, university colleges and research institutions, in social science disciplines. Researchers and students from other institutions and disciplines, particularly from health enterprises and hospitals, are also using NSD's services more and more.

NSD is a permanent national research infrastructure organised as a limited liability company fully owned by the Norwegian Ministry of Education and Research. The funding structure is based on major allocations from Research Council of Norway, the ministries, the research and education sector and the EU. Only a minor portion of NSD's income comes from trades (outside the sectors listed above). NSD are not charging academic users.

Ever since NSD was established, the Research Council of Norway has, through its basic allocations of funding and project allocations, played a key role with respect to both the scope and quality of the services that NSD offers. The Research Council is the largest contributor of funds to NSD and its investments have contributed to the establishment of a number of joint resources, which have improved opportunities and working conditions for empirical research.

NSD has a collaboration with Statistics Norway (SN) that is unique in the international context. SN use NSD for managing data. This has resulted in the establishment of procedures for safeguarding the protection of personal data while allowing for relatively extensive use of individual data for research purposes.

NSD is the Data Protection Official for Research and has signed agreements with all Norwegian universities, specialised university colleges and state university colleges and a number of health enterprises and research institutes. This means that the Norwegian Data Inspectorate has

[47] This segment was provided by Dag Kiberg, NSD.

delegated responsibility to NSD in relation to the Personal Data Act and the Personal Health Data Filing System Act.

Major services offered by NSD:

- Data for research, teaching and student dissertations
- Guidance and assistance in connection with data collection
- Research design and data analysis
- Guidance on legal and ethical rules relating to data collection
- Use and filing of personal data
- Fling of data from research projects
- Training
- Delivery of Nesstar software
- Documentation of and information about research

7.1.1.2 Collection policy

In collaboration with the Research Council of Norway, NSD collects and archives data from research projects that are subject to legal deposit as a result of the Council's- allocation terms and conditions.- NSD has established systems and procedures for compiling and making data available to ensure that the archiving duty is met and that data are deposited in line with contractual terms and conditions. The work is a continuation of the responsibility that NSD has had for almost 40 years to archive data from projects that receive funding from the Research Council of Norway.

NSD also collects data from other projects conducted by researchers and students at universities, specialised and state university colleges, institutes and other research institutions. The data that can be archived comprises anonymous machine-readable quantitative data and machine-readable data with personal identification. NSD's work on archiving research data is in line with the major investment in research infrastructure and better utilisation of research resources that are seen in Europe and worldwide.

7.1.1.3 Criteria for evaluating data

NSD's criteria for evaluating data follow 4 dimensions;

a) Is the data set relevant to NSD's mission, remit and priorities?

b) Does the data set meet~~s~~ applicable regulations and ethical codes, i.e. regarding personal data? The data are checked for anonymity based on 4 categories:

- Direct person-identifiable data file which contains information that directly identify individuals.
- Pseudonymous data and unique personal encrypted data where identification is modified according to a specified key.
- Indirect identifiable data without directly identifying information, but where one can identify individuals using the combination of background variables.
- Completely anonymous data where it is not possible to identify individuals.

c) What is the technical status? The files are checked regarding readability, file format , data format, and if the files contain viruses etc.

d) What is the metadata status? Data are checked for documentation on value -level, survey questions and, if applicable, references to publications based on the data. The metadata shall support sharing, access and re-use of the data.

7.1.2 Legal and Regulatory Framework

As a limited liability company, and thus a legal entity, NSD is itself responsible to comply with all actual laws, regulations, agreements, etc. With respect to compliance with national laws under which NSD operates, in Norway there are several regulations related to the privacy and protection of research participants. Of particular note is the Personal Data Act and Health Register Act.

NSD has a "Data Protection Official for Research"[48]. The Data Protection Official scheme implies that the requirement for obtaining licenses from the Data Inspectorate for a greater part of processing of individual data at NSD are replaced by a notification requirement where the Data Protection Official for Research is the last instance for reviewing applications for licenses. This means that the Data Inspectorate basically has delegated part of its responsibility to NSD itself in relation to the Personal Data Act[49].

The Research Council of Norway and NSD have an agreement that makes NSD responsible for the archiving of data from projects financed by Research Council of Norway. The filing requirement is incorporated into the Research Council of Norway's funding contract terms. The depositors thus have a contract with the Research Council of Norway and not with NSD. -For

[48] http://www.nsd.uib.no/personvern/en/index.html

[49] NSD is also Data Protection Official for Research for most of the research sector in Norway.

depositors from projects not funded by Research Council of Norway, NSD and the depositor sign a special agreement.

7.1.3 Funding and Resource Planning

Based on its basic allocations of funding and project allocations, Research Council of Norway has played a key role for decades with respect to both the scope and quality of the services that NSD offers to its users. Their investments in NSD have contributed to the establishment of a number of joint resources, which have improved opportunities and working conditions for the research communities. Further main allocations come from the ministries, the universities and university colleges sector and the EU.

7.1.4 Long Term Preservation Policy

In order to effectively prevent loss of deposited data, NSD has implemented a set of strategies involving a focus on hardware, software and human knowledge. The goal is to preserve data and be able to make them available in perpetuity. Procedures are established regarding challenges related to hardware, software, and human knowledge.

As an permanent national institution based on public funding and governmental ownership, NSD is able to schedule its operations on long time terms and to accumulate necessary human knowledge.

NSD applies processes and procedures for guaranteeing the storage of data. These include a backup strategy and plan that reflects requirements for availability and integrity of the ICT services and information assets. Quality control and quality assurance procedures are used especially in the pre-ingest, ingest, archival storage and access functions.

NSD's long time preservation includes a Nesstar[50] solution combined with the use of portable media stored in a safe place outside the NSD's premises. As Nesstar is developed by NSD's development team NSD has full control on future compatibility and software obsolescence.

NSD's routines for long-term preservation are included in internal data curation documents which are in line with the overall ICT strategy of NSD. The objectives are to secure confidentiality, integrity and access/availability.

7.1.5 Access Policy

NSD aims to provide access to data and other resources in as open a manner as possible without compromising the privacy. A large part of NSDs data sets can be accessed through NSD's Nesstar solutions. In addition to direct on-line access, data can be downloaded and/or exported to preferred format. For some of NSDs data sets only metadata can be accessed and explored. Access to the data are then given upon application.

[50] http://www.nesstar.com/

7.2 Technological Environment

7.2.1 IT Architecture

Data are stored and processed on NSD’s servers, data are included in NSD’s backup routines. Backups, both incremental and full, are stored and secured outside NSD’s premises. All system on NSD are protected against disasters etc.

Nesstar is the preferred and major system. It allows the user to have remote access to data and/or metadata. When data are published along with metadata (some types of data are publishes with metadata only) the users are allowed to download data as well.

NSD are now, in a joint project with Statistics Norway, developing a system for remote *processing*. The primary goal for this project is to provide access register data. However, the system could, when completed, also cover survey data.

7.2.2 Standards and Formats

NSD's policy is that all dataset shall be documented in Nesstar. Nesstar is supporting most/all known data formats used for quantitative data set. The fact that NSD is the developer of the software also secure future compatibility. The aim is to make all documentation in a consistent format based on approved templates. NSD also uses Nesstar as a publishing channel for all data sets. NSD accepts all major file formats and sets no restrictions in this regard. Data are however stored in the original format *and* in a format proper for archiving. At NSD this is usually Nesstar format. The objective is to make all information on all levels in a uniform manner with pre-defined templates.

7.2.3 Security and Risk Management

NSD receives, process, store and distribute large amounts of data to and from the users and partners. NSD values thus lies in the amount of information that the institution administers. The main focus is on the protection and maintenance of the necessary access to information assets. NSD is working continuously to have an IT management that support the NSD's goals, values and main purposes.

7.3 Data Curation

7.3.1 Pre-Ingest Function

7.3.1.1 Information and guidance given to data producer

NSD provides information and guidance online about the package of information that should be deposited to facilitate data assessment and reuse. NSD also works with researchers before data deposit to ensure that the research data are fully documented and optimally usable. With respect to ethical and legal norms, NSD obtains information from data producers about their compliance with these obligations. While assessments of data quality are ultimately the task of the secondary analyst, NSD attempts to gather enough information about any given study and its investigators to make quality assessments possible.

NSD does not publish any list of preferred formats. The research communities use in almost all cases standard formats, and NSD has the knowledge and ability to convert any known format to our standard archival format (Nesstar).

7.3.1.2 Ingest Function

NSD has developed and implemented an internal administration system for all quantitative datasets deposited at NSD. The purpose of the system is to compile all relevant information about the data sets, including documentation of versions and processing. In order to be able to recreate data sets, all versions as well as the processes leading up to the versions are stored. NSD documents received data by using the complete metadata authoring tool Nesstar Publisher, which consists of data and metadata conversion and editing tools enabling the user to prepare data and metadata for publication to a Nesstar Server. The Publisher is DDI compliant.

NSD performs disclosure control on all incoming data. The data sets are categorised according to four classifications:

1. Direct identifiable personal data – split of information;
2. Pseudonym direct identifiable personal data – the key is stored separate from the data;
3. Indirect identifiable personal data;
4. Anonymous data.

In general, the descriptive and structural metadata deposited by data producers is sufficient. NSD uses a simple deposit form for describing the data collections. In addition the researcher submit questionnaires, relevant publications etc. In cases of missing information, NSD will contact the data producer for supplements.

7.3.1.3 Information and documentation from data producer

NSD requests the data producers to submit:

- Information necessary for measuring scientific content and value
    - Enough information for other to evaluate the scientific value of the data
    - Enough information to know that the data are compiled and can be used by scientists
- Metadata

        - *Descriptive metadata*; i.e. when and who has compiled the data, what are the data about, etc
        - *Structural metadata*; code book etc. that are necessary to process the data
        - *Administrative metadata*: information necessary to access the data

7.3.1.4 Quality assurance and data checking

Received data, metadata and other relevant material are checked:

    - Technical check
        - Viruses and other harmful software
        - Accessibility
        - Format
        - Readability
    - Check for personal identifications
        - Directly personally identifiable data (direct); in case the file is split
        - Pseudonymous data and unique personal encrypted data
        - De-identified data (indirectly identifiable)
        - Anonymous data

7.3.1.5 Data documentation and enhancement

All received data are documented in Nesstar in order to maintain a unified documentation standard with pre-defined templates based on DDI.

- General
    a. Abstract
    b. Full Title
    c. Identification Number
    d. Authoring Entity
- Data production

a. Producer

b. Date of Production

c. Place of Production

d. Funding Agency/Sponsor

e. Data Distributor

f. Contact Person

- Depositing

  a. Depositor

  b. Series Name

  c. Series Information

  d. Version

  e. Notes

  f. Bibliographic Citation

- Content

  a. List of Keywords

  b. Topic Classification

- Period and geographical coverage

  a. Time Period Covered

  b. Date of Collection

  c. Country

  d. Geographic Coverage

- About the dataset

  a. Unit of Analysis

  b. Universe (Population)

    c. Kind of Data
    d. Time Method
    e. Data Collector
    f. Sampling Procedure
    g. Mode of Data Collection
    h. Type of Research Instrument
    i. Response Rate

- Accessibility
    a. Data Set Availability
    b. Location
    c. Availability Status
    d. Extent of Collection
    e. Completeness of Study Stored
    f. Number of Files
    g. Restrictions
    h. Citation Requirement
    i. Deposit Requirement
    j. Conditions
    k. Disclaimer

In addition to published metadata, also structural and administrative metadata is added for internal use.

7.3.2 Archival Storage and Preservation

7.3.2.1 Physical data preservation and storage

Original data and metadata are stored off-line as they are received. New versions and editions are stored at dedicated server. All data and metadata are stored at a Nesstar server. Non-anonymous data are secured and stored at portable media outside NSD.

Data stored in-house at NSD are subject to daily incremental backup and weekly full backup. Backup are secured and stored outside NSD. Portable storage media are checked and migrated on regularly basis.

7.3.2.2 Preservation strategy

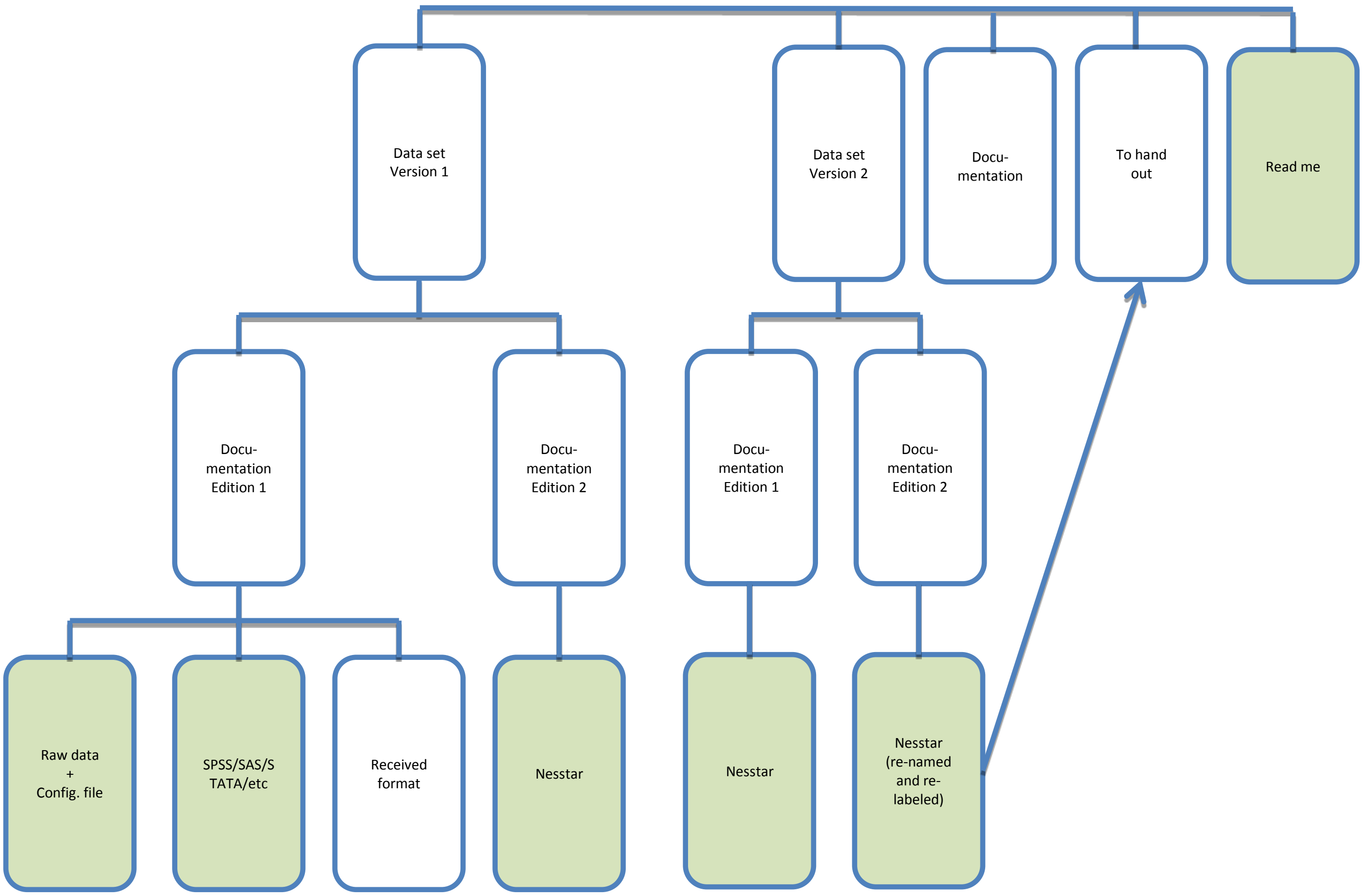
7.3.2.3 Version control/change procedures

NSD distinguishes between different versions of a dataset. Different versions occurs when changes in the data matrix are made. That is either differences in the number of units, number of variables or a variable value category.

NSD also distinguishes between different editions. What distinguishes one edition from another is change in metadata, and a variation of a data set can have multiple editions. Primary data as delivered from the researcher are always the first version of the data set (variation 1), and accompanying documentation is the first edition (edition 1). If the file is documented in Nesstar it becomes the next edition (edition 2). When a data set is ready for dissimination, this marked "to hand out." This will always be the latest version / edition.

All changes made are documented, and all versions and editions are stored together with the documentation. This is done in order to have the possibility to roll back to previous versions/editions if necessary.

The diagram on the next page shows a typical storage structure for a two-version.



7.3.3 Dissemination

Data archived at NSD are as a rule available for reuse for researchers and students from academically institutions, and NSD's access policy is closely related to the work of archiving data, provided in all the main principles of the OECD guidelines[51] on access to publicly funded data. These can briefly be summarised to that publicly funded data is a public good, that within the current legislation, should be open and freely available for research. NSD emphasized that storage is done in a flexible manner to ensure accessibility regardless of changes in technology or staff.

NSD aims to make information about data producers, data documentation, and~~,~~ as far as possible, data as visible and available as possible. Any lack of such visibility contributes largely to poor or lack of data availability. NSD is therefore continually developing systems and procedures that contribute to increased visibility. In 2010 NSD developed a new bi-lingual web-service that makes it possible to search more than thousand Norwegian surveys. This makes-NSD's collection of national surveys one of the largest in Europe.

[51] http://www.oecd.org/science/scienceandtechnologypolicy/38500813.pdf

In case NSD gives online access to metadata only, the user will have access to data upon application (in some cases after consultations with data owner). When access is granted, the researcher or student will be given a temporarily access - usually for two years. Before data are delivered the receiver has to sign a user agreement. If the data will be used by a student the supervisor has to sign a supervisor agreement. According to the user agreement the user pledge secrecy, cannot pass onto third parties and are obliged to give credit to the owner of the data, and to NSD.

Another example is NSD's development (on assignment for the Research Council of Norway) of a portal for educational research. The portal provides an overview of projects, data and research communities linked to educational research. The major objective of the portal is to ensure that the available data is used more, and that it is used repeatedly to a greater extent.

7.3.3.1 Visibility

NSD major system for online visibility is Nesstar. All data generated from research projects are published in the Nesstar based system "*Norske spørreundersøkelser*" (Norwegian surveys). The system allows four options for a data set's visibility:

1. metadata and data;
2. metadata and variable frequencies;
3. metadata;
4. study description.

NSD promotes data sets and data collections in different ways. This can either be to inform on one particular study or program like the NSD special publication of different data series and cumulative files. In addition, develops and publishes NSD specific web pages for data within a certain category, eg. within specific fields of research. The intention is to make it easier for users to be aware of and to find data in this field.

7.3.3.2 Availability and accessibility

Regarding availability and accessibility NSD categorize data into three main categories:

1. Open online access to data and metadata
2. Open online access to frequencies and metadata – access to data upon application
3. Open online access to metadata - access to data upon application

Category 1 applies mainly to data owned by NSD or by agreement with data owner

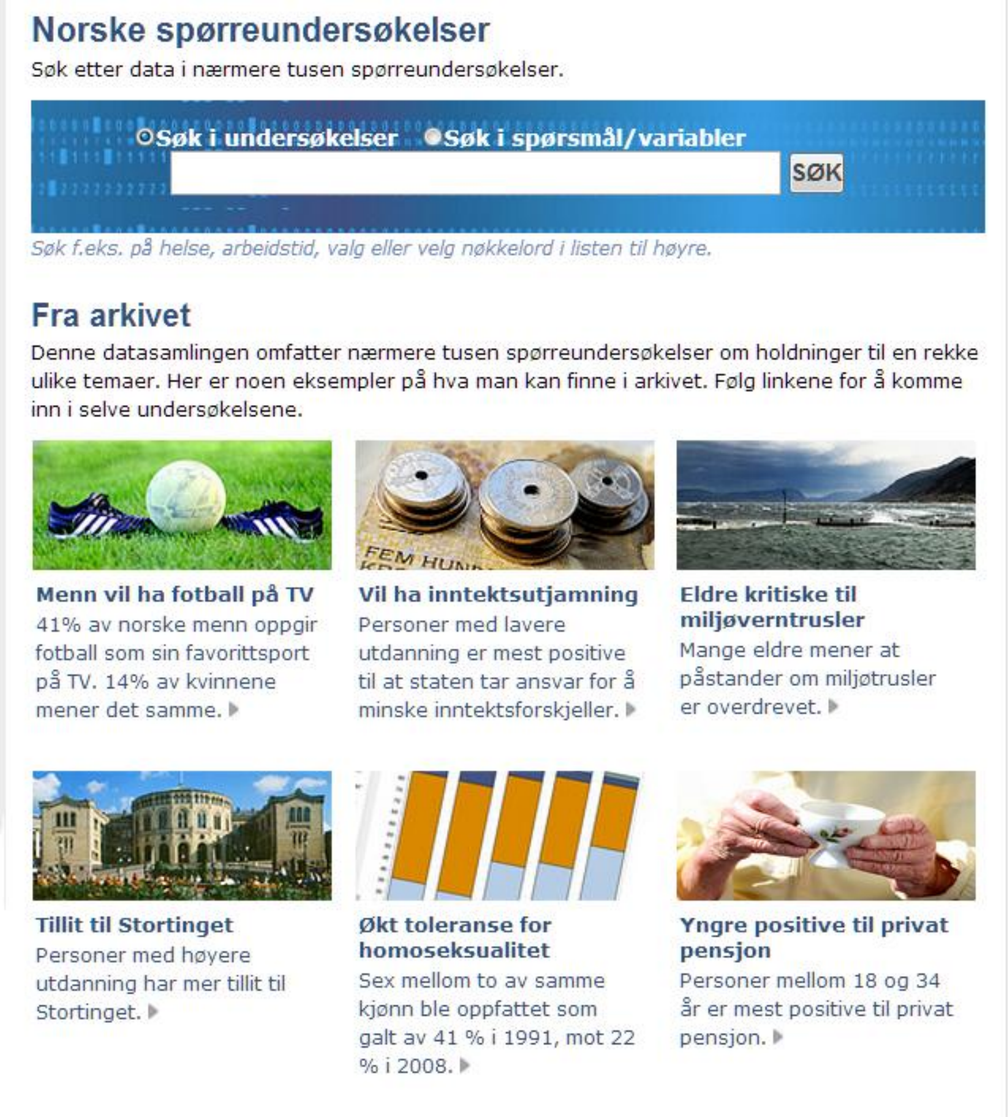
Category 2 applies to data from research projects

Category 3 applies to qualitative data

7.3.3.3 Tools and interfaces

All archived data at NSD are in one form or another published Nesstar via a web interface. NSD has also developed interfaces on top of Nesstar. For search among archived data NSD has made an online system (see illustration below) for searching among more than one thousand surveys.

Nesstar that is used in the most of online access and browse services at NSD's web is a software system for data publishing and online analysis. The software consists of tools which enables data providers to disseminate their data on the Web. Nesstar handles survey data and multidimensional tables as well as text resources. Users can search, browse and analyse the data online.



Norske spørreundersøkelser

Søk etter data i nærmere tusen spørreundersøkelser.

Søk f.eks. på helse, arbeidstid, valg eller velg nøkkelord i listen til høyre.

Fra arkivet

Denne datasamlingen omfatter nærmere tusen spørreundersøkelser om holdninger til en rekke ulike temaer. Her er noen eksempler på hva man kan finne i arkivet. Følg linkene for å komme inn i selve undersøkelsene.

Menn vil ha fotball på TV
41% av norske menn oppgir fotball som sin favorittsport på TV. 14% av kvinnene mener det samme. ▶

Vil ha inntektsutjamning
Personer med lavere utdanning er mest positive til at staten tar ansvar for å minske inntektsforskjeller. ▶

Eldre kritiske til miljøverntrusler
Mange eldre mener at påstander om miljøtrusler er overdrevet. ▶

Tillit til Stortinget
Personer med høyere utdanning har mer tillit til Stortinget. ▶

Økt toleranse for homoseksualitet
Sex mellom to av samme kjønn ble oppfattet som galt av 41 % i 1991, mot 22 % i 2008. ▶

Yngre positive til privat pensjon
Personer mellom 18 og 34 år er mest positive til privat pensjon. ▶

It is now possible to build custom applications to harness or to do tailor made operations on metadata stored in a Nesstar server. The Nesstar team has built a public API for this purpose.

7.3.3.4 Monitoring, review and feedback

NSD has no systematic plan for feedback to its archiving service. The archivists are, however, in their daily work regularly in close contact with many of the users and receive important signals regarding the service. NSD keeps track of the use either by automatic counting, recording of users' e-mails, or orders that come by mail, NSD’s online web-based order form, or e-mail. Users are required to submit publications that are written on the basis of data from NSD.

8 DANS[52]

8.1 Organizational Framework

8.1.1 Purpose and Requirements

8.1.1.1 Scope and objectives

DANS promotes sustained access[53] to digital research data. For this purpose, DANS encourages researchers to archive and reuse data in a sustained manner, e.g. through the online archiving system EASY[54]. DANS also provides access, via NARCIS[55], to thousands of scientific datasets, e-publications and other research information in the Netherlands. In addition, the institute provides training and advice, and performs research into sustained access to digital information.

DANS provides services for data access and preservation for all sciences. EASY presently contains large collections for the disciplines of Humanities (history and archaeology), Social and behavioural sciences, Geospatial sciences. The DANS staff includes data managers with expertise in these disciplines, but DANS is not exclusive to these fields of research. All disciplines are welcome to deposit their data at DANS. The DANS strategic plan for the period of 2011-2015[56] includes a further development to a discipline-independent data-organisation; a data archive which encompasses all fields of science where there is a demand for long-term data preservation and access.

DANS takes part in numerous projects[57] with the aim of promoting the research data infrastructure in the Netherlands and Europe. Sometimes as project leader, on other occasions as a consortium member, work package leader, main contractor, sub-contractor or supplier of a specific service[58].

8.1.1.2 Collection policy

DANS applies the concept of *self-archiving*: data producers deposit their research data with DANS themselves using the deposit service in the EASY archive. DANS archivists can provide assistance during the depositing procedure and will process all submitted datasets.

Certain data projects have an obligation to deposit data due to the requirements from their subsidiary. For example, research funded by the Netherlands Organisation for Scientific Research (NWO). Dutch archaeology adheres to a national regulation (the Kwaliteitsnorm voor

[52] This segment was provided by Mike Priddy, DANS.
[53] long-term preservation, archiving, accessibility and availability
[54] https://easy.dans.knaw.nl
[55] http://www.narcis.nl
[56] http://www.dans.knaw.nl/en/content/about-dans/more-information/strategy
[57] http://www.dans.knaw.nl/en/content/projects
[58] http://www.dans.knaw.nl/en/content/services

de Nederlandse Archeologie, KNA quality norm[59]), which rules that all digital documentation from archaeological research projects needs to be deposited for long-term preservation.

Apart from the obligatory deposits, all scientists are welcome to deposit their research data at DANS. A Data Management Plan for scientific research is available on the DANS website[60] containing instructions and addressing questions important for researchers in the early stages of a data collection project. Additionally, a Data Management Plan Checklist is provided within the document.

Researchers are instructed to prepare their data collection with:

- general information
- an overview of previously collected data
- a choice of software and hardware to be used
- a determination of intellectual property and legal requirements
- user information
- interoperability requirements

When the data collection project is implemented, further arrangements are made regarding the data management needs and tasks and the choice of metadata to be used.

The specifics of the data archiving can then be decided on. The DANS EASY Electronic Archiving System meets with the guidelines of the internationally established Data Seal of Approval (see *1.4*) and DANS recommends the use of EASY for archiving datasets.

8.1.1.3 Criteria for evaluating data

A checklist for storing and selecting research data[61] is provided on the DANS website. It may be obligatory to store research data for reuse or verification. Alternatively, a researcher may wish to deposit their data because of the data's importance, uniqueness or reusability.

When datasets are submitted to DANS, an archivist will check the dataset for completeness and understandability. The main goal of archiving datasets is to make future research possible. Anyone who is not familiar with the data needs to be able to understand it. If files in the dataset, or codes or variables within the file, cannot be understood, the depositor is requested to provide additional metadata.

Researchers are recommended to deposit their files in either preferred or accepted formats (see *2.2*). The archivist may contact the depositor of files that are not in a preferred or accepted

[59] http://www.erfgoedinspectie.nl/archeologie/wet-en-regelgeving/kna

[60] http://www.dans.knaw.nl/sites/default/files/file/Datamanagementplan%20UK%281%29.pdf

[61] http://www.dans.knaw.nl/sites/default/files/file/archief/Factsheet_Checklist_storing_and_selecting_DEF.pdf

format. If it is not possible to deliver the files in a preferred format, other formats can still be chosen for publication, but DANS will not be able to guarantee their long-term preservation.

8.1.2 Legal and Regulatory Framework

Both to depositing and using data, agreements apply which are based on Dutch and European legislation and the 'Code of conduct for use of personal data in scientific research' of the Dutch Association of Universities (VSnU[62]). This code of conduct is an elaboration of the Dutch Data Protection Act. The agreements are incorporated in the DANS Licence Agreement[63] and the DANS General Conditions of Use[64].

Before being able to submit a dataset to DANS, the dataset depositor is required to acknowledge acceptance and understanding of the Licence Agreement. A licence agreement PDF is generated for the specific deposit, including the depositor account details and the title of the dataset. This PDF is made available to the depositor on the final screen of the deposit procedure ('Overview and submitting'). Two inclusions of the agreement are emphasized:

- DANS is granted a non-exclusive licence to store and make available to third parties the digital dataset, in according with the access conditions indicated by the depositor.
- The depositor declares to be the holder of rights to the dataset and/or entitled to act in the present matter with the permission of other parties that hold rights.

The license is non-exclusive: the owner of the data is at liberty to deposit and/or make available the data in other places as well. Copyright is not waived when data are deposited. The depositor, or stated copyright holder, retains the copyright.

The rules for data use are laid down in the DANS General Conditions of Use, in conjunction with the licence terms. Every time a (set of) file(s) is downloaded from EASY, the downloaded file(s) are compressed into a ZIP-file including the DANS General Conditions of Use in PDF.

The General Conditions of Use concerns the following elements in particular:

- Personal Use: Users are only allowed to download the dataset for personal use. Copying (parts of) the dataset for other purposes is not allowed. Users are not allowed to make any commercial use of the dataset. Copyrights of third parties must always be respected.
- Bibliographic reference of the dataset: When publications make use of datasets, which originate from EASY, this must be indicated by a bibliographic reference.
- Publications based on the dataset: DANS requests every user to inform DANS of bibliographic information on all publications for which datasets from EASY were used.

[62] http://www.vsnu.nl/code-pers-gegevens.html (in Dutch)
[63] http://www.dans.knaw.nl/sites/default/files/file/archief/Licence_agreement_DANS_UK.doc
[64] http://www.dans.knaw.nl/sites/default/files/file/archief/DANS_General_Conditions.pdf

- Personal data: Special restrictions apply to datasets with personal data according to the Dutch Data Protection Act (WBP). Therefore, these data cannot be consulted via EASY in practice. Under strict conditions, they can be made available for scientific research. In such cases, the user is obliged to guarantee confidentiality of the data.

DANS processes privacy-sensitive and personal data according to DANS' privacy regulations[65]. Notice has been given of the DANS Privacy Regulations to the Personal Data Officer of the Royal Netherlands Academy of Sciences (KNAW) on April 22th 2009.

Portrait right, which concerns the identifiable representation of persons, applies to visual data. In addition, visual and audio recordings may contain privacy-sensitive information. Files that contain this type of data may only be made available with the explicit permission of the person(s) concerned. If there is no permission, data can only be used if they do not invade privacy. In practice, this means that DANS completely anonymises the data, for instance replacing names and other identifying data in audio files by 'beeps'.

Interview data can only be disclosed if there is a statement including the explicit permission of the interviewee.

Downloading of the information in NARCIS, or copying it in any other fashion, is permitted. One exception applies to this regulation. Re-use of information in the sections "Persons" and "Organizations" is restricted. Due to contractual and legal reasons information on persons and organizations presented in these sections may not be crawled, or copied in any other fashion. Re-use of information in the sections "Persons" and "Organizations" is allowed solely and in very small measure for personal study or use.

Harvesting the information in NARCIS by OAI-PMH protocol in order to collect and index metadata is permitted only after registration. It is required to contact the functional manager of NARCIS to discuss the use of information in the sections "Persons" or "Organizations". Registering can be done by submitting a completed online form to DANS.

8.1.3 Funding and Resource Planning

Two-thirds of all revenues are structural grants by the Royal Netherlands Academy of Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO). Financial plans by KNAW/NWO are guaranteed for every 4 years. DANS has to deliver annual reports and every 5 years a roadmap about the future strategy.

One-third of all revenues are derived from third-party funding for specific projects.

[65] http://www.dans.knaw.nl/en/content/dans-privacy-regulations

A costing model is being developed to charge depositors/funding bodies for large deposits of data sets.

8.1.4 Long-Term Preservation Policy

The license agreement between the data depositor and DANS states that the data archive:

- shall ensure that the deposited dataset is archived in a sustainable manner and remains legible and accessible
- shall preserve the dataset unchanged in its original software format, taking account of current technology and the costs of implementation;
- has the right to modify the format and/or functionality of the dataset if this is necessary in order to facilitate the digital sustainability, distribution or re-use of the dataset

To ensure that archived data can still be found, accessed and used in the future, DANS developed the Data Seal of Approval[66] (DSA). This data seal can be requested and granted to data repositories that meet a number of clear criteria in the field of quality, preservation and accessibility of data.

8.1.5 Access Policy

In line with the policy of the Dutch Universities, NWO and the KNAW, DANS promotes the use of Open Access but understands the occasional need for access restrictions: “open if possible, protected if necessary”.

All access agreements are as much as possible written in the spirit of the principles of the Open Access movement[67] as established in recent years. The most explicit government policy statement on research data from public funding can be found in the internationally agreed upon OECD Principles and Guidelines for Access to Research data from Public Funding, OECD (Paris, 2007)[68].

DANS also adheres as much as possible to the Creative Commons[69] licences. However, these licences are aimed at publications and not yet applicable to data files.

Data producers decide on the access level of their own datasets. A general access level is selected during the deposit procedure, but it is possible for the data producer to specify the access level per individual data file. See *3.4.2* for a detailed overview of the accessibility options.

The access level is set for the data files, not the metadata. EASY metadata (Dublin Core and file details) can always be viewed by anyone, even without needing to log into EASY. The metadata is checked by an archivist for privacy-sensitive information and made anonymous if necessary.

[66] http://datasealofapproval.org
[67] http://www.openaccess.nl/
[68] http://www.oecd.org/dataoecd/9/61/38500813.pdf
[69] http://creativecommons.nl/

It is required to register as an EASY-user before data can be downloaded. User actions will be registered in an Activity log with each dataset. All users who are logged in can view the Activity log. Users can choose to have their activity log registrations displayed with their username or as ‘anonymous’, the full information remains available to the DANS archivists.

The DANS General Conditions of Use adhere to Dutch and European regulations and conducts for the use of scientific data. The ‘Code of conduct for use of personal data in scientific research’ of the Dutch Association of Universities (VSnU) applies to the use of datasets with personal details.

8.2 Technological Environment

8.2.1 IT Architecture

The online archiving system EASY (Electronic Archiving SYstem) is used as a content management system for data users and data depositors.

EASY provides users with direct access to the data contained in the collections of the DANS data archive. Depositors have the choice of having their datasets published in Open Access, restricted to specific groups, or restricted to individual permission requests. If a user meets the download access conditions, data and documentation files can be downloaded free of charge.

EASY is developed by DANS and built on Fedora Commons repository software. Users can make use of faceted browsing and advanced search options to find datasets on their metadata given by the depositor in (Qualified) Dublin Core. All metadata can be viewed without needing to log into the system. A log-in is necessary for all downloads, even when the data is published in Open Access. Anonymous access to files is technically possible using Fedora digital objects, but the legal agreements in place require the need for user authentication and activity registration. Anonymous access can be arranged for, as long as the use of data files from EASY can still be measured. DANS is working towards enabling a selectable Anonymous Access rights condition for data depositors.

EASY-registration is free and accessible for everyone.

EASY contains a user interface facility for easy manual deposits of datasets. Depositors are instructed to prepare their datasets, then describe their datasets within (Qualified) Dublin Core metadata fields (see *3.2.1*). The files belonging to the dataset can be uploaded during the deposit procedure. It is possible to compress many files into a ZIP-file and upload this single ZIP instead of the separate files; EASY will extract the ZIP upon uploading. Alternatively, data can be agreed to be sent to DANS outside of EASY, for example on a CD or DVD. A data manager will add the files to the dataset by FTP during dataset processing.

After the depositor finishes describing the dataset and accepts the non-exclusive DANS Licence Agreement, the dataset can be submitted to DANS. A DANS datamanager will then further process and publish the dataset.

Upon submitting a dataset, a unique Persistent Identifier is generated and assigned to the dataset. The depositor will receive an automated deposit confirmation e-mail including the Persistent Identifier reference for the dataset.

Data files are stored on servers of SARA (Stichting Academisch Rekencentrum Amsterdam); an independent ICT services organisation founded by the University of Amsterdam with high expertise in data storage.

Since 2011, the access portal to scientific information NARCIS has been part of DANS. NARCIS has been developed since 2004, starting as a cooperation project of the Royal Netherlands Academy of Sciences (Koninklijke Nederlandse Akademie van Wetenschappen, KNAW) Research Information, the Netherlands Organisation for Scientific Research (de Nederlandse organisatie voor Wetenschappelijk Onderzoek, NWO), the Association of Universities in the Netherlands (Vereniging van Samenwerkende Nederlandse Universiteiten, VSNU) and the universities' research information and registration system Metis, as part of the development of services within the DARE (Digital Academic Repositories[70]) programme of SURFfoundation (Samenwerkende Universitaire Reken Faciliteiten[71]). NARCIS contains (open access) publications from all Dutch universities, scientific institutes, KNAW and NWO, datasets from EASY as well as datasets from the archive of the data centre of the technical universities (3TU.Datacentrum).

All researchers listed in NARCIS have a personal page providing their contact details and information about their research projects and open access publications. Many personal pages also list the researcher's Digital Author Identifier (DAI). This number makes it easy to integrate information from different systems.

NARCIS provides its users with options for RSS feeds as well as a programming code ('widget') to embed on a website in order to display NARCIS content. Registered users are able to harvest the information in NARCIS by OAI-PMH.

8.2.2 Standards and Formats

Metadata standards

[70] http://en.wikipedia.org/wiki/Darenet
[71] http://www.surf.nl/en/

A dataset in EASY is described in Dublin Core[72] fields, with additional options from Qualified Dublin Core.

Subject and coverage Dublin Core fields for deposits of archaeological datasets offer optional selections from a national dictionary of standardised archaeological codes (ABR, Archeologisch Basisregister[73]).

Upon ingesting a file in EASY, technical metadata is added to each file on the size, the basic mime-type, the file-ID in the Fedora system. The mime-type is identified using Aperture[74].

File-descriptive information can be added to separate files to describe various attributes of the file. Descriptive metadata is submitted by a depositor using attributes from the standards of the Social Sciences' Data Documentation Initiative[75] (DDI) with additions from the geographical sciences' Federal Geographic Data Committee[76] (FGDC). Optionally, non-standardized fields can be added to the file-specific metadata.

File formats

In 2008-2009, a work group at DANS researched various possibilities for storing file types for long-term preservation and accessibility. A list of Preferred Formats was compiled: a selection of file formats which, according to the assessment of DANS, have a high chance of remaining usable in the far future, or have a high chance of being convertible to different formats (without the loss of significant properties) which will be accessible and usable when the original format becomes obsolete.

The list is revised depending on developments in software/technical possibilities or in communities. The latest version of the DANS Preferred Formats list[77] (Table 1) dates from September 2012 and is published on the DANS website.

Two levels of file formats are distinguished in the Preferred Formats list: preferred formats (best choice at this moment) and acceptable formats (good choice). The list is mainly used for recommendations; depositors are encouraged but not restricted to using the formats presented in the list. If it is not possible to deliver files in preferred or acceptable formats, or to convert the files to such formats, then files may still be published in EASY, be it without any guarantees for long-term preservation.

[72] http://dublincore.org/
[73] http://www.den.nl/standaard/166/Archeologisch-Basisregister
[74] http://aperture.sourceforge.net/
[75] http://www.ddialliance.org/
[76] http://www.fgdc.gov/metadata
[77] http://www.dans.knaw.nl/sites/default/files/DANS%20preferred%20formats.pdf

Type of data	Preferred format(s)	Acceptable format(s)
Text documents	• PDF/A (.pdf)	• OpenDocument Text (.odt) • MS Word (.doc, .docx) • Rich Text File (.rtf) • PDF (.pdf)
Plain text	• Unicode TXT (.txt,…)	• Non-Unicode TXT (.txt…)
Spreadsheets	• PDF/A (.pdf) • Comma Separated Values (.csv)	• OpenDocument Spreadsheet (.ods) • MS Excel (.xls, .xlsx)
Databases	• ANSI SQL (.sql, …) • Comma Separated Values (.csv)	• MS Access (.mdb, .accdb) • dBase III or IV (.dbf)
Statistical data	• SPSS Portable (.por) • SAS transport (.sas) • STATA (.dta)	• R (*)
Pictures (raster) (**)	• JPEG (.jpg, .jpeg) • TIFF (.tif, .tiff)	
Pictures (vector)	• PDF/A (.pdf) • Scalable Vector Graphics (.svg)	• Adobe Illustrator (.ai) • PostScript (.eps) • PDF (.pdf)
Video	• MPEG-2 (.mpg, .mpeg, …) • MPEG-4 H264 (.mp4) • Lossless AVI (.avi) • QuickTime (.mov)	
Audio	• WAVE (.wav)	• MP3 AAC (.mp3) (***)
Computer Aided Design	• AutoCAD DXF version R12 (.dxf)	• AutoCAD other versions (.dwg, .dxf)
Geographical Information	• MapInfo Interchange Fomat (.mif/.mid)	• ESRI Shapefiles (.shp and accompanying files) • MapInfo (.tab and accompanying files) • Geographic Markup Language (.gml)

(*) 'R' is currently under investigation

(**) TIFF is regarded as the best archival format for raster images, JPEG is regarded as a preferred publication format
(***) DANS is to be contacted in case of MP3 audio file deposits

Table 1: DANS Preferred Formats, Version 2 – September 2012

8.2.3 Security and Risk Management / Media Monitoring and Refreshing Strategy

All sorts of security aspects are continuously being monitored by the DANS security officer as well as the security officer of the KNAW ICT service (I&A).

DANS participates in (self-)assessments in order to achieve international standards for digital archiving, such as the Digital Repository Audit and Certification[78] (RAC) and the Digital Repository Audit Method Based on Risk Assessment[79] (DRAMBORA). The on-going work within these assessments led to the identification of weaknesses in the DANS archive and the formulation of plans to work towards solving these issues and improving on the services of DANS.

DANS strives to keep up to date with the standards of the Data Seal of Approval. A DSA is valid indefinitely but needs to be updated in order to stay compliant with newly released standards by the Board. DSA repositories are contacted automatically when an update is obtainable.

Media monitoring: All data is stored at an external service provider with high-level security. The central server storage disks are under constant surveillance and are automatically replaced by the storage appliance. Back-up tapes are replaced every 5 years.

The software used within DANS is kept up to date via monitoring by the DANS Application Officer and the Office Manager. If software is used for specific file conversions, the conversion method will be checked and if necessary, revised with a new version of the software.

The archivists at DANS monitor developments in digital archiving, by participating in international conferences, keeping in touch with the designated communities and checking deposited datasets on the emergence of new file formats or standards. As stated under 2.3, the list of preferred formats is revised depending on outside developments. If developments signify the need for new functionality of the EASY archive, the DANS archivists will propose this to the coordinator of the archive, who can accept the proposal and request it for the EASY development.

[78] http://wiki.digitalrepositoryauditandcertification.org/bin/view

[79] http://www.repositoryaudit.eu/

8.3 Data Curation

8.3.1 Pre-Ingest Function

8.3.1.1 Information and guidance given to data producer

The EASY deposit screen includes direct hyperlinks to instructions per discipline and to the main page on the DANS website for information on depositing data[80]. The instructions guide a data producer through the entire depositing procedure.

A factsheet serves as the model for all elaborate instructions by use of the following seven-step procedure for long-term data preservation at DANS:

- Preparing data
- Go to EASY
- Start deposit
- Document dataset (including selecting the access level)
- Upload data files
- Submit dataset
- Publication of dataset by DANS

DANS regularly distribute analogue copies of the factsheets and guidebooks at conferences, symposia and similar events.

During the deposit procedure, a researcher will describe their dataset in Dublin Core fields. Each field is accompanied by a help-icon providing the depositor with extra information and examples for the specific field.

DANS is very active in organising presentations, courses and symposia and also actively participates in many (international) workshops and conferences including workshops and conferences on archiving strategies and metadata. Courses on depositing archaeological datasets have been organised as part of the PASTA (post-academic schooling trajectories for archaeologists) programme and have also been presented to individual archaeological project groups

Members of the DANS staff are always available for questions, comments and additional assistance. The Help and Support[81] section of the website lists specific names, e-mail addresses and telephone numbers for specific subjects.

Additionally, a list of Frequently Asked Questions[82] is available on the website (presently only in Dutch).

[80] http://www.dans.knaw.nl/en/content/data-archive/depositing-data
[81] http://www.dans.knaw.nl/en/content/data-archive/help-and-support

8.3.2 Ingest Function

8.3.2.1 Information and documentation from data producer

Through *self-archiving*, the data producers describe their own dataset in Dublin Core metadata. Six Dublin Core fields are mandatory (Creator, Title, Description, Date created, Access rights, Audience), up to 15 other fields are optional but recommended to promote the visibility and the reusability of the dataset.

Depositors are required to ensure that other researchers will be able to understand their data. A data producer may be required to submit (a) codebook(s) with their dataset. It could be necessary, for example, to list the variables used in a spreadsheet.

Submitted datasets are checked by an archivist for understandability. The depositor is contacted should additional information or codebooks be deemed required.

A depositor can submit a file-list with the dataset for adding descriptive metadata to the files of the dataset. A DANS archivist can deem file descriptive metadata to be required for the understandability of the dataset. Depositors of archaeological datasets are generally required to include a file-list with their dataset.

8.3.2.2 Quality assurance and data checking

DANS archivists work according to a standard protocol in order to provide long-term preservation, findability and accessibility of the data, as well as checking for and anonymising privacy-sensitive data. This protocol[83] is available on the DANS website (presently only in Dutch).

The DANS archivist standard protocol prescribes the archivist to check:

- the completeness of the dataset (files and documentation)
- the readability/accessibility of the files
- the file formats, options to deliver or produce preferred formats or accepted formats (see *2.2*) if other formats are deposited
- the completeness and correctness of the metadata
- if the files or the metadata contain privacy sensitive information
- the clarity of the dataset structure (use of file folders)

The archivist is free to apply small changes to the metadata or the structure in order to improve the clarity or accessibility of the dataset.

[82] http://www.dans.knaw.nl/sites/default/files/FAQ-data-deponeren-in-EASY.pdf
[83] http://www.dans.knaw.nl/sites/default/files/Provenance_document_DEF.pdf

Arrangements can apply that require the DANS archivist to convert the data files to preferred formats. The standard protocol includes an appendix on file conversions, which are all performed to standards. If files are converted, the archivist will ensure that a copy of the original dataset remains stored in the archive.

Datasets in EASY include an administrative section wherein the archivists log their actions with the date when a specific action was performed.

8.3.2.3 Data documentation and enhancement

EASY archivists have full access to the editing mode for the Dublin Core metadata of a dataset and can edit or add information where deemed necessary.

An archivist adds file-specific metadata deposited together with a dataset to the files in EASY. The metadata is delivered in a spreadsheet format; the archivist will convert this file-list to an XML and have it ingested by EASY. Spreadsheets are converted to ingestible XML files using a database export script. The programming script used for having the XML ingested in EASY is written as part of a self-developed toolkit for archivists.

Users can download a dataset's Dublin Core metadata in XML. After adding the file-specific metadata to a file, it can be displayed in EASY by any user by use of the 'View details' button. The file-specific information will also be written into a metadata PDF every time the file is downloaded and added to the downloaded ZIP.

8.3.3 Archival Storage and Preservation

8.3.3.1 Physical data preservation and storage

- Data deposited in EASY is stored on a virtual server at SARA, with two back-up copies on different locations (Amsterdam, Almere).
- A selection of large volumes of specific datasets is stored outside of EASY on servers of SARA subsidiary Vancis.
- Archivists working on archaeological datasets store an extra copy of the original and the processed data on a Networked Attached Storage drive.
- Datasets larger than 100MB are commonly sent to DANS outside of EASY on an external medium (CD/DVD/HD/USB) or via a file-sharing programme (WeTransfer). An archivist will upload the files to EASY. Optical devices are kept in storage.

8.3.3.2 Preservation strategy

The DANS list of preferred formats (see *2.2*) serves as a recommendation to depositors to provide their data in file formats suitable for long-term preservation. While other formats are accepted, depositors are contacted by the archivists and guided to deliver formats included in the list, if possible.

It can be arranged or agreed to that the DANS archivists convert deposited files to preferred formats. Data processing for the deposits of archaeological datasets in EASY have always included file conversions to preferred formats as part of the eDNA project (e-Depot for the Dutch Archaeology).

In the course of the coming two years, DANS plans to start on projects to obtain inventories of the different file formats in the archive and to bulk-convert formats that are archived in non-preferred formats to preferred versions.

DANS activities oriented toward making research data accessible

- Archiving research data and making them accessibly by means of the online archiving system EASY;
- Making agreements with organizations that finance research, give orders for research to be carried out or carry out research themselves, with the purpose of making data available to others;
- Targeted acquisition of research data;
- By participating in research into the need for data archiving in certain disciplines and thus developing contacts among representatives of those disciplines;
- By granting subsidies to small-scale data-archiving projects (KDP – Kleine Data Projecten[84], in Dutch);
- By issuing a quarterly in conjunction with other institutes: e-Data&Research with a circulation of 5000;
- Developing an activity-based cost model;
- Selecting academic heritage and making it available, by means of the ADA approach[85] (Archiving Digitial Academic heritage, in Dutch).

DANS activities oriented toward keeping research data usable

- Development of preservation strategies;
- Converting research data into other formats.

DANS activities oriented toward international standardization and cooperation

- By participating in the development of data infrastructures, for example CESSDA PPP[86], DARIAH[87], CLARIN[88], CARARE[89] and other explorations;

[84] http://www.dans.knaw.nl/content/projecten/kleine-data-projecten
[85] http://www.dans.knaw.nl/sites/default/files/file/publicaties/ADA.pdf
[86] http://www.cessda.org/
[87] http://www.dariah.eu/
[88] http://www.clarin.eu/external/

- Developing and establishing a Data Seal of Approval;
- Developing and operating a Persistent Identifier infrastructure;
- By contributing to the DDI Tools Foundation.

8.3.3.3 Version control/change procedures

When a dataset is published, users need to be able to regard the dataset as being complete. Any changes after publication of a dataset will be regarded as a change in version. Older versions of data files or datasets will remain archived. This also ensures that the Persistent Identifier of the older version will remain valid.

If a data producer deposits a new version in EASY, version numbers will be added to both the new and the old dataset titles. The Dublin Core relations field is used to refer the versions to another. A message will be added to the description of the old version to warn users that the dataset does not contain the most recent version of the data, with a reference to the relations field.

If a data producer submits additional data files to add to an existing dataset, they can be added to the existing dataset provided that additional metadata is included. In case of file replacements, version numbers are added to the filenames. Changes in datasets are documented in the dataset administration section.

8.3.4 Dissemination

8.3.4.1 Visibility

Each file in EASY has settings for its Visibility. An archivist can change the settings. Options for a file’s visibility are:

- anonymous: every user can see the file, even without needing to log into EASY. This is the standard setting for ‘published’ files.
- known: users need to log into EASY before being able to see the file. A message is displayed to anonymous users that a login is required to see all of the published files.
- none: the file is not visible to users. Files kept in EASY for storage purposes may have this setting. If files from a dataset are converted to other formats, the original deposit will be kept in storage with the dataset with its files set to visibility:none.

A depositor will retain the full view of all files of their own dataset, regardless of the visibility settings. DANS archivists have full view of all datasets.

[89] http://www.carare.eu/eng

The Dublin Core descriptive metadata is always visible to all users, even without needing to log into EASY. Similarly, file-specific metadata for visible files can be displayed using the ‘View details’ button without needing to log in.

8.3.4.2 Availability and accessibility

Persistent access to a dataset is guaranteed by means of a Persistent Identifier (PID), which is automatically generated and assigned to a dataset when it is submitted to DANS. The depositor will receive an e-mail with their dataset’s PID. A PID is used for source-referencing a dataset.

The Fedora system for storing digital objects with a dataset-, directory- and file-ID allows for direct access to files from sources outside of EASY, like the use of archived files by enhanced publications.

Like visibility, each file in EASY has settings for its Accessibility. An archivist can change the settings. Options for a file’s accessibility are:

- anonymous: every user can access the file, even without needing to log into EASY.
- known: users need to log into EASY before being able to access the file. A message is displayed to anonymous users that a log-in is required to access the published files. ‘Known access’ is the standard setting for Open Access datasets, as present legal agreements require the need for user authentication.
- none: the file is not accessible to users. This setting may be chosen for large files where it is deemed user-friendlier to provide the data on an optical device (CD/DVD). Files made inaccessible for storage purposes will typical be made invisible to users as well (see *3.4.1*).
- restricted (request): a ‘request permission’ link is provided within the dataset file browser. Before being able to access the file, a logged-in user is first required to send a permission request to the depositor, describing their research title and theme. An e-mail is generated and sent to the depositor who needs to accept or deny the request. The user can only access the file after permission is granted.
- restricted (group): the files are only accessible to users whose account is registered within a specific user group. Archivists assign accounts to groups. This option is used within the discipline of archaeology, where some data is made available only to students and professional archaeologists working under national regulations. The option is mainly chosen in order to keep potential treasure hunters, or others who may damage the archaeological record, from accessing the data. Users can send an e-mail to DANS to request group registration, the archivists will check if the user e-mail belongs to a university or an accepted professional organisation.

The depositor selects a general access rights setting for the entire dataset, but can specify file-specific settings.

Additionally, an embargo for up to two years from the date of submitting the dataset can be placed on the entire dataset. The data files will not be accessible until after the embargo period, after which the specific accessibility setting will apply.

A depositor will retain full access to all files of their own dataset, regardless of the accessibility settings. DANS archivists have full access to all datasets.

8.3.4.3 Tools and interfaces

All datasets in EASY open on a front page or *jumpoff page*, displaying a header of the dataset creator, date and title, followed by the abstract description.

Depositors or archivists may choose to have dataset relations emphasized, which will place the relation in an emphasized box on the jumpoff page.

Jumpoff pages can be edited by an archivist in HTML or TinyMCE, to add extra text or visualisation to the dataset: images, logos, hyperlinks, references to the dataset location on a web-based map, etc.

DANS archivists create overview pages for easy accessing of datasets belonging to the same source or organisation. These pages collect all Persistent Identifier references to the relevant datasets.

Additionally, datasets can be part of a *collection*, wherein datasets are automatically collected based on specified requirements. A collection can be targeted for specific purposes, such as harvesting.

A user of the EASY system can find datasets using a (advanced) search option, which will search on text in (a selected field of) the Dublin Core metadata of all published datasets. Alternatively, a faceted browser is available. Search and browse results can be refined using categories of Audience, Collections or general Access rights, or by additional free text search.

DANS offers a service to track data for users who cannot find what they are looking for: the Datactive[90].

In addition to the EASY archiving system, DANS offers researchers and research organisations the option to publish survey data integrally with documentation via the DANS EASY Online Analysis Tool[91]. This application with deeper access to data offers extensive search options of all variables and queries in a database. Creating and analysing the tables online as well as downloading the data and documentation belong to the possibilities.

[90] http://www.dans.knaw.nl/en/content/categorieen/diensten/datactive-not-found-what-you-were-looking

[91] http://nesstar.dans.knaw.nl/webview/

The DANS Data Portal[92] allows for very specific searching of the Internet for data from several other data collections. It concerns collections of data from the following disciplines: Humanities, archaeology, environmental sciences and behavioural and Social Sciences. At the moment, it is possible to search data collections of about sixty national and international organizations with the Data Portal.

NARCIS serves as a portal to all scientific information in the Netherlands, including the EASY datasets.

8.3.4.4 Monitoring, review and feedback

Since late 2010 DANS is conducting a pilot study to promote the review of research data. Nearly 300 people have so far assessed the quality of a dataset, which they had downloaded from EASY. Within the dataset reviews, a dataset is scored on different aspects, such as the quality of data and documentation and structuring within the dataset. The reviewers state if the data helped to answer their research questions, and whether they would recommend the dataset to other users. A publication on the peer-reviewed research data project[93] is available through the DANS website.

DANS intends to continue the peer-reviewed research approach as well as explore alternatives for adding user review and feedback options to EASY.

Quarterly internal reports are produced on deposit and download/user statistics. DANS evaluates these reports on trends and the course towards reaching its strategic goals on data collection and the data use.

DANS activities oriented toward reuse of research data

- Giving scientific credits to researchers who make their data available to others by means of registration in the universities' research information and registration system Metis;
- Putting persons or organizations in the spotlight that encourage data sharing, by means of a data prize;
- Coupling of publications, data and research information;
- Creation of a demonstrator for enriched publications;
- Linking of data sets to articles in journals. (JALC[94]);
- Linking of components of data sets to articles in journals. DataPlus;
- Organization of symposia around certain data sets.

DANS provides e-mail and telephone contact options for direct user support.

[92] http://dansdataportal.nl/

[93] http://www.dans.knaw.nl/en/content/categorieen/publicaties/dans-studies-digital-archiving-5

[94] http://dpc.uba.uva.nl/cgi/t/text/text-idx?c=jalc

9 The Language Archive[95]

9.1 Organizational Framework

9.1.1 Purpose and Requirements

9.1.1.1 Scope and objectives

The Language Archive (TLA) has been established with joined forces of The Max Planck Society (MPG), the Berlin-Brandenburg Academy of Sciences (BBAW) and the Royal Netherlands Academy of Sciences (KNAW) at the Max Planck Institute for Psycholinguistics in Nijmegen, Netherlands (MPI-PL). The new unit continues and consolidates the well-known work by the Technical Group (TG) at the MPI-PL. The TG functioned already as an archive and technical center of the DOBES program and had an important role in several others developments and initiatives, in particular in CLARIN.

The primary goals of TLA are:

- To store and preserve digital language resources,
- To give access to researchers and other interested users,
- And to develop and integrate new technologies advancing language research.

Although TLA will be primarily grounded on the research needs of the MPG, BBAW and KNAW, it is open to researchers and to all requests for depositing any suitable language related data.

The current focus is on observational data from languages all around the world (typically manually annotated audio and video recordings), such as data from the DOBES program or other data resulting from ethno-linguistic field research, or observational data from language acquisition studies, mainly for major better-studied languages.

TLA also hosts data resulting from experiments: psycholinguistic studies with response patterns to stimuli, eye-tracking data, and more recently also neurological imaging data connected with language production or perception, and even genetic data related to linguistic topics. In principle, any well-structured digital data (no physical objects, data carriers are returned to the depositors) with long-term linguistic scientific relevance are accepted.

9.1.1.2 Collection policy

The process of uploading data to The Language Archive is performed by each user by means of the LAMUS tool, developed at the TLA. It makes use of IMDI for archive management, allows users to upload new resources or resource collections, set access permissions based on

[95] This segment was provided by Przemyslaw Lenkiewicz and Paul Trilsbeek, MPG.

linguistic needs and carries out checks on metadata correctness, on the consistency of all links and on the adherence of resources to the set of accepted formats. In the meantime persistent identifiers (PID) automatically registered with a Handle System server were added to make the references independent of all changes in the storage configuration, i.e. also when new resources or collections are uploaded every object will be associated with a PID and the PID itself is associated with an MD5 checksum information to allow authenticity checks. To strictly maintain achievable formats no encapsulation was accepted, i.e. all resources including the metadata descriptions are stored in standard formats in the file system. Only for fast access purposes databases and indexes are created and used. This makes access to resources and their interpretability completely independent of layered software, which is important for long-term access.

9.1.1.3 Criteria for evaluating data

The Language Archive does not perform the evaluation of the quality of the content to be deposited. The decision of whether to accept a given set of data or not is taken based on the person of the depositor and the research that they are carrying out. In the case of a positive evaluation, the researcher is granted access to the Archive and is free to use it in their preferred way, uploading new content, editing the data and meta-data and so on.

The evaluation of the deposited data happens only on the technical level. The files need to be meeting the format requirements, which are strict. This is important in order to guarantee the future usability of the deposited data and to avoid expensive curation.

9.1.2 Legal and Regulatory Framework

The depositors and the users of The Language Archive need to adhere to the policies of TLA and the Max Planck Institute for Psycholinguistics. These policies may differ depending on the affiliation of the given depositor or user.

In principle the researchers who collected the data retain the right to use the data and to be credited with its collection. TLA will undertake measures to maintain and protect the data for scientific purposes to the best of its ability while protecting the rights of consultants and communities. In principle, all data should be available to the entire scientific community, except for ethically sensitive data or data collected for a PhD program research.

Release of data for non-scientific purposes is not generally permitted. Exceptions are only possible by decision of the Directorate.

For ethical considerations, the Institute follows the guidelines developed in the framework of the project Documentation of Endangered Languages – DOBES. Fieldworkers, archivists, users, and funding agencies have to accept its CoC as the basic guideline for their activities. Any violation of these guidelines will be taken seriously and may have the consequence of loss of permission to contribute to the program or to use the data.

In case that the depositor is a researcher affiliated with the Max Planck Institute for Psycholinguistics, the internal MPI policy for data is applied. It states that all data collected by researchers at the Max Planck Institute for Psycholinguistics belong to the Institute and are archived there. Researchers are obliged to describe the data collected by them and to archive an index/log of them at the Institute. As the meta-data is open to the general public, the data should be described carefully so as not to compromise the rights of the consultants or the relevant communities. Researchers are obliged to describe the data collected by them and to archive an index/log of them at the Institute. The researchers retain the rights to access the data themselves after they leave the Institute. To this end, the Institute will attempt to provide copies of essential material to them after leaving the Institute.

9.1.3 Funding and Resource Planning

The Language Archive is funded by three bodies:

- The Max Planck Society (MPG)
- The Berlin-Brandenburg Academy of Sciences (BBAW)
- and the Royal Netherlands Academy of Sciences (KNAW)

A burst of momentum of language documentation activities of the TLA came with the first call for projects aiming at documenting endangered languages, the DOBES programme by the German Volkswagen Fundation stiftung, in 1999. With its background in building digital data repositories and its experience with valuable data from field research, the TLA was in a perfect position for serving as the central archive of the DOBES initiative which would over the years accumulate terabytes of priceless data on more than 60 languages worldwide. The development and maintenance of TLA Archive and developed tools was financed by a growing number of external research projects, DOBES being one important source among others.

The more and better tools and infrastructure components were developed, the more the group grew, and with more expertise, the TLA was again capable of attracting more challenging projects and to extend and broaden their activities to an international scale – a self-reinforcing process. However, with the eminent end of the DOBES programme (for more than a decade a very reliable and stable source of funding), it became clear that the archive, the tool development and the expertise achieved by the TG could not rely entirely on short-term project funding (although this will continue to play an important role).

The three major funding bodies of the TLA (BBAW, KNAW and MPG) recognized the importance of the language archive and related activities. In an exemplary international cooperation they agreed to give it a more secure basis, providing funds for employing about 7 core members. By now, TLA has more than 20 staff members, most of them still being funded by external research projects.

As for the pure bit-stream preservation of research data, in 2005 a long-term guarantee had been given by the Max-Planck-Gesellschaft for replicas stored at the computer centers – over 50-years, much more than most comparable institutions have. Still, long-term availability requires much more, at least the continuation of the administration of the archive and the maintenance and further development of the core tools needed for its functioning.

9.1.4 Long-Term Preservation Policy

All the software components of the Archive are developed by the TLA team and are open-source. It is possible to obtain the full software solution as a single package and the TLA team assists with the deployment process as well as offers support for the later use.

The data is stored directly in the file system, without the use of any database system and not dependent on external applications.

The sustainability of the TLA Archive is guaranteed by the president of the Max Planck Society and by meeting the requirements of the Data Seal of Approval.

9.1.5 Access Policy

TLA promotes a culture of free sharing of data and believes that in principal and wherever possible data should be made freely accessible via central online repositories. However, the personal and privacy rights of the speakers (and sometimes the intellectual property rights of the depositors) have to be taken into account so that controlled or regulated access may be necessary for certain parts and/or types of data (for instance sensitive material such as sacred rituals). Therefore efficient mechanisms have been created that allow depositors and managers to easily associate rights with individual data objects or branches in a metadata tree, possibly restricted to certain data types (audio, video, annotation etc.).

9.2 Technological Environment

9.2.1 IT Architecture

The Language Archive is based on three essential pillars:

- A data archive holding resources on languages and cultures worldwide.
- Management and access tools developed and maintained in collaboration with a wide variety of projects.
- Archiving and software expertise for collaborative projects.

The IMDI metadata framework was built as a result of discussions between linguists and technologists that satisfied all criteria in so far as it:

- makes use of an element set and vocabularies that emerged from linguistic considerations and semantics
- allows to build hierarchies and collections for management and virtual collection building purposes
- allows to browse in hierarchies and search on descriptions
- offers a gateway to Dublin Core to allow OAI-PMH based harvesting

LAMUS and its components for access management and access requesting are acting as gate keepers for the archive to ensure consistency and coherence, to associate PIDs, to create presentation formats such as MPEG4 for video streaming, to update fast search indexes, etc. A first component called COSIX has been integrated to do data replication and synchronization based on logical level (in contrast to the replication at physical level used for example by rsync), which allows us to properly exchange sub-collections with the regional archives. Together with the DEISA project that brings together the high-performance computer centers in Europe we are working on the REPLIX framework for safe data replication, which is being based on policies at various levels.

A number of web-applications have been developed to be able to access the archived material via the web. The metadata is offered in various ways:

- as the IMDI catalogue;
- via IMDI search (simple and complex),
- as an overlay in Google Earth and
- via a faceted browser in the Virtual Language World.

Metadata selections can be used by these techniques, which then can be used to carry out a content search via the TROVA search engine. As well annotated media streams and multimedia lexica can be viewed and manipulated to a certain extent. VICOS allows users to create conceptual spaces by drawing relations between lexicalized concepts, to navigate in this semantic domain and to open related archived resources from every node.

With the open source TLA software suite we have been developing software components that cover the whole lifecycle of language resources of different types without claiming that (a) these need to be used and (b) they include all functionalities. There are tools to create multimodal annotations for media recordings which can include time series such as eye tracking data, EEG data, etc. as well, complex lexica allowing to include multimedia fragments and syntax trees. The IMDI components allow users to create metadata descriptions that adhere to the IMDI schema and the associated vocabularies. The old IMDI components are currently being replaced by modern tools such as ARBIL that combine metadata creation with organization capabilities and thus increase metadata creation efficiency. In addition we stepped away from a fixed schema approach, but let people now create their own profiles as long as they are using elements registered in ISOcat – the certified concept registry – which is of course important for semantic interoperability.

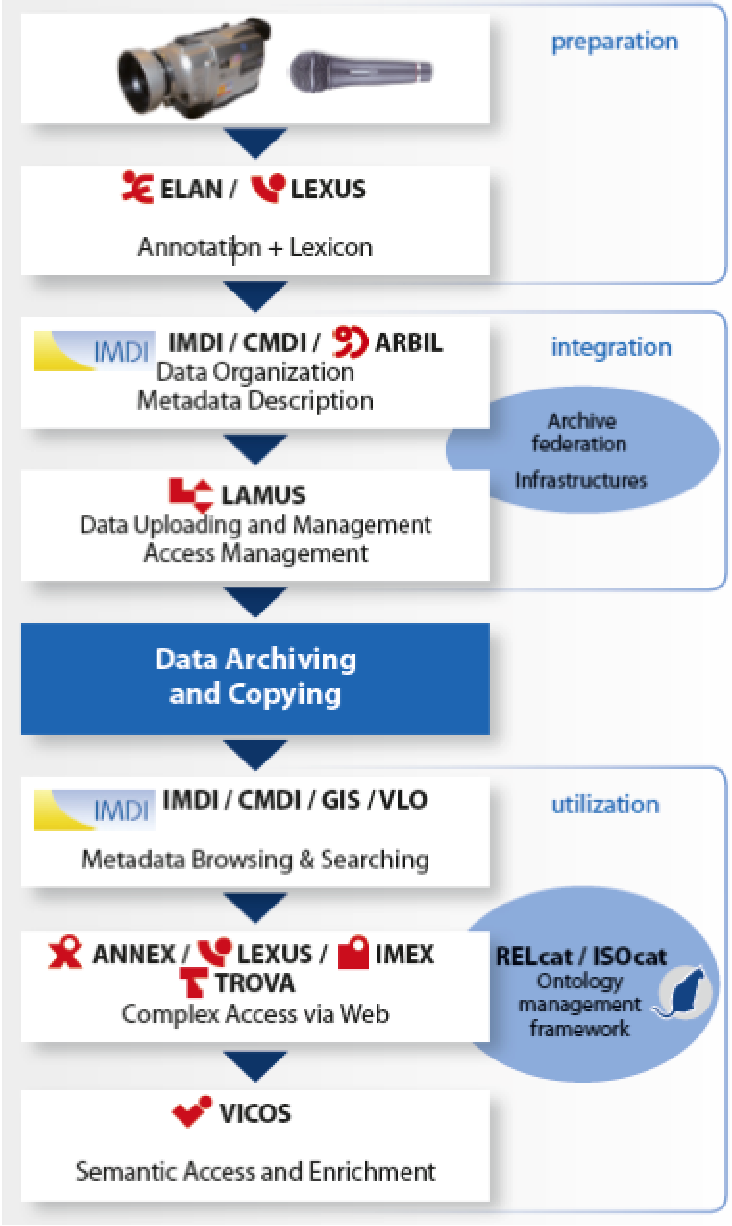


FIGURE 3. THE STRUCTURE AND TOOLS COMPOSING THE TLA ARCHIVE

9.2.2 Standards and Formats

The adherence to standards is very important for long-term interpretability. Here the MPI team participated in particular in the ISO TC37/SC4 committee to work on the following issues: (1) ISO 12620 as a model for registering data categories (formal concepts) and building the ISOcat software to host the definitions many of which have already being entered; (2) Lexical Markup

Framework to have a generic model to represent all kinds of digital lexicons; (3) Establishing principles for associating persistent identifiers with linguistic resources; (4) Defining a set of generic guidelines for annotation formats. Of course widely accepted vocabularies such as the ISO language codes ISO 639-3 are supported.

9.2.2.1 Meta-data

The IMDI Metadata file can describe one of three types of content:

- Corpus – these are only meant to define the structure of metadata, resulting in a tree structure that is reflected in the IMDI browser and Arbil;
- Catalog – these files are used to give an overview over a certain corpus, including statistics;
- Session – these files contain the actual metadata and links to the resources.

The XML Schema is available online at:

http://www.mpi.nl/IMDI/Schema/IMDI_3.0.xsd

Most fields in IMDI sessions are free text fields and for some are restricted by controlled vocabularies. They exist in two varieties:

- Open, where the user may choose from the list, but also add their own values (e.g. 'Genre' or 'Task');
- Closed, where the user can only choose from the pre-defined list (e.g. 'Continent' or 'Gender').

The 'Date' field is expecting values in ISO date format (yyyy-mm-dd), which can be underspecified (i.e. it is possible to leave out the day or day and month). All fields that use a Controlled Vocabulary and the Date field also offer the values 'Unspecified' and 'Unknown'.

The 'Language' fields are using an open Controlled Vocabulary populated with a subset of the ISO 639-3 list. IMDI and TLA provide OAI-PMH support.

9.2.2.2 Standards and Formats

The TLA has restrictions on which file formats are allowed in the Archive. These restrictions are made by the archive management, they are not a part of the IMDI format, even though the main IMDI editor – Arbil – can only automatically recognize the format and pre-populate the metadata if the given format is on that list.

The focus is on open or at least well-documented formats to keep the archive future-proof. We decided to adhere to a number of basic standards such as UNICODE and XML for texts, MPEGx for video representation, linear PCM with high quality for audio streams. The dynamics in the area of video codecs made it necessary to change our strategy 3 times in the last decade. When we started just MPEG1 was usable. Then we turned to MPEG2 as archiving format and are using increasingly often MPEG4/H.264 as presentation format. Recently after deep investigation we have chosen to smoothly turn over to lossless mJPEG2000 to finally have a master format from which we can create other formats without risking concatenation effects.

A detailed list of accepted formats can be found in Table 1 and Table 2.

Type	**IMDI Format**	**MIME Type**	**File Extension**	**Comment**
Audio	audio/x-wav	audio/x-wav	.wav	waveform audio
	audio/x-aifc	audio/x-aiff	.aifc	not accepted for new data, tolerated for legacy data for the moment
	audio/x-aifc	audio/x-aiff	.aiff	not accepted in the archive
	audio/x-mp3	audio/mpeg	.mp3	not accepted for new data, tolerated for legacy data for the moment
	audio/mp4	audio/mp4	.m4a	mpeg4 audio, needs hinted track for streaming
	audio/x-mp2	audio/mpeg	.mp2	not accepted for new data, tolerated for legacy data for the moment
	application/ogg	application/ogg	.ogg	not accepted for new data, tolerated for legacy data for the moment
Video	video/x-mpeg1	video/mpeg	.mpg	mpeg1
	video/x-mpeg2	video/mpeg	.mpeg	mpeg2
	video/mp4	video/mp4	.mp4	mpeg4, needs hinted track for

				streaming
	video/quicktime	video/quicktime	.mov	not accepted for new data, tolerated for legacy data for the moment
	video/x-msvideo	video/x-msvideo	.avi	not accepted in the archive
	application/smil+xml	application/smil	.smil	smil multimedia format
Image	image/jpeg	image/jpeg	.jpg	jpeg image
	image/png	image/png	.png	Portable Network Graphic
	image/tiff	image/tiff	.tiff	tiff encoded image
	image/gif	image/gif	.gif	gif encoded image
	image/svg+xml	image/svg+xml	.svg	Scalable Vector Graphics
Docu-ment	application/pdf	application/pdf	.pdf	Portable Document Format
	text/html	text/html	.html	web page

IMDI Type	**IMDI Format**	**Web server MIME type**	**File Ext.**	**Comment**
Annotation	text/plain	text/plain	.txt	Unstructured annotation file
	text/html	text/html	.html	Unstructured annotation file
	application/pdf	application/pdf	.pdf	Unstructured annotation

				file
	text/x-esf	text/plain	.tr	ESF annotation file
	text/x-chat	text/plain	.cha	chat annotation file
	text/x-eaf+xml	text/xml	.eaf	Eudico Annotation Format (ELAN)
	text/x-pfsx+xml	text/xml	.pfsx	ELAN XML preference file
	text/x-shoebox-text	text/plain	.sht	Shoebox annotation file
	text/x-toolbox-text	text/plain	.tbt	Toolbox annotation file
	text/x-cgn-bpt+xml	text/xml	.bpt	CGN annotation file
	text/x-cgn-lxk+xml	text/xml	.lxk	CGN annotation file
	text/x-cgn-pri+xml	text/xml	.pri	CGN annotation file
	text/x-cgn-prx+xml	text/xml	.prx	CGN annotation file
	text/x-cgn-skp+xml	text/xml	.skp	CGN annotation file
	text/x-cgn-tag+xml	text/xml	.tag	CGN annotation file
	text/x-cgn-tig+xml	text/xml	.tig	CGN annotation file

	text/x-trs	text/xml	.trs	Transcriber is not accepted, should be converted to eaf
	text/praat-textgrid	text/praat-textgrid	.TextGrid	TextGrid annotation file
Primary Text	text/plain	text/plain	.txt	plain text
	text/html	text/html	.html	web page
	application/pdf	application/pdf	.pdf	Portable Document Format
Study	text/plain	text/plain	.txt	Plain text
	text/html	text/html	.html	web page
Lexical Analysis	text/x-shoebox-lexicon	text/plain	.shx	shoebox lexicon file
	text/x-toolbox-lexicon	text/plain	.tbx	toolbox lexicon file
	text/x-cut	text/plain	.cut	chat lexicon files
	text/x-lmf+xml	text/xml	.lmf	Lexical Markup Framework
	text/plain	text/plain	.txt	Plain text
	text/html	text/html	.html	web page
Unspecified	text/x-shoebox-type	text/plain	.typ	shoebox type file (4)
	text/x-shoebox-language	text/plain	.lng	shoebox language file (4)
	Text/x-shoebox-sortorder	text/plain	.set	Shoebox sort order file (4)
	text/x-lexus-config+xml	text/xml	.conf	LEXUS configuration file (4)
	text/xml	text/xml	.xml	XML file
	text/xml	text/xml	.xsd	XML Schema

				file
	text/xml	text/xml	.dtd	XML DTD
	text/x-imdi+xml	text/xml	.imdi	IMDI metadata
	application/vnd.google-earth.kml+xml	application/vnd.google-earth.kml+xml	.kml	Google Earth kml file

Since it is known that curation costs grow over time we apply an immediate conversion policy where possible. Since many tools still do not support standards and are not restrictive with respect to structures the conversion of for example complex lexicons is fairly cost intensive and not feasible without manual intervention.

9.2.3 Security and Risk Management / Media Monitoring and Refreshing Strategy

In order to assure the security of all files stored in The Language Archive each of them is kept in 6 copies, which are distributed among 3 locations in Netherlands and Germany. The Archive is based on a hierarchical file system constructed on disk arrays and magnetic tapes, which are

In order to assure that no information is lost due to deteriorating storage media, each file needs to be touched periodically. This is performed in part by the SAM-FS file system and in part by the TLA corpus management team. Additionally, each file present in the Archive has its checksum calculated and stored upon uploading. Regular checks of the sum are performed for each file.

9.3 Data Curation

9.3.1 Pre-Ingest Function

9.3.1.1 Information and guidance given to data producer

The Language Archive provides precise information about the procedure of file deposit in the Archive. Since this is based on software tools develop by the TLA, this procedure is described in the user manuals for the said tools, as well as in short user guides. These documents can be found on TLA web-site.

Two steps are required when adding a corpus to the MPI Archive. The first involves creating the corpus, editing the metadata and converting it into an archivable format using ARBIL (Archive Builder) tool, while the second involves exporting the created/edited corpus to the archive using LAMUS.

Arbil allows viewing all the corpora in the Archive or any other remote location (provided the user has the access rights).

The Arbil tool allows the following functionality:

1. View the Archive
2. Import metadata from the Archive & edit.
3. Create new corpora, with new metadata and media files.
4. Export the IMDI files, ready to be archived

The LAMUS tool allows depositors themselves to manage their specific part of the archive with as little intervention of corpus- or system managers as possible. Various kinds of resources (e.g. video-, audio data, pictures and annotations) can be linked into the tree structure of the archive, enriching the content of the corpus. LAMUS is designed around virtual workspaces, which provide safe working environments for its users. Managing the resources and the structure of the sub-corpus in this workspace will not affect the actual archive itself. Only when the researcher has finished working on the workspace and the data is submitted, will such data be incorporated into the actual archive. After the transfer, search indexes will be updated for the metadata and the content. During the whole process various checks are being performed on the metadata and on the resources to ensure the consistency and the coherence of the archive.

LAMUS allows the depositors to organize and update the content of their corpora via the web by following functionality:

1. Upload files and add new content to the workspace;
2. Request additional storage
3. Show the content of the workspace.
4. Submit workspace to incorporate it into the archive
5. Delete current workspace
6. Report a bug

9.3.2 Ingest Function

9.3.2.1 Information and documentation from data producer

The depositor is requested to provide a proper description of the files uploaded to the Archive with the IMDI meta-data schema. The MPI Archive can ONLY accept corpora which are in the .IMDI format. By creating a corpus through Arbil, the depositor creates an IMDI file ready to be exported to the Archive. As this meta description is open to the general public, the data should be described carefully so as not to compromise the rights of the consultants or the relevant communities.

9.3.2.2 Quality assurance and data checking

The condition of using only the file formats that are accepted in the Archive is very strict and cannot be broken. The LAMUS tool performs type-checking of the files uploaded to the Archive to assure that the format specified by the depositor is actually true. This is necessary to assure the long-term preservation of the content without the need for expensive curation.

In order to assure high technical quality of the content stored in the Archive, the TLA team organized the DOBES workshops twice every year. They are addressed to anyone interested in digital content creation and archiving it at the TLA.

The scientific quality of the content is not assessed by the TLA corpus management team.

9.3.2.3 Data documentation and enhancement

All the meta-data for the deposited media files is created and managed in the Archive by the depositor. The depositor has full control over the content and is able to use the Archive as their workspace, uploading new media files, adding or changing their meta-data, and so forth.

9.3.3 Archival Storage and Preservation

9.3.3.1 Physical data preservation and storage

The Language Archive operates primarily on a local storage system, which stores two copies of all resources, i.e. at upload immediately two copies are being created. The core of the storage system is based on the SAM-FS hierarchical storage management system[96] which manages:

- fast disk array caches for the small textual resources, indexes etc.,
- slow disk array caches for the media files
- and a tape library based on LT04 technology.

With two servers and a double path SAN configuration single points of failures have been avoided. Since the two local copies will not be sufficient to address seriously the matter of long-term preservation, dynamic copies are created at two large computer centers in Germany at distinct locations. Furthermore, each of these centers has an agreement with another computer center about long-term archiving. Thus all archived objects are available in 6 copies. In order to further decrease the probability of failures, two different dynamic replication protocols at physical level are used: with one center rsync[97] and with the other Andrew File System[98] based exchange are being used. In addition it is important that the president of the Max Planck Society has given a 50 years institutional guarantee for all resources stored at the computer centers.

The archive currently stores more than 80 Terabyte of data contained in about 1 million objects.

[96] https://www.hlrn.de/home/view/System/SamFS
[97] http://rsync.samba.org/
[98] http://its2.unc.edu/dci_components/afs/

9.3.3.2 Preservation strategy

Since all components operated very smoothly and as a whole error free for several years, we can claim that we indeed take care of long-term preservation – at least as good as it is possible these days. This claim was confirmed by getting the Data Seal of Approval after a peer-reviewed assessment of our procedures.

9.3.3.3 Version control/change procedures

The Language Archive stores all the versions of the files ever uploaded.

9.3.4 Dissemination

9.3.4.1 Visibility

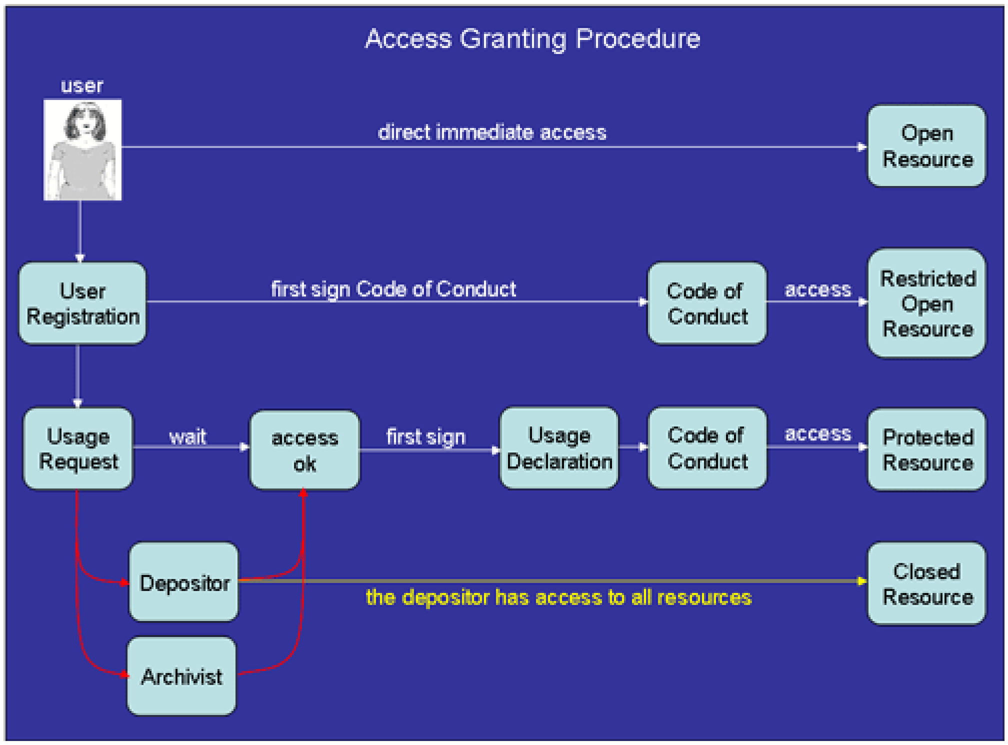


FIGURE 4. FOUR LEVELS OF ACCESS IMPLEMENTED AT THE TLA

All data present in the Archive is visible to any user by terms of the meta-data, which can be found when performing a search or when browsing the IMDI tree. The meta-data is also available through the Open Archives Initiative Protocol for Metadata Harvesting [99] and through the Open Language Archives Community[100] protocols.

[99] http://www.openarchives.org/pmh/

[100] http://www.language-archives.org/

The visibility is naturally restricted to the meta-data and description of the recording, the media itself if controlled with the access granting procedure.

9.3.4.2 Availability and accessibility

Access to the contents of the TLA Archive and most important functionality is provided with web-applications. The software components are tested and documented by their developers and assistants.

Four levels of access have been established:

1. Open resources can be accessed immediately.
2. Restricted open resources can be accessed by registered users which possibly (as in the case of DOBES) have to agree with a Code of Conduct.
3. In addition to the conditions that hold for restricted resources, protected resources can be accessed on request only. The responsible (usually the depositors) will examine the request and, if they grant access, they may do so for a specific use or limited amount of time, which may have to be agreed upon in a usage declaration.
4. Some sensitive “closed” resources can be accessed only by the depositors (and, e.g., members of the respective speech community).

Except for the first, all may require signing a Code of Conduct (such as in the case of DOBES data). Currently, about 25% of the resources are of level (1) or (2). Access to most other material can be requested. These protection mechanisms are seen as sufficient – for various reasons, logo introduction, watermarking, encryption etc. are not applied.

9.3.4.3 Tools and interfaces

The MPI Archive is organized in a hierarchical structure with corpus nodes containing sessions. The sessions, in turn, contain multi-media data, annotations, lexica, etc. The archive can be searched or browsed for annotations. These annotations can then be explored on-line with the ANNEX tool[101], or their contents can be searched through with the help of TROVA[102].

ANNEX provides different views of the annotations (much like the standalone tool ELAN) and can stream media through the QuickTime plugin in the browser. Annotations can be selected, and the selected media intervals can be played. Because it is web-based, it provides a quick and easy way to view annotations directly in the archive without having to download software or files from it. At the moment, annotations can be viewed, but not edited.

Trova is a search engine for annotation content archived at The Language Archive. Searchable formats include ELAN EAF, Childes CHAT, Toolbox, PDF, SubRip, Praat TextGrid and others.

[101] http://tla.mpi.nl/tools/tla-tools/annex/
[102] http://tla.mpi.nl/tools/tla- tools/trova

To use Trova, the user can select ‘annotation content search’ in the node context menu when browsing the archive tree, or do a metadata search and proceed from the search results to a content search in the found sessions and resources.

9.3.4.4 Monitoring, review and feedback

The Language Archive team performs regular checks for statistical overview over the contents of the Archive. The tool developed for this purpose takes a *node id* of a given corpora in the Archive and produces a sorted table as output, which has the following form:

> extension | number of files | total file size (for some video/audio types an estimation of running time)

These statistics can be viewed by any user of the Archive.

10 European Social Survey – Data Archiving[103]

10.1 Organizational Framework

10.1.1 Purpose and Requirements

10.1.1.1 Scope and objectives

The European Social Survey (ESS) is an academically-driven large-scale cross-national social survey designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations. The first round of the survey was conducted in 2002 and it has been conducted every two years ever since. Data are currently available without charge for five rounds covering almost 250,000 respondents from 34 participating countries across Europe.

The Norwegian Social Science Data Services (NSD) has been the official archive for the ESS since 2001. NSD is responsible for providing the ESS National teams with specifications for data and metadata; overseeing the deposit, adjustment, archiving and dissemination of data and documentation for each round of the ESS as well as providing resources to enhance analyses for data users.

10.1.1.2 Collection policy

The 'European Social Survey Project Specification for participating countries' (European Social Survey, 2011) outlines the procedures that countries must adhere to when participating in a round of the ESS. The Specification includes requirements for translation, sampling and fieldwork as well as data preparation and deposit. In order to collect ESS data, a source questionnaire is developed in British English and then translated into languages spoken by 5% or more of the population in a given country. Survey agencies appointed by the National Coordination team in each participating country then collect the data. Interviewers visit selected respondents in their own home (or public venue) to complete the questionnaire. Computer Assisted Personal Interviewing (CAPI) or Pencil and Paper Interviewing (PAPI) can be used depending on the technical capacity in a country. Data from the respondent and the interviewer as well as paradata such as timing variables, call record data (of each contact attempt made at a house/address) and information about neighbourhood characteristics are collected at the time of interview.

[103] This segment was provided by Sally Widdop, CITY. As of November 2012, the ESS had not secured ERIC status. However, the archiving procedures described in this report are likely to be the same under the ESS-ERIC.

10.1.1.3 Criteria for evaluating data

The main criteria for evaluating ESS data and documentation that are deposited to the ESS Archive are adherence to both the general Project Specifications and the data specifications outlined in the Data Protocol. National teams are responsible for delivery of data to the ESS Archive. If the deliverables do not adhere to the specifications in the Data Protocol, the National teams may be asked to supply new deliverables. All files must be deposited online via the 'data deposit' option accessible from the ESS Intranet one month after the official end of fieldwork. The ESS Archive at NSD aims to produce harmonised and standardised data files that are as accurate, consistent and user-friendly as possible. In addition, data should be comparable across countries and time and reflect the original reliability and quality of the data (Kolsrud et al., 2010:63).

There are four stages that are completed when evaluating ESS data, namely:

1. Data ingest and initial checks
2. Data edit, edit control and data approval
3. Processing of metadata
4. Archiving, preservation, feedback and maintenance

More details on these stages can be found in sections 3.2.2 and 3.3.

10.1.2 Legal and Regulatory Framework[104]

Since round 4 of the ESS (2008), the ESS Archive has secured a license from the Norwegian Data Inspectorate, which defines NSD as the Data controller of ESS data at the *International* level. The license also defines the organisations responsible for the collection of data in each country as the data controllers at the *national* level. An additional agreement between NSD and the national organisations regulates the transfer, storage and dissemination of data that could possibly indirectly identify individuals (e.g. non-anonymous raw data and sample design data). All EU and EEA countries are covered by the licence. Data that could indirectly identify individuals is not released publicly, but stored securely by the ESS Archive team in accordance with NSD's licence subject to the Norwegian Personal Data Act and the 95/46/EC Data Protection Directive.

10.1.3 Funding and Resource Planning

The work of the ESS Archive team at NSD has, to date, been funded from Central Coordination grants provided by the European Commission under its various framework programmes. This is true of all seven institutions that constitute the Core Scientific Team of the ESS. Future funding is expected to be secured through the ESS-ERIC.

[104] Text in this section is based on European Social Survey (2011:27-28) and Kolsrud & Kalgraff Skjåk (2010:44).

10.1.4 Long-Term Preservation Policy

The capacity for updating the data and metadata is vital to the long term preservation of the ESS repository. The long-term curation and preservation of data and metadata from new rounds of ESS data will continue under the EU FP7 project ESS-DACE. Under the DACE program, appropriate internal procedures, production protocols, tracking and versioning systems will be retained and enhanced to ensure systematic and continuous improvement in preservation.

So far, 10 years' worth of ESS data and documentation have been securely stored at the ESS Archive. These digital assets have lasting value for both the project team and external researchers and need to be preserved for future users. A migration strategy supported by emulation will be implemented to ensure this. Under this approach, 'digital objects [can be] converted into current or more widely accessible formats' (Kolsrud and Kalgraff Skjåk, 2010). In particular, office documents (e.g. spreadsheets, presentations etc) could be 'migrated to an open format, for example the xml-based Open Document Format for Office Applications (ODF)' (Kolsrud and Kalgraff Skjåk, 2010:46). See section 3.3.2 for more details.

10.1.5 Access Policy

All data and documentation from each round of the ESS are made publicly available as soon as the necessary data integration and data cleaning processes have been completed. "The data are made simultaneously available to all, regardless of their prior involvement in the project, their affiliation or their nation" (Kolsrud, Skjåk and Henrichsen 2007:142). The first data release for a round usually takes place approximately 12 months after the start of the official fieldwork period. The ESS Archive is accessible via http://ess.nsd.uib.no/. All ESS data and documentation is freely available without restriction for not-for-profit purposes. To access data files, individuals have to register as an ESS data user. This requires them to supply a few basic personal details (name, institution, country and type of activity[105]) as well as a valid email address. Once registered, users have immediate access to the data online, and are also able to download files.

There are five 'conditions of use' that apply to those wishing to use the ESS data and documentation. Namely:

1. Restrictions:
   The data are available without restrictions, for not-for-profit purposes.

2. Confidentiality:
   In accordance with data protection regulations in participating countries, only anonymous data are available to users. Before depositing data to NSD, each national team is responsible for checking their data with confidentiality in mind and to undertake the

[105] Faculty and research, PhD student, Student, Government, Organisation (non-government), Private enterprise, Private individual, other.

necessary measures to ensure anonymity of the data files and to foresee that anonymity is also maintained after merging of data files.

3. Citation requirements (for data and for documents):
To ensure that such source attributions are captured for social science bibliographic utilities, citations must appear in the footnotes or in the reference section of publications.
e.g. ESS Round 5: European Social Survey Round 5 Data (2010). Data file edition 2.0. Norwegian Social Science Data Services, Norway – Data Archive and distributor of ESS data.
e.g. ESS Round 5: European Social Survey (2012): ESS-5 2010 Documentation Report. Edition 2.0. Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services.

4. Disclaimer:
The Core Scientific Team (CST) and the producers bear no responsibility for the uses of the ESS data, or for interpretations or inferences based on these uses. The CST and the producers accept no liability for indirect, consequential or incidental damages or losses arising from use of the data collection, or from the unavailability of, or break in access to the service for whatever reason.

5. Deposit requirement:
To provide funding agencies with essential information about the use of ESS data and to facilitate the exchange of information about the ESS, users of ESS data are required to register bibliographic citations of all forms of publications referring to ESS data in the ESS on-line bibliography database at http://ess.nsd.uib.no

From ESS round 4 (2008) onwards, registered ESS data users in countries that have implemented Directive 95/46 “On the Protection of individuals with regard to the Processing of Personal Data” or that have bilateral agreements with the EU, can access the indirectly identifiable (non-anonymous) data according to a special license from NSD. Researchers from elsewhere can get access to the data at NSD. More specific information about accessing the ESS data and documentation can be found in section 3.4.

10.2 Technological Environment

10.2.1 IT Architecture

All ESS data and documentation can be accessed via the ESS Archive website at NSD. “The only requirement for the user to access the ESS data and metadata repository is a standard web browser and an internet connection with modem bandwidth or higher. In other words, the infrastructure can be accessed from anywhere at any time, and the access is not affected by updates or services being performed.” (European Social Survey 2010:44).

The internal networks utilised by the Archive consist of: Microsoft Active Directory Domain Services c-nets (work stations, in-going data protected and three servers, in-going data allowed); with a capacity of 1 GB/s and a firewall administered by the IT-department at the University of Bergen. The external network is UNINETT[106] (member of GEANT, TERENA and PRACE).

10.2.2 Standards and Formats

The ESS Data Protocol is a comprehensive document that outlines the specifications and procedures for data depositors to use when producing national data files. It also defines what files are required and in what format (see Table 1).

Table 1: Accepted data file formats (based on European Social Survey, 2012)

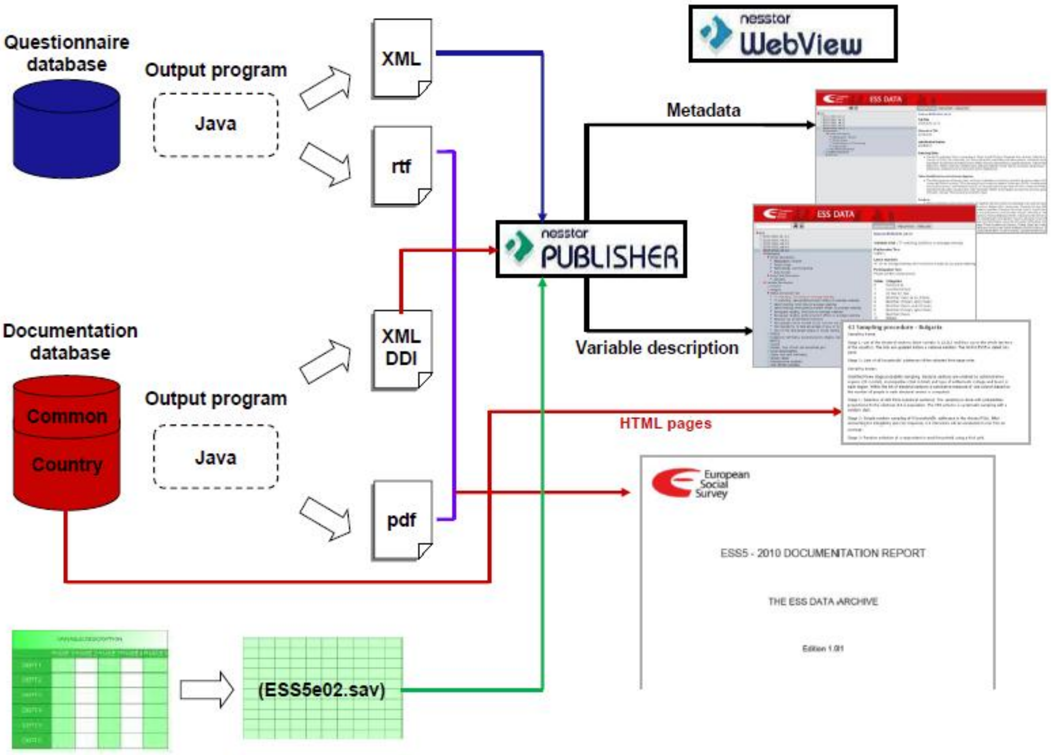
Files	Accepted File Formats (deposited)	Dissemination formats
Main and Supplementary questionnaire data (responses), with any additional ESS variables e.g. timing variables	SPSS system file SPSS portable file SPSS set up, reading data file SAS system file, Windows SAS portable file SAS set up, reading data file	SPSS/PASW & Stata (.por and .sav) SAS, NSDstat, Statistica, DIF, Dbase, Text delimited
Interviewer data (responses to Interviewer questionnaire)		
Contact form data (call records and neighbourhood characteristics)		
Parent's occupation data	As above, delimited text format possible for verbatim recorded answers	
Raw data		
Sample design data file (SDDF)		

ESS metadata are stored in two distinct databases - a 'question database' and a 'documentation database'. Both of these use an MS SQL server and Java Programming. The Java programmes allow various parts of the contents from the databases to be outputted to several file formats

[106] UNINETT: a state owned company responsible for Norway's National Research and Education Network (NREN). NSD is connected to UNINETT.

serving different uses. PDF reports can be made available for end users, XML files and rtf files can be produced for further editing (before being published as appendices to the overall ESS Documentation Report). See Figure 1.

Figure 1: Metadata work flow in the ESS (taken from European Social Survey, Data team (2012:5)



The ESS Data Protocol also defines the documents ESS National Teams are required to deposit to the Archive and in what format (see Table 2).

Table 2: Accepted file formats for documents (based on European Social Survey, 2012)

Documents	Accepted File Formats (deposited)	Dissemination formats
National Technical Summary (documenting the implementation of the ESS; completed by each country team)	PDF form (recommended) Word Processing Format	MS SQL, PDF, HTML, MS Office
Population statistics	Word Processing Format	

Main questionnaire and showcards	ASCII text if text is the only possible output from the CAPI system	
Supplementary questionnaire and showcards		
Interviewer questionnaire		
Contact form		
Interviewer and fieldwork instructions	Word Processing Format PDF	
Advance letters, brochures and other written information to the respondents		
If CAPI: CAPI programmes	ASCII text	

10.2.3 Security and Risk Management / Media Monitoring and Refreshing Strategy

Security - The ESS Archive designates a secure area of the ESS Intranet to the deposit and processing of data files and documentation. This is divided into a common area and a country-specific area for each participating country's National Team (NT). The CST has administrative rights to the common area and the country-specific fieldwork log, and read/download rights to all country-specific areas. The NTs have read/download rights to the common area and upload and read/download rights to their own country-specific area. Access is controlled through assigned usernames and passwords.

All data and documents that are uploaded to or downloaded from the ESS Intranet are transferred using cryptographic protocols (Hypertext Transfer Protocol Secure - HTTPS) 'to ensure a secure channel for transfer of data and metadata' (Kolsrud and Kalgraff Skjåk, 2010: 32).

Refreshing - The current refreshing strategy for the ESS Archive is to review all data and metadata files every second year. Following the review, corrections are made to files as necessary. Any other errors or corrections that have previously been detected are also corrected during this period.

10.3 Data Curation

10.3.1 Pre-Ingest Function

10.3.1.1 Information and guidance given to data producer

The ESS Intranet is a secure area where all the information, specifications, standards and documents needed by National Coordination teams (the data producers) in order to produce standardised and harmonised cross-national data files is made available (Kolsrud et al. 2010). All National teams are required to apply the same specifications to their data as outlined in the ESS Data Protocol. This document specifies the production and delivery of data files, in particular it outlines:

- procedures for collaboration between the National Teams and the ESS Archive
- electronic deliverables required
- principles of variable definitions
- standards and classifications to be applied
- country-specific, identification and administrative variables to be included
- specific variable definitions and coding of data

The variable names, labels, categories and formats found in the ESS Data Protocol are copied into programmes and made available as "empty" data files (dictionaries) in SAS and SPSS format to data producers. For countries using CAPI, these can be used in building the CAPI programmes. They are also used by the National Teams to build data entry programmes and the national data files (Kolsrud et al., 2010). Data Producers are encouraged to apply the Data Protocol variable names, labels and categories to the data files before the files are submitted to the ESS Archive.

International standards are used in the ESS in order to code verbatim responses, such as occupation (ISCO), industry (NACE), citizenship and country of birth (ISO 3166-1 country codes) and language (ISO 639-2). The use of these standards is outlined in the ESS Data Protocol, while the coding standards themselves are available for download from the ESS Intranet (Kolsrud et al. 2010). In addition to the international standards, the ESS utilises ESS coding frames for questions on religion and education, which are included in the questionnaire used during the survey interview and then used to code the data. These are also outlined in the ESS Data Protocol.

10.3.2 Ingest Function

10.3.2.1 Information and documentation from data producer

The ESS National Teams (data producers) are required to deposit electronic deliverables to the ESS Archive one month after the end of the official fieldwork period (see Tables 1 and 2 in Section 2.2). With regard to the deposit of data, National teams (and their appointed survey agencies) are responsible for ensuring that ‘the data are suitably anonymised to comply both with their national laws and regulations and with the ISI Declaration on Ethics’ (European Social Survey 2010: B:29) in other words, any direct personal identifiers are removed from the data before it is delivered. In addition, National Teams are asked to confirm ‘that the data have been anonymised in accordance with national or EU regulations on protection of individuals with regard to the processing of personal data’ European Social Survey (2012:11).

With regard to metadata, all National Teams are required to fill in and deposit a National Technical Summary form (NTS) which documents how the ESS was implemented in a particular country. It includes information about fieldwork dates and procedures used during interviewer briefings as well as information about sample design, response rates, quality control back checks, verification of CAPI programmes and scanning or keying of questionnaires. In addition, country-specific variables such as political parties, education system, electoral system, legal marital status’ and the demographic composition of the population are also recorded.

10.3.2.2 Quality assurance and data checking

All National Teams are required to check their data before it is deposited to the archive. In particular, they are asked to ‘check and edit the data with respect to uniqueness and consistency of identification number(s), wild codes and data consistency’ European Social Survey (2012:10-11). In addition, information about the coding and cleaning procedures implemented by the National Team is recorded in the NTS.

Following deposit of the data files, the Archive team ‘check the data for logical errors, filter errors and for breaches of the ESS standards and specifications [and proceed by] checking and processing the documentation accompanying the data’ (Kolsrud et al., 2010).

In particular, this involves ’normalisation of file formats, evaluation and assessment of the content, structure and format of datasets and documentation and resolving possible confidentiality issues and signing agreements on handling of indirectly identifiable data when necessary’ (European Social Survey, 2010). The following areas are also checked:

- Duplicate or missing identification numbers and consistency of identification numbers across files
- Content of variables present in the data files
- Variable names
- Wild codes (invalid, out-of-range, extreme and missing values)
- Data consistency (limited to the technical quality of data attributes i.e. not responses to attitude or behaviour questions)
- Consistency over time in selected background variables (religious denomination, education level, industry, occupation, language(s) spoken at home, citizenship and country of birth

- Structural consistency in routed variables[107]

During data processing, the National Teams have full access to the data and programmes for their country. Following the completion of the checks outlined above, a data processing report is sent to the National Coordination Teams for their feedback. When agreement has been reached on "how to correct and upgrade the data [...] re-coding of data is done by NSD, or the National Team delivers corrected variables or a complete data file replacing the original. [This] stage is therefore often an iterative process because new data have to be controlled to the same extent as the data originally delivered" (Kolsrud et al., 2010).

Following the completion of stage 1, data edit, edit control and data approval activities take place. This consists of:

- Data edit – recoding of wild codes and corrections of identification numbers
- Data edit – assignment of missing values
- Control of data edit - if a variable has a different distribution after the editing than before, this is reported in the output from the control programme
- Approval of data and integration in the cross-national data files[108]

Once these checks have been completed, a second report is produced for the National Coordinators. At the same time, a draft national data file is produced for the National Coordinator to validate. 'Before the data file is transferred to the National Coordinator, she/he has to agree upon a confidentiality agreement, stating that the draft file isn't disseminated outside the National Coordinator's team' (Kolsrud et al., 2010).

For each round of the ESS, the Archive team are responsible for processing the data files and documentation from the participating countries. 'To ensure as coherent and consistent processing as possible, a processing handbook specifying control and editing rules and a set of common processing programmes have been developed. Decisions about editing of data that are not straight forward are discussed in the group. Consistency is also ensured by the balance between automatic and manual procedures' (Kolsrud et al., 2010).

10.3.2.3 Data documentation and enhancement

As noted in Section 3.2.1, all National teams are asked to complete a National Technical Summary form (NTS) documenting the implementation of the ESS in their country. The information in the NTS is subsequently used to produce the ESS Documentation Report, which accompanies the data when it is released. Although the specific elements adhere to the

[107] Kolsrud, et al. (2010).
[108] Kolsrud, et al. (2010).

international standard Data Documentation Initiative (DDI) Document Type Definition (DTD), the organisation of the items in the Documentation Report differs from the DDI standard in the downloadable report. (Kolsrud and Kalgraff Skjåk, 2010:9).

The Documentation report contains information about the study as well as detailed country-specific reports (based on the information provided in the NTS). This information is enhanced through documentation from the ESS Core Scientific Team and the ESS sampling team.

10.3.3 Archival Storage and Preservation

10.3.3.1 Physical data preservation and storage

At the ESS Archive, metadata, anonymous data, applications, databases, programmes, and production guidelines are stored in c-net 2. Non-anonymous data are stored in stand-alone, password secured PC and as encrypted files in c-net 2 using TrueCrypt.

The ESS Archive uses automated back up routines. Differential back-ups take place every night on weekdays and 'total back-ups' are performed every weekend. Back-ups are stored on tape, both in-house and off-site. The back-up system used is Backup Exec.

10.3.3.2 Preservation strategy

The ESS Archive has adopted a policy on preservation of ESS digital assets. 'Taking into account that the majority of the digital assets of the ESS consist of digits and text, migration will in general be sufficient, and it is also the most cost-efficient approach. By storing the digital objects in generic formats, as the ESS archive to a large extent already does, the task of preservation of the continuous increasing holdings of data and metadata is considered as manageable.' (Kolsrud et al, 2010:79).

10.3.3.3 Version control/change procedures

All processing stages for ESS data are recorded in program files. The input files are retained in their original state and the final versions of output files are also retained. There are two main principles guiding the ESS Archive versioning:

1. Each individual file (data or document) is the starting point for each version. A change in the version number of a file can only occur if there has been a change made to that specific file. Therefore, different files from the same ESS round will have different version numbers[109].
2. If the version number of the Documentation Report and the Main and Supplementary data files are not the same, the Documentation Report must clearly state which data file it corresponds to.

[109] This may seem illogical but it does mean that end users can be sure whether there have been changes made to the files that they have already downloaded.

A detailed versioning system is also applied to distinguish between major and minor versions. A major version (position X) refers to a change in the inclusion of countries in the data file and/or the Documentation Report. A minor version at the first minor level (y) refers to a change in at least one of the variables in the data file (but no change in the countries included in the data file). The second minor position (z) is used for changes in the metadata that did not involve any changes in the data file (X.yz). See Table 4 for an example.

Table 4: Example of file version changes

Original file version number	*Change made*	*New file version number*
1.00	Country added	2.00
2.00	Variable changed	2.10
2.10	Spelling error corrected	2.11

Only the most up to date versions of ESS data files and accompanying documents are made available online to data users. All data files and documents related to that version of the data file are clearly labelled with the edition number, month and year of release. Any subsequent new version of a document also includes a note at the start detailing changes from the previous version. An 'alert' section on the archive website informs data users of amendments that are made to the data files.

10.3.4 Dissemination of data and metadata

All ESS data files and documentation are publicly available from the ESS Archive website - http://ess.nsd.uib.no/. In addition, some project documents and research reports are also available from the main survey website www.europeansocialsurvey.org. A new and integrated website, is currently under construction. This will be available as soon as the ESS-ERIC launches. The focus for this section is only on dissemination from the Archive website.

The ESS Archive website tries to meet different user requirements and in particular their needs for different levels of documentation. It is divided into sections on data download, fieldwork documents, survey documentation, contextual data, and on-line browsing and download. Integrated within this website is the on-line analysis and distribution tool *Nesstar* (Kolsrud et al., 2010).

ESS data are available as direct downloads or through an online analysis tool. In terms of 'direct downloads', large-scale data-sets containing responses to survey questions from individual respondents in each participating country are available as integrated data files for each of the five rounds to date - in either SPSS or SAS format. The integrated data files contain the case and variable counts shown in Table 5 below.

Table 5: Data files and variable counts

Round	**Overall case count**	**Overall variable count**
Round 1 (2002)	42,359	565
Round 2 (2004)	49,066	603
Round 3 (2006)	47,099	517
Round 4 (2008)	58,799	662
Round 5 (2010)	50,781	662
Total	248,104	3009

Country-specific files are also available for each round. Furthermore, individuals can create their own customised data file including 'core' variables from the first four rounds of the survey using a Cumulative Data Wizard. They can then download this file for their own use. Metadata and paradata are also available from the ESS Data website. Individuals can browse through or download documents. All documents are available as PDFs. See Table 6 below.

Table 6: Metadata and paradata available

Metadata *(Information or data that can be relevant to interpret or use the data)*	**Paradata**: *Information or data that can be used to interpret the data collection process*
• Survey Documentation *e.g. Data protocol; Documentation report and appendices; Response based quality assessment report* • Fieldwork Documentation *e.g. main & supplementary questionnaires; showcards; interviewer questionnaire; Fieldwork instructions; contact forms; translated materials e.g. advance letters, brochures and questionnaires for every language fielded* • Fieldwork Summary overview – *Fieldwork dates; sample size; response rate & problems with the data*	• Interview timing data • Data from Interviewer's questionnaire • Call record data (of each contact attempt made at a house/address) • Neighbourhood characteristics data • Sample design data files

In addition, auxiliary (contextual) data is available in the form of a 'Multi-level data resource', which is available for online browsing or multi-level download (see section 3.4.3). Finally, indirectly identifiable data (such as parental occupation) are not made public but are available on request from the Archive (see section 1.2).

In terms of 'online analysis', an electronic codebook and data analysis resource is available via the *nesstar* system (http://nesstar.ess.nsd.uib.no/webview/). *nesstar* is an online analysis tool "that is built on the standardised tag library Document Type Definition (DTD), developed by the Data Documentation Initiative (DDI)" (Kolsrud et al, 2010: 86)[110]. The *nesstar* system enables users to view information about variables; to select specific variables to be displayed as simple frequencies or cross-tabulations as well as produce correlations and conduct regression analyses. See section 3.4.3 for more information.

10.3.4.1 Visibility

Several European data archives link from their websites to the ESS Archive at NSD thereby widening the visibility of ESS data - see for example, the social science data archives for Sweden, Finland, Denmark and the UK.

All members of the ESS Core Scientific Team as well as the National Teams in each country are responsible for promoting the visibility and use of ESS data, documentation and methodology. Current CST outreach activities (funded under FP7 ESS-DACE program) include: promoting the use of secondary analysis of ESS data by academics and to provide supporting materials for public and policy communities (e.g. through dissemination of topline and key findings booklets); maintaining the ESS bibliography as a record of outputs related to or inspired by the ESS (see section 3.4.2); promoting the use of the existing ESS EduNet online series (see section 3.4.2) and to organise a series of policy seminars as a mechanism for knowledge exchange between academics and policy makers.

10.3.4.2 Availability and accessibility

The ESS Archive compiles monthly statistics which record registered users, their activity and the number of downloads amongst other things. The latest statistics from 2 November 2012 reveal that there are currently 51,302 registered users from over 100 countries across the world. Of these, the 'Top 5' countries are Germany (4,938), Belgium (4,484), United Kingdom (3,990), Slovenia (3,053) and the United States (2,863). The statistics show that the most common activity reported by users (when they register) is 'Student' (30,024) followed by 'Faculty and Research) (11,082) and PhD Student (4,490). This clearly demonstrates that the main users of ESS data are from academia.

[110] Refer to http://www.ddialliance.org/what for information about DDI.

The statistics also reveal the number of times specific data files have been downloaded by registered users. The most commonly downloaded file is from the first round of the survey (from 2001) – 13,237 downloads for the 'Static R1' file (see Table 7).

Table 7 – Number of downloaders for each data file

File type	Static Round 1	Static Round 2	Static Round 3	Static Round 4	Static Round 5	Static cumulative rounds 1-3	Cumulative data wizard
N of downloaders	13,237	10,155	9,064	11,288	6,231	4,265	2,890

10.3.4.3 Tools and interfaces

The *nesstar* system[111] (mentioned in Section 3.4) facilitates online analysis of ESS data. It is built on the standardised tag library Document Type Definition (DTD), developed by the Data Documentation Initiative (DDI). *nesstar* allows analysis to be performed on full ESS datasets from a single round or specially prepared ESS EduNet datasets online. Users can choose between 'Description' (of the study, data file, variables, etc.), 'Table' (create frequency tables and cross-tabulations) and 'Analysis' (perform correlation or regression analysis). All results can be exported as either an excel file or PDF.

The ESS teaching resource ESS EduNet can also be accessed via the Archive website[112]. EduNet aims to make it simpler for researchers and higher level social science students to utilise ESS data. It allows theoretical questions to be explored using high quality empirical data from the ESS and by combining theory, data and methodology. ESS EduNet covers a range of substantive topics – e.g. immigration, well-being, family, gender and work, social and political trust and human values as well as methodological areas - e.g. weighting ESS data and regression analysis. Videos are also available explaining how to work with EduNet and how to use *Nesstar*.

The online ESS Bibliography[113] provides information about publications based on the ESS – including analysis of data, methodological research and descriptions and documentation of the ESS. It documents the output of ESS based research, enhancing the visibility of the ESS. The bibliography relies on authors to submit information about their own publications. It is possible to search by author's country, publication type or year published and to download the entire bibliography in Excel format or view it online. As of 24 October 2012, the bibliography held 878 publications.

[111] http://nesstar.ess.nsd.uib.no/webview/

[112] http://essedunet.nsd.uib.no/

[113] http://ess.nsd.uib.no/bibliography/

The ESS Cumulative Data Wizard[114] gives access to cumulative data from countries that have been included in the integrated ESS files in two or more rounds. Users can generate and download a customised subset of the ESS cumulative file by selecting rounds, countries and variables. The data file can be downloaded in SPSS, Stata, SAS and other formats. A Study Description and documentation of the selected variables is generated for each customised data file and is included in the download. All variables that have been asked in more than one round are included in the cumulative data file.

The ESS Multilevel Data resource (ESS MD)[115] contains data at the individual level (the survey data from ESS respondents), the country level and the regional level. It incorporates contextual variables on a number of themes, including demography, geography, economy, health, education and crime. The MD data can be accessed via online browsing or via multilevel download.

10.3.4.4 Monitoring, review and feedback

"Even though data and documentation are handled carefully and with vigilance throughout the data processing, there is always a chance that deviations and errors may prevail in the published dataset or in its documentation" (Kolsrud et al., 2010:79). These errors are usually detected through secondary analysis by data users and reported back to the Coordinator's office based at City University (UK) or to the ESS Archive Team at NSD. Feedback from users can provide an opportunity to improve the quality of the data, metadata and documentation available. The errors detected vary enormously. Table 3 summarises the range of errors detected to date and the solutions available for dealing with these. Note that solutions are dependent on the nature of the error detected and the availability of information to rectify these.116

Table 3: Errors and solutions

Type of error	**Data 'fixes'**	**Keeping records**
Error in variable labels	Correct data in collaboration with National teams / Recode data / Restore data / Remove variable from integrated data file / Make variable available in a country-specific data file only	Add a note to the Documentation report for the specific round and notify users via the 'Deviations and Fieldwork Summary' page on Archive website[117]
Coding errors		
Reversed scales		
Data merging errors		
Inaccurate or erroneous		

[114] http://ess.nsd.uib.no/downloadwizard/

[115] http://ess.nsd.uib.no/essmd/

[116] Text based on Kolsrud, et al. (2010).

[117] http://ess.nsd.uib.no/ess/round5/deviations.html

documentation		

The Archive Team do not release files with known errors unless it is not possible to correct these. Such errors are reported as ‘deviations’. Any errors that are detected between data releases are either released as separate corrected files (if possible), or data users are made aware of the error in an ‘alert’ section of the archive website in anticipation of a new release of the data file(s).

Sources

European Social Survey (2010). *The European Social Survey - Data for A Changing Europe - Annex 1/ Description of Work*, FP7 Research Infrastructures, Internal Project Document.

European Social Survey, (2011). *Round 6 Specification for Participating Countries*. London: Centre for Comparative Social Surveys, City University London.

European Social Survey, (2012). ESS-6 2012 Data Protocol. Edition 1.1. Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services.

European Social Survey, Data Team. (2012) ESS Archive Architecture Report to DASISH WP2, Task 2.1 “State of the Architectures Report” Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services.

Kolsrud, K., Kalgraff Skjåk, K. and Henrichsen, B. (2007) Free and immediate access to data. In: Measuring Attitudes Cross-nationally: Lessons from the European Social Survey, R. Jowell, C. Roberts, R. Fitzgerald and G. Eva. London: Sage Publications, pp.139-156.

Kolsrud, Kirstine. 2009. *Access to Survey Data on the Internet.* Presentation at the Third ESRA Conference. Warsaw June 29–July 3 2009.

Kolsrud, K. and Kalgraff Skjåk, K. (2010) A Strategy Report for the ESS e-infrastructure. The European Social Survey Infrastructure Preparatory Phase (ESSPrep), Deliverable 18.

Kolsrud, K., Midtsæter, H., Orten, H., Kalgraff Skjåk, K., and Øvrebø, O. (2010) Processing, Archiving and Dissemination of ESS data. The Work of the Norwegian Social Science Data Services ASK Research Methods 19, 1 51–92.

Skjåk, Knut Kalgraff. (2007). *Clean Data or Cleaned Data? Data Editing Procedures and*

Experiences of the ESS Data Archive. Presentation at the Second ESRA Conference, Prague June 25–29 2007.

11 Survey of Health, Ageing and Retirement in Europe – Data Archiving[118]

11.1 Organizational Framework

11.1.1 Purpose and Requirements

11.1.1.1 Scope and objectives

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks of more than 60,000 individuals from 19 European countries and Israel aged 50 or over.

The data are available to the entire research community free of charge. The data is deposited at two data archives. Applications can either be submitted through the SHARE website (www.share-project.org) or the Data Archive for the Social Sciences, run by GESIS in Cologne, Germany. The administration of the data archive at the SHARE website is managed by CentERdata.

The data is collected by survey agencies appointed by the local scientific committees in SHARE. To collect the interview data, software developed by CentERdata is used to administer the questionnaires. This software also collects and sends the data to a central server at CentERdata. Here the first data processing is done.

11.1.1.2 Collection policy

The SHARE data is collected by local survey agencies that are appointed by the country teams of SHARE. Interviewers visit respondents to fill in questionnaires, the answers given by the respondents are stored in data files. Next to the data collected in these interviews, para data and process data is collected by the SHARE software and the survey agency administration.

SHARE makes for every country team and participating agency a software package available that includes the fieldwork management and questionnaire software in national language(s). Next to the software several manuals are offered that describe how the data collection should occur. All countries participate in so-called train the trainer sessions in which the fieldwork management is trained to train the local interviewers.

Most of the information and documentation from the data producer is included in the exports of the SHARE fieldwork management software. For example, interviewers can add remarks to

[118] This segment was provided by Eric Balster, CenterData.

questions in the interview while the interview is administered. But also local administrators can add information to fieldwork cases. These are all collected in a central uniform way. Next to this automatic collection, the central coordination team asks fieldwork agencies to supply them with an evaluation form of the fieldwork. This form is used to evaluate the complete fieldwork process.

11.1.1.3 Criteria for evaluating data

During and after the fieldwork in SHARE several quality checks are done. Among these are interview and individual question time length check as well as checks whether or not the standard procedures were applied properly. Amongst others, keystroke records and random check-calls are used to assess these quality issues.

11.1.2 Legal and Regulatory Framework

In March 2011, the Survey of Health, Ageing and Retirement in Europe (SHARE) became the first European Research Infrastructure Consortium (ERIC). This gives it many of the same advantages and tax exemptions enjoyed by major international organisations.

Although SHARE-ERIC is hosted by Tilburg University/Netspar in the Netherlands, SHARE is centrally coordinated at MEA (Munich Center for the Economics of Aging), Max-Planck-Institute for Social Law and Social Policy, Germany. The project aims to help researchers understand the impact of population ageing on European societies and thus to help policy makers make decisions on health, social and economic policy.

Austria, Belgium, the Czech Republic, Germany and the Netherlands are the founding members of SHARE-ERIC, with Switzerland having an observer status. Italy joined in June 2011. Denmark, Spain, France and Portugal are expected to follow soon.

The SHARE country team leaders are responsible for observing legal requirements such as data confidentiality and safety regulations in their country. Only the survey agencies who administer the fieldwork have contact information like addresses and names. These are never collected by the central SHARE co-ordination team.

Privacy related information will be removed from the data as soon as it leaves the survey agencies. Within the agencies only the interviewers themselves can view the answers given by the respondents. Sample managers can only see the addresses and some statistics about the progress of the fieldwork, not the answers themselves. Country team operators can only see the interview data and not the address data of respondents.

11.1.3 Funding and Resource Planning

The SHARE data collection has been primarily funded by the European Commission through the 5th Framework Programme (project QLK6-CT-2001-00360 in the thematic programme Quality of Life), through the 6th Framework Programme (projects SHARE-I3, RII-CT-2006-062193, COMPARE, CIT5- CT-2005-028857, and SHARELIFE, CIT4-CT-2006-028812) and through the 7th Framework Programme (SHARE-PREP, N° 211909, SHARE-LEAP, N° 227822 and SHARE M4, N° 261982). Additional funding from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, R21 AG025169, Y1-AG-4553-01, IAG BSR06-11 and OGHA 04-064) and the German Ministry of Education and Research as well as from various national sources is gratefully acknowledged (see www.share-project.org for a full list of funding institutions).

11.1.4 Long-Term Preservation Policy

All SHARE data (including questionnaires, setup packages, data and paradata) is currently stored at the SHARE archive located in Tilburg. This archive is regularly backed-up by servers of the Dutch Supercomputing Center for Scientific Research (SURFsara).

Next to this, all data is archived by the Data Archive for the Social Sciences, run by GESIS, located in Cologne. Regarding this, we refer to the GESIS chapter in this document to GESIS's long-term preservation policy.

11.1.5 Access Policy

Access to the SHARE data is provided free of charge to all scientists globally, subject to European Union data protection regulations. The SHARE data is available for registered users via the website www.share-project.org. Registered users can download the data of Wave 1-4 in various formats.

To register as a SHARE data user, the following conditions have to be met:

1. Applicants must have a scientific affiliation and have to sign a statement confirming that under no circumstances the data will be used for other than purely scientific purposes.

2. Data will only be made available after these documents have been received, by mail or fax (care of Josette Janssen; address: CentERdata, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands; e-mail: jjanssen@uvt.nl; fax: +31 13 4662764). The required forms can be downloaded here. Upon request a copy by fax can be sent. Upon receipt of the signed statements users will receive a username and password enabling them to download the data. Registration is free of charge. Registered users are allowed to use data of the SHARE project as long as the scientific affiliation indicated in the user

statement is valid. The original login code and password persist for all subsequent releases of the data. A new statement has to be filled, however, when any of the specifications given in the statement (incl. e-mail address) change. If users forgot their password, they should go to http://centerdata.nl/link/sharedata. A password can however only be resent if the e-mail address typed in is the same as used when applying for the data. Additionally, the data are available via the GESIS Data Archive for the Social Sciences.

3. Data users are not allowed to make copies of the data available to others and/or enable any third party access to the database. Anyone wanting to use the data must contact CentERdata directly to request a copy of the data free of charge.

4. Users are requested to provide references to all papers based on the SHARE data to the SHARE co-ordination team. Whenever a paper is being written using SHARE data, a disclaimer and an acknowledgement have to be included in the following form:

5. "This paper uses data from SHARE wave 4 release 1, as of November 30th 2012 or SHARE wave 1 and 2 release 2.5.0, as of May 24th 2011 or SHARELIFE release 1, as of November 24th 2010. The SHARE data collection has been primarily funded by the European Commission through the 5th Framework Programme (project QLK6-CT-2001-00360 in the thematic programme Quality of Life), through the 6th Framework Programme (projects SHARE-I3, RII-CT-2006-062193, COMPARE, CIT5- CT-2005-028857, and SHARELIFE, CIT4-CT-2006-028812) and through the 7th Framework Programme (SHARE-PREP, N° 211909, SHARE-LEAP, N° 227822 and SHARE M4, N° 261982). Additional funding from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, R21 AG025169, Y1-AG-4553-01, IAG BSR06-11 and OGHA 04-064) and the German Ministry of Education and Research as well as from various national sources is gratefully acknowledged (see www.share-project.org for a full list of funding institutions)."

6. Registered users of the data will be included in the list of users of the SHARE project. By signing the user statement users agree to be informed about updates of data via e-mail.

7. In case of doubt whether or not the data have been used for purely scientific research, the Coordinator of SHARE will decide whether the password will be removed and legal action will be taken.

11.2 Technological environment

11.2.1 IT Architecture

Several survey agencies in different countries collect data for SHARE. To enable these agencies to do this fieldwork in a harmonized way, SHARE supplies the agencies with software to administrate the fieldwork and interview the respondents. This software package exists of two programs: a Sample Distributor (SD) and a Sample Management System client (SMS).

The Sample Distributor is the program used on the central office of a survey agency. With this program the sample can be managed and assigned to interviewers to interview the respondents. The SMS client is the program that is installed on the interviewer's laptop. Once some sample is assigned to an interviewer laptop, the interviewer can register the interview workflow paradata (like contacts, contact attempts, changes to the sample) and perform the interview with the respondents. The SMS can communicate with the local survey agency through a (secure) FTP connection. Once the SMS is synchronized, the SD administrator can view the progress and send the interview information to the SHARE servers.

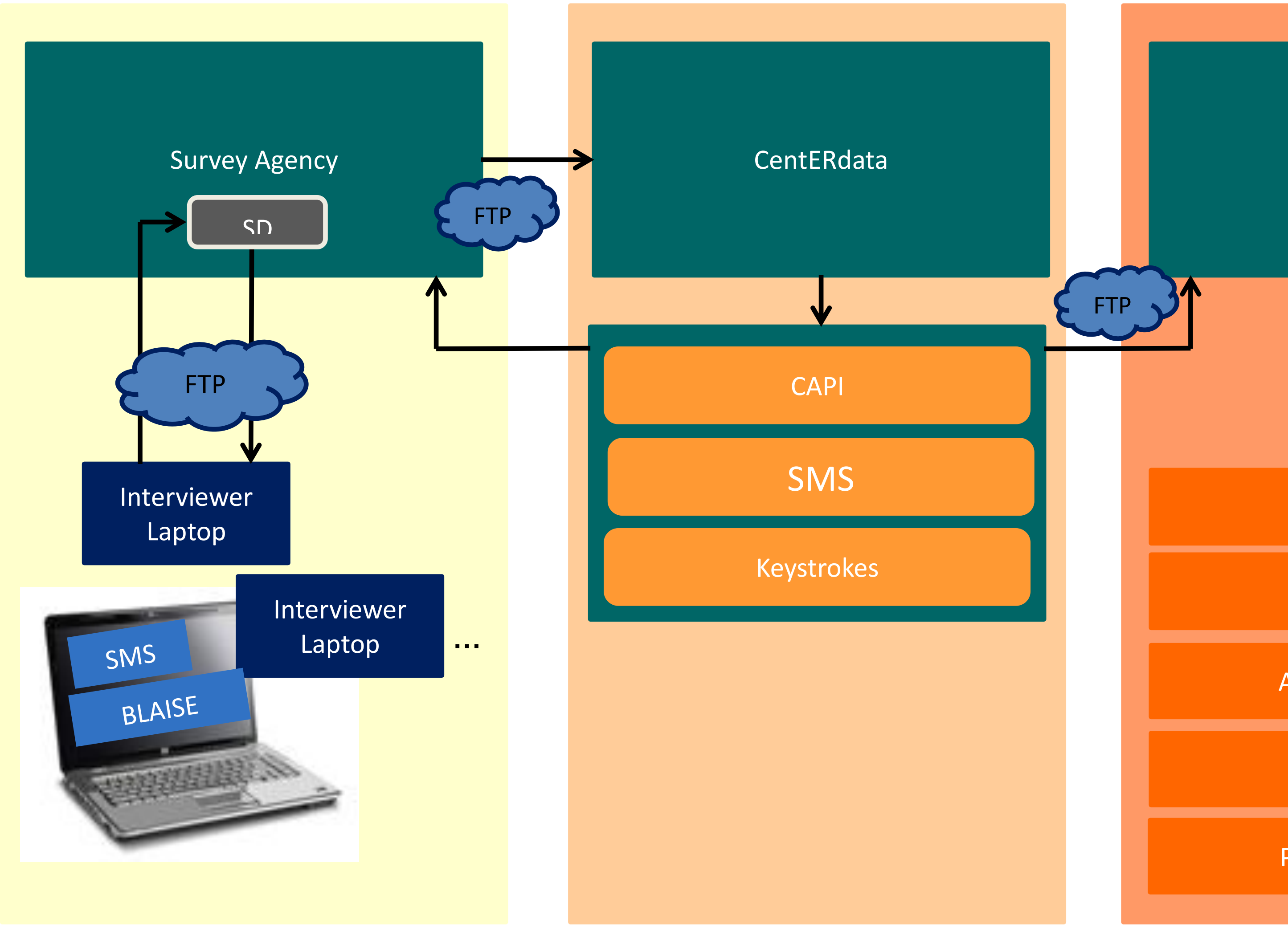


Figure 5 Overview of the SHARE data collection IT architecture

Once the data is received at the CentERdata servers it is immediately archived. Once the data is archived, it can be processed to several data sets. These include the interview data (also called CAPI files in SHARE) in SPSS and STATA format, keystroke data in SPSS and STATA format and SMS data (contact information) in excel and SPSS format.

These files are used by the central coordination team at MEA to make the different releases for both the other SHARE teams and the public data users.

11.2.2 Standards and Formats

The only file format accepted for the input is the format defined by the SHARE Fieldwork software package. This data is used to generate other data files like SPSS, STATA and text-files. In SHARE we have the following types of data:

1) Main and supplementary questionnaire data (CAPI data)
    - Interview data are released up to wave 4 data
    - Generated variable modules
    - Weights and imputations
    - Available for scientific use
2) Paradata
    - Keystrokes (interview length, time stamps, time for reading out items)
    - SMS data (eg. call record data, area information, refusal information, screening process if applicable)
    - Interviewer profiles (demographics of interviewer and information on previous interviewing experience)

      [Note: all information mentioned in this paragraph is not made available for scientific use. Preparation and Publication is work in progress.]
    - Duration of fieldwork
3) Metadata
    - Information on the sampling design is available for each country separately (Börsch-Supan & Jürges 2005, Malter & Börsch-Supan 2013)
    - Fieldwork documentation (questionnaires, sample size, sample composition, response rates are partially available on www.share-project.org, and are published in Börsch-Supan & Jürges 2005 and Malter & Börsch-Supan 2013)
4) Auxiliary Data
    - An initial collection of contextual data on country level can be found online (the collection was part of SHARElife and will be updated in the future)

11.2.3 Security and Risk Management / Media Monitoring and Refreshing Strategy

Security in SHARE is very important because personal information is collected. For this reason several measures have been taken. Among these are that all communication between different computers and servers in the collection process is done via encrypted files. These files can be only opened by entitled entities. Also the information like names and addresses is filtered out of datasets before they leave the Sample Distributor server. For this reason no address information is available in the data stored at the SHARE servers.

All data is archived in encrypted packages that are backed up on a daily base. Only a special designed tool is able to unpack these data files. This ensures that this data cannot be altered. For users that don’t have remote desktop access, the only way to download the data is via a username/password protected ftp or http/https connection. FTP upload is only allowed for encrypted new data. The deletion and alteration of data is not allowed.

11.3 Data curation

11.3.1 Pre-Ingest Function

11.3.1.1 Information and guidance given to data producer

The fieldwork of every country is managed by a program called the SHARE Sample Distributor. In this program fieldwork managers can control the fieldwork by assigning and unassigning cases to interviewers in the field. Also the system helps to monitor the progress of the fieldwork. On the background the system collects the interview data from the different interviewer laptops in the field. This data is bundled into exports to CentERdata. This central organization processes the data to be ready for the next steps in the data life cycle.

11.3.2 Ingest Function

11.3.2.1 Information and documentation from data producer

Next to the data that is automatically collected by the Sample Distributor, agencies and country teams are asked to supply a sample design document if they use a refresher sample for their SHARE fieldwork. On the end of the fieldwork period the agency is asked to complete a questionnaire about how the fieldwork process went. The results of this questionnaire are used to optimize the SHARE data collection process and the training sessions.

11.3.2.2 Quality assurance and data checking

After every wave, several actions are done to improve the data quality. Amongst these are:

- Updating of gross sample information after fieldwork;
- Checking and cleaning of gross sample and SMS data;
- Detailed reports of all problems that were documented during the fieldwork, e.g. problems with preload data, exchanged/wrong IDs;
- Confirmation of IDs and demographic information;
- Checking interviewer remarks and any open questions for relevant information;
- Checking and cleaning of CAPI and ‘drop off’ data;
- Coding of variables and other information;
- Providing input for and checking the results of imputations;
- Providing any information or data that is needed for weights.

Only after these actions have completed, the data will be made available for public release.

11.3.2.3 Data documentation and enhancement

Due to their cross-national, multidisciplinary and longitudinal nature, the SHARE data were complex from the very start. With the release of SHARELIFE that focuses on people's life course histories, complexity was even extended. As the need for information increases with data complexity, user support becomes more and more necessary in the course of the project. SHARE has reacted to these needs and makes all relevant information accessible via www.share-project.org. Documentation of every wave and release is provided under "Documentation". For a quick but extensive overview of the data, SHARE gives answers to all typical and frequently asked questions under "Frequently Asked Questions: About the SHARE Data Set". This contains basic information about accessibility, data structure, countries, content variables, merging, loops, weights, imputations etc. Under "Sample", users can find information about the sample size for each wave by country and sex as well as household and individual response rates.

The SHARE documentation concept is organized in a three-fold structure:

a) SHARE Guide to Release X
b) Item correspondence tool
c) Tailored user support

a. The SHARE Release Guides contain all relevant information for working with the respective release X of the data: participating countries, eligibility rules, the additional drop-off questionnaires and vignette studies, as well as general issues in the composition of datasets and types of respondents. It also provides information on how to merge different modules and across panel waves, and on how to merge SHARELIFE data with wave 1,2 and 4. The guide furthermore documents treatment of missing codes, the conversion of currencies into comparable Euro values as well as unfolding bracket questions and multiple answer questions into "dummy" variables. There is helpful information on specific issues like coding open answers, nationality or country of birth. As a cross-national project, SHARE has to deal with different institutional contexts (see "Item Correspondence" below). Furthermore, the release guide holds information on the additionally generated datasets on imputations, weights, housing, health, social support and household composition as well as alive-status. Therefore, it is also documented how the generated variables ISCO, ISCED and NACE were coded.

   The original questionnaires, as a basic part of survey documentation, are also available on the website. For every wave, the originally used instruments are to be found online, either in the generic English version or in specific language/country versions. They hold all the technical information on filter rules, interviewer instructions, accepted answer ranges and looping rules.

b. As usual for an international panel survey, there are cross-national deviations of questions and/or answer categories as well as differences between variables across waves. SHARE provides an online tool for an easy overview on item correspondence across countries. It allows the user to generate specific views on single countries, modules or questions. Currently, there are two cross-sectional correspondence tools available for the two released waves yet that document country specific deviations, as well as a third tool that documents longitudinal changes in the generic questionnaires between the two waves. For all deviations, an English translation is provided.

c. Finally, tailored user support by SHARE is designed to help users to exploit the richness of the data. This is done by email contacts to the central database management team of SHARE in Munich, Germany, as well as to all the teams in the different countries participating in the SHARE project. This leads to a two-fold user support: on the one hand support by the SHARE team that implemented the survey in the respective country, and on the other hand support by the central team that was in charge of producing the released data version. Most user questions are answered within one week or less by both the central SHARE team and the different country teams. Questions on specific issues are directed to the respective working group. Additionally, SHARE organizes regular scientific user conferences, where both users and members of the various SHARE teams meet and provide comments and assistance, as well as data workshops at international conferences and universities. To further enhance user support and documentation, the central database management team sent around user feedback questionnaires concerning technical as well as content issues before wave 4. This feedback was very positive, especially concerning the smooth data access and comprehensive documentation. Also, generated variables and SHARE modules in general were highly appreciated. Suggestions such as the implementation of codebooks and a deeper insight into the technical background (e.g. imputations) were realized by the database management team.d. A generic and French codebook for wave 2 was made by IRDES with inputs from MEA.

Moreover, SHARE provides several publications that give further information on methodology and sampling, all available online via www.share-project.org -> “Publications” and is about to set up a new regular working paper series on technical SHARE issues -> “SHARE working paper series”.

A version 1 of a generic QbyQ (question by question instruction, definitions, explanations, background for questions) has been prepared by the area coordinators. They are be translated/adapted in each country as suits the need of the local team and the use of the interviewers before and during the interview. A QbyQ is the source of many improvements:

(1) the question itself might be better framed once it is better understood (caveat: we want to keep longitudinal comparability).

(2) the short IWER instruction appearing on the screen will be made clearer.

(3) The training will be easier and more precise as the trainers will be able to answer interviewers’ questions or concern about the questions.

(4) The manual will have a QbyQ section, or a separate QbyQ manual can be edited and printed in each country.

(5) The data users will be aware of the meaning of questions.

As such a QbyQ is integral part of the documentation of any survey. In SHARE it gets another layer because of country specificity. It plays in many directions:

(1) Translation issue: a technical generic word has to be translated by a technical country specific word, most importantly all those concerning income/pension/benefits/assets. For this purpose there is no need to translate the QbyQ.

(2) What was formerly known as item correspondence will be more easily documented from the country specific QbyQ. So part of them will have to be re-translated into English, to the extent that they are useful for data users. In SHARE, because of the various institutional contexts, we cannot get for granted that meaning of questions is straightforward. QbyQ provides some country institutional background.

11.3.3 Archival Storage and Preservation

11.3.3.1 Physical data preservation and storage

Once the data leaves the survey agencies, CentERdata is responsible for the archiving and processing of the files. This however is not the final data that is published for scientific use. This data is first cleaned by SHARE and only after this stored at the CentERdata Archive and the GESIS Data Archive for the Social Sciences.

Besides that the data is stored at two archives, the CentERdata site is backed up by a service provided by SARA, an independent organization which supports scientific research by offering the most advanced services and expertise in the areas of computing, data storage, visualization, networking, cloud and e-Science. At the GESIS site also backup strategies are applied.

11.3.3.2 Preservation strategy

All data in SHARE is stored on both the SHARE Archive at CentERdata and the GEISIS Data Archive for the Social Sciences in Cologne. At the CentERdata Archive the data is available in different forms, ranging from the encrypted source packages to the data files suitable for data users to download and analyse. All files are also backed up by an external backup system located at SARA.

11.3.3.3 Version control/change procedures

In SHARE version control exists in a simple way. Datasets get labeled version numbers that point to different stages in the data life cycle.

11.3.4 Dissemination

11.3.4.1 Visibility

Visibility of the SHARE data is important. Many channels are used to increase the exposure of SHARE data, not least the project`s main website with rich information about the data and results and web 2.0 (twitter, facebook). Among other channels are conferences, booths and workshops where SHARE teams present data and results, , regular newsletters to all users and supporters, information material like brochures, posters and flyers, press information and reviews, - and external links like for example on the GESIS or European Commission sites.

11.3.4.2 Availability and accessibility

Our focus on the needs of users encompasses a release policy that gives free, speedy and convenient access to all world-wide scientific users subject to EU data confidentiality restrictions which must be signed prior to access. Data has been released extremely timely (the first preliminary release took place about 6 months after data collection, a second cleaned and thoroughly documented release about one year later).

The data is deposited at two data archives. Applications can either be submitted through the SHARE website (www.share-project.org) or the GESIS Data Archive in Cologne, Germany.

The typical new user will visit the SHARE website and will from there submit a request for data access at the University of Tilburg, where the user platform is physically located. First, we check the scientific background of the applicant. This will be done upon faxing a letterhead of the involved academic or publicly funded research institution, together with a short description of the institution and a signed form declaration of data confidentiality according to the form provided by the European Commission (Official Journal of the European Communities, L6/52, 10.1.2002). Second, upon acceptance of the credentials (normally within 24 hours on working days), access will be given to a secure website via a user ID and a password. Third, the data can then be downloaded to the user’s institution.

Users are supported by the SHARE website with all public information. It includes

(a) the questionnaire in all languages in a synoptical form, using a new software that is currently being developed. This software directly uses the CAPI instrument such that changes in the instrument are immediately reflected in the user interface,

(b) “metadata” such as technical data (e.g. response rates) and institutional data (e.g. pension replacement rates) for all participating countries.

The data dissemination system will enforce three stages of access restrictions. Raw data from experiments and pilots will remain strictly confidential. The recoded data sets will be purged from sampling identifiers and have restricted access according to national requirements. Publicly available data will be factually anonymised.

SHARE is a distributed website with no physical site. Hence, there is no restriction on the number of data users at any time once a user has downloaded the data and uses at the user’s institution. After queries and download, there is no “crowding” or “interdependence” and no “scheduling” is necessary. There is also no need for a peer review procedure selecting users and no need for the installation of a User Selection Panel.

11.3.4.3 Tools and interfaces

Regarding the access to SHARE data, further steps may be appropriate with regard to considering more up-to-date technologies like, e.g., the use of electronic signatures. Furthermore, new ways of data dissemination involving tools like, e.g., Nesstar should be considered for the future and, in particular, regarding “*easy*SHARE”.

11.3.4.4 Monitoring, review and feedback

In December 2012, seven years after the first public release of the SHARE data base to the scientific community, nearly 2900 users had registered as users at the SHARE Research Data Center located at the University of Tilburg. It is remarkable that the speed of registration has not slowed down but rather increased: currently, some 35 new users register each month (see Figure).

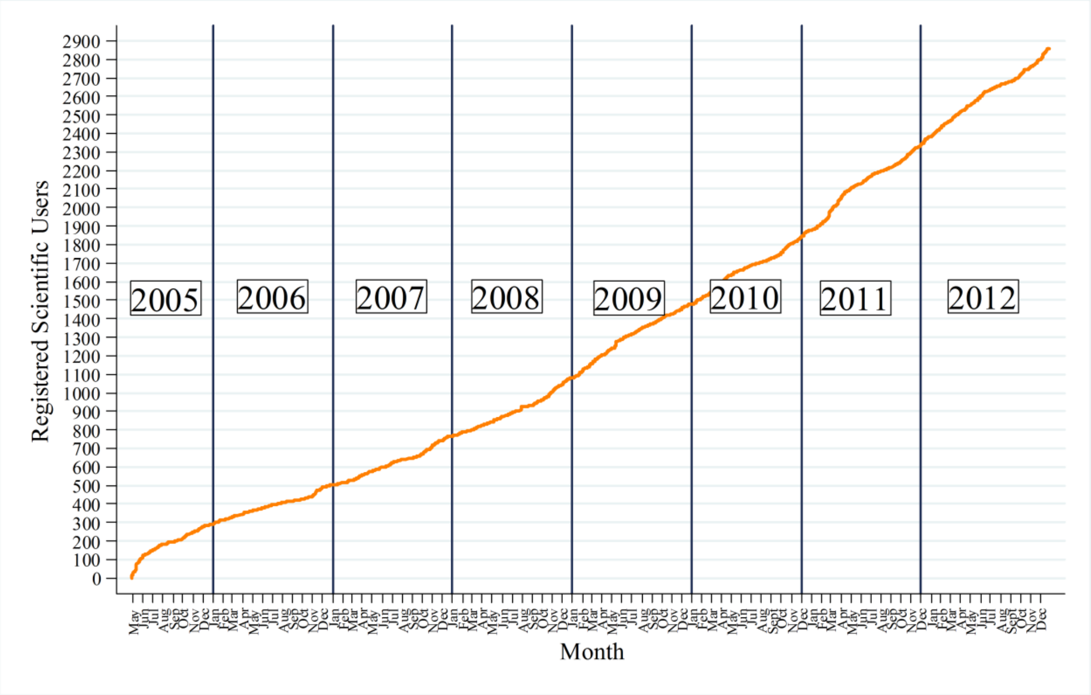


Figure 2: SHARE user registrations since April 2005

As demonstrated in Figure 3, the SHARE database has attracted users from all over the world. Users from a total of 44 different countries have so far registered. SHARE data are primarily used in SHARE countries, most notably Germany and Italy, but also in new member states of the EU such as Poland and the Czech Republic. A significant number of registered users is based in the US and the UK where more than 200 researchers asked for access to the SHARE data set. These are the countries where the SHARE sister studies, the U.S. Health and Retirement Study (HRS) and the English Longitudinal Study on Ageing (ELSA) are taking place. This shows that SHARE has reached acceptance in those countries that have pioneered ageing and retirement surveys.

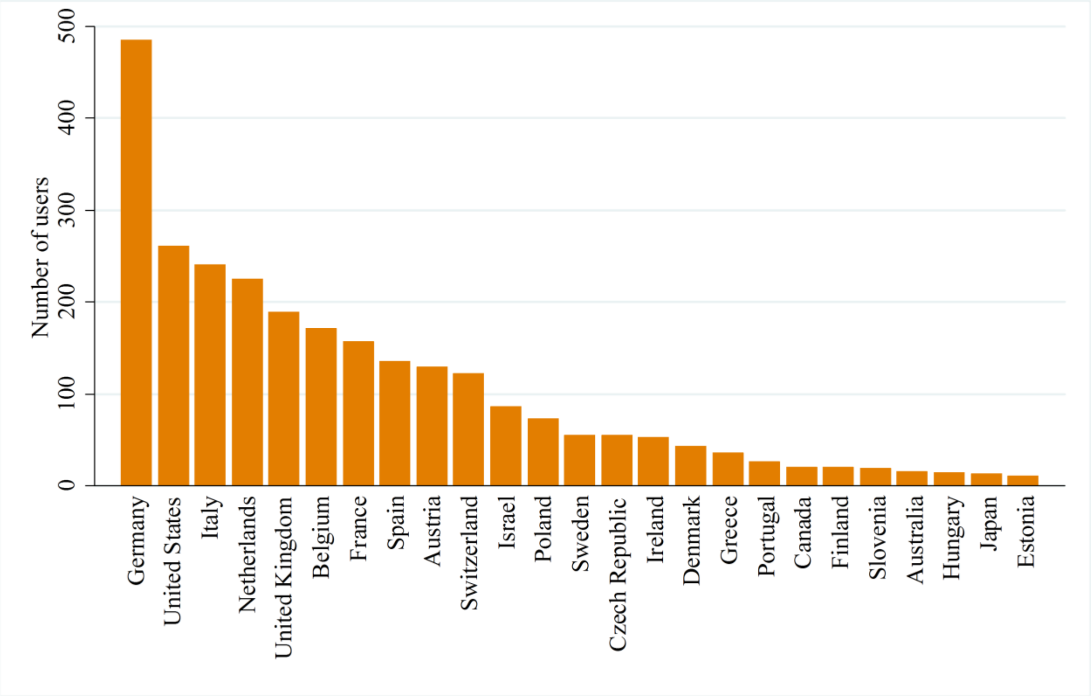


Figure 3: The number of SHARE data users per country

Note: Version August 2012, only countries with 10+ displayed

Sources:

The SHARE project website: www.share-project.org

Memo : Standards and procedures (COORD): codebooks, QbyQs and instrument

Survey of Health, Ageing and Retirement in Europe (SHARE) - Technical description

SHARE-Wiki

Report on SHARE data access, SHARE MEA

Börsch-Supan, Axel and Hendrik Jürges (2005): SHARE –Methodology, MEA, Mannheim.

Malter, Frederic and Börsch-Supan, Axel (2013): SHARE Wave 4: Innovations & Methodology, MEA, Munich.

12 Commercial Deposit Services[119]

12.1 Summary

In addition to officially recognized deposit services destined for academic research, many researchers frequently use various commercial upload services for uploading, storing and sharing their research data. This sub-report examines a selection of commercial deposit services; Dropbox, Figshare, Flickr and Youtube, that are used by researchers.

The services have been analysed with reference to four dimensions: organizational framework (including factors such as purpose and requirements, legal and regulatory framework and licensing), technological environment (including information on IT architecture, data storage, security and risk management and standards and formats), ingest and dissemination. Finally, a brief overall evaluation of each service is given.

Disclaimer: the main purpose of this report is to provide an overview of commercial services and the features that may be attractive to researchers as well as possible drawbacks. It does not however, provide a full overview of the services in question, and we refer readers to their web sites for a full description.

12.2 Dropbox

12.2.1 Organizational Framework

12.2.1.1 Purpose and Requirements

Dropbox is a commercial enterprise located in California that provides an online deposit service for photos, documents, and other files. When saving a file to Dropbox, the file automatically saves to all computers and devices you have connected to your Dropbox account. Dropbox also offers easy, intuitive options for file sharing. According to Dropbox' own figures, the service has 100 million users and 1 billion files are saved every day, making them a major player in data depositing. Although Dropbox is not specifically oriented towards research data, it is widely used for (informal) storing and sharing (via e-mail and social media) of research data. Dropbox has no specifically defined user group, except for the fact that users need to be more than 13 years of age.

12.2.1.2 Legal and Regulatory Framework

Use of Dropbox is regulated by the document "Terms of Service", which may be modified at any time. The terms of service and use of the software is governed by California law.

[119] This segment was provided by Anje Gjesdal, UiB; Bartholomäus Wloka, OEAW; and Dr. Astrid Recker, GESIS.

12.2.1.3 Business Model

Dropbox membership offers 2GB of storage for free. If more storage space is needed, users can upgrade to Dropbox Pro, available in 100 GB ($9.99/month, $99.00/year), 200 GB ($19.99/month, $199.00/year), and 500 GB ($49.99/month, $499.00/year). Dropbox also offers the service Dropbox for Teams for users aimed at educators, non-profits and businesses and needing more than 500 GB of storage.

12.2.1.4 Licensing

Users retain full ownership and Intellectual Property Rights to deposited documents, however there is no option for licensing uploaded data. Dropbox has no obligation to verify IPR rights of deposited material, but has channels for addressing misuse and infringement of copyright[120].

12.2.2 Technological Environment

12.2.2.1 IT Architecture

The user interface of Dropbox consists of three strands:

“The **Dropbox Desktop Application** is software that watches a folder on your desktop computer and syncs any changes to the web and to your other computers.

The **Dropbox Website** allows you to access your files on any computer from a web browser. You can also use the Dropbox website to share your files or folders with others.

The **Dropbox mobile website** and **Dropbox for mobile devices** allow you to connect to your Dropbox from your pocket, so you can take your files with you wherever you go.”

In order to simply access a file shared by others, it is sufficient to use a browser and it is not necessary to install the Desktop Application.

12.2.2.2 Data Storage, Security and Risk Management

Data uploaded to a Dropbox account synced to Dropbox’ online severs and is stored on Amazon's Simple Storage Service (S3) in multiple data centers located across the United States.Storage facilities are provided by Amazon via Amazon Simple Storage Service (Amazon S3), however Dropbox offers a simplified version, and it is unclear how many of the Amazon S3 functionalities are actually available for end users. Amazon specifies the following:

“Amazon S3 is intentionally built with a minimal feature set. Write, read, and delete objects containing from 1 byte to 5 terabytes of data each. The number of objects you can store is unlimited.

Each object is stored in a bucket and retrieved via a unique, developer-assigned key.

[120] https://www.dropbox.com/privacy#dmca

A bucket can be stored in one of several Regions. You can choose a Region to optimize for latency, minimize costs, or address regulatory requirements. Amazon S3 is currently available in the US Standard, US West (Oregon), US West (Northern California), EU (Ireland), Asia Pacific (Singapore), Asia Pacific (Tokyo), Asia Pacific (Sydney), South America (Sao Paulo), and GovCloud (US) Regions. The US Standard Region automatically routes requests to facilities in Northern Virginia or the Pacific Northwest using network maps.

Objects stored in a Region never leave the Region unless you transfer them out. For example, objects stored in the EU (Ireland) Region never leave the EU.

Authentication mechanisms are provided to ensure that data is kept secure from unauthorized access. Objects can be made private or public, and rights can be granted to specific users. Options for secure data upload/download and encryption of data at rest are provided for additional data protection.

Dropbox uses standards-based REST and SOAP interfaces designed to work with any Internet-development toolkit.

It is built to be flexible so that protocol or functional layers can easily be added. The default download protocol is HTTP. A BitTorrent™ protocol interface is provided to lower costs for high-scale distribution.

It also provides functionality to simplify manageability of data through its lifetime. Includes options for segregating data by buckets, monitoring and controlling spend, and automatically archiving data to even lower cost storage options. These options can be easily administered from the Amazon S3 Management Console.

Reliability is backed with the Amazon S3 Service Level Agreement."
(http://aws.amazon.com/s3/)

Dropbox stores file metadata in a separate facility.

Dropbox claims to be a secure service for data uploading. The security precautions are specified in the following way by Dropbox:

The security specifications are stated as follow:

"Dropbox uses modern encryption methods to both transfer and store your data.

Secure Sockets Layer (SSL) and AES-256 bit encryption.

Dropbox website and client software are constantly being hardened to enhance security and protect against attacks.

Two-step verification is available for an extra layer of security at login. You can choose to receive security codes by text message or via any Time-Based One-Time Password (TOTP) apps, such as those listed here.

Public files are only viewable by people who have a link to the file(s).

Dropbox uses Amazon's Simple Storage Service (S3) for storage, which has a robust security policy of its own."[121] Furthermore, security is password-based, and users are warned to safeguard their login details and to not disclose them to third parties.

Amazon S3 does offer several strategies for safeguarding data. It is, however, unclear, to what extent theses functionalities are available via Dropbox, in any case they are not clearly enough stated to make it a viable alternative for storing sensitive research data, or data provided by third parties, for which IPR needs to be strictly respected.

The security functionalities for Amazon S3 are specified as follows[122]:

"Amazon S3 supports several mechanisms that give you flexibility to control who can access your data as well as how, when, and where they can access it. Amazon S3 provides four different access control mechanisms: Identity and Access Management (IAM) policies, Access Control Lists (ACLs), bucket policies, and query string authentication. IAM enables organizations with multiple employees to create and manage multiple users under a single AWS account. With IAM policies, you can grant IAM users fine-grained control to your Amazon S3 bucket or objects. You can use ACLs to selectively add (grant) certain permissions on individual objects. Amazon S3 Bucket Policies can be used to add or deny permissions across some or all of the objects within a single bucket. With Query string authentication, you have the ability to share Amazon S3 objects through URLs that are valid for a predefined expiration time.

You can securely upload/download your data to Amazon S3 via the SSL encrypted endpoints using the HTTPS protocol. Amazon S3 also provides multiple options for encryption of data at rest. If you prefer to manage your own encryption keys, you can use a client encryption library like the Amazon S3 Encryption Client to encrypt your data before uploading to Amazon S3. Alternatively, you can use Amazon S3 Server Side Encryption (SSE) if you prefer to have Amazon S3 manage encryption keys for you. With Amazon S3 SSE, you can encrypt data on upload simply by adding an additional request header when writing the object. Decryption happens automatically when data is retrieved.

[121] https://www.dropbox.com/help/27/en

[122] http://aws.amazon.com/s3/

Amazon S3 also supports logging of requests made against your Amazon S3 resources. You can configure your Amazon S3 bucket to create access log records for the requests made against it. These server access logs capture all requests made against a bucket or the objects in it and can be used for auditing purposes."

12.2.2.3 Ingest

Dropbox is a repository for user-generated content, and has no restrictions on the nature of the uploaded content. There are no specifications of the standards and formats accepted. Dropbox does however request that user respect Intellectual Property Rights and do not upload spyware or other malicious software. In this sense, Dropbox imposes fewer restrictions on content than i.e. Figshare (for instance, Figshare refuses users to upload content that is 'obscene, offensive or profane'). It follows that Dropbox offers no criteria for evaluating the data deposited there.

Data is uploaded via the desktop application or online via a computer or mobile device. The only information provided for uploading data is on how to install and use the software. No information is given on requirements for standards and formats.

There seems to be no procedure for quality assurance or data checking. No information or documentation is required from the data producer beyond that required for signing up for a Dropbox account.

12.2.2.4 Dissemination

Data depositors are free to share their data as they see fit. There is no catalogue that gives an overview of the data stored in Dropbox, so it cannot be used to freely access data.

12.2.2.5 Evaluation

Overall, for limited quantities of data, Dropbox is free, very easy to use, it is highly accessible (single sign-on) and availability is very good. Tools and interfaces are very intuitive and easy to use. Using Dropbox requires very little previous technical knowledge and is accessible to people with basic technical skills.

However, some weaknesses could be noted. Importantly, there is no explicit preservation strategy available via the Dropbox web site. Since Dropbox is a private company, they are at liberty to terminate the services at their convenience. While Dropbox will not likely suspend it services in the near future, a long-term preservation status is not assured. Furthermore, Dropbox takes no responsibility for loss or corruption to uploaded data.

The storage solution is convenient and easy to use. Version copies are stored on Dropbox and can be retrieved by the data depositor, which is a very convenient feature. Dropbox stores file metadata, but it is difficult to find any information on whether or to what extent this can be retrieved by the data depositor.

Although Dropbox claims to offer a very secure storage service, several research institution in the US advise against using Dropbox for storing data of a sensitive or critical nature, and also raise concerns as to whether IPR is properly protected using cloud-based services (see e.g. the University of Delaware[123])

For uses other than informal storing and sharing of research data, Dropbox will likely be of limited use. Dropbox will not likely be very useful for dissemination of research, due to the lack of quality control in the ingest of data.

Problems with very slow uploads have been reported anecdotally, so on this basis the reliability of the service may to some extent be questioned.

12.3 Figshare

12.3.1 Organizational Framework

12.3.1.1 Purpose and Requirements

FigShare is a non-commercial content sharing platform, based in London and supported by Digital Science[124], [125] . The platform is aimed towards researchers from virtually every scientific field. Data sharing for promoting the research and raise the profile of the researcher is the main purpose of the platform, though it functions include safe data storage and collaborative research capabilities. The uploaded content is Creative Commons[126] and can be either kept private, or made publicly accessible to the FigShare community. In order to browse public data no registration is required, in order to share and store data registration of a free account is necessary. Any file format can be uploaded and categorized in several levels of categories, starting with the type: Media, Dataset, Figure, Poster, Paper. FigShare is used by researchers from Yale, Stanford, Imperial College London and UCLA.

12.3.1.2 Legal and Regulatory Framework

The data uploaded on FigShare is either classified as public or private. Private data is only accessible to the account holder who uploaded it and hence does not require any licensing. Data which is uploaded publicly remains public, i.e. cannot be set to private afterwards and falls under the Creative Commons license. A more detailed description of the license types is elaborated in the Licensing section below. FigShare's privacy policy[127] states in detail which information about the account holder is stored and used for service improvement,

[123] http://www.udel.edu/it/security/facultystaff/cloud2.html

[124] http://www.digital-science.com/pages/press-releases#figshare

[125] http://www.digital-science.com

[126] http://creativecommons.org/

[127] http://figshare.com/privacy

advertisement, business related issues and contact, as well as the account holder's capabilities of editing and customizing the way the information is handled. The terms and conditions of FigShare state the obvious restrictions on posting data that breaks copyright laws or can be in some other way regarded as inappropriate. The site itself and the service falls under the U.K. copyright laws, international conventions and other copyright laws. The user is bound to local, state, national and international laws and regulations. Further terms, which are fairly common for most filesharing services are listed on the terms and conditions website[128].

The location of the FigShare headquarters, including address and telephone and fax number is available at their website[129].

12.3.1.3 Licensing

Figures, media, posters, papers and filesets are licensed under CC-BY, which enables the FigShare community to distribute, remix and tweak, even commercially, as long as the license holder is credited for the original creation. Datasets are licensed under CC0, which is the Creative Commons recommendation for datasets, since “Databases may contain facts that, in and of themselves, are not protected by copyright law.” as it is stated in the FigShare license information.

12.3.2 Technological Environment

12.3.2.1 IT Architecture

The content is stored in a cloud based system. The API for the service is available, which offers the possibility to develop modules as well as full transparency of the entire serviceA collaborative workspace and a desktop upload integration are currently under development.

12.3.2.2 Data Storage, Security and Risk Management

FigShare is hosted through Amazon Web Services and claims to ensure the highest level of security and stability for the stored data, by storing several redundant, time-stamped copies. To guarantee persistence of data, FigShare uses the CLOCKSS$_7$ archive, which is a geographically and geopolitically distributed network of redundant archive nodes, located at 12 major research libraries around the world.

12.3.2.3 Standards and Formats

All file formats and any type of data is accepted for upload.

12.3.2.4 Ingest

FigShare offers 1GB of free storage for private data. Data which is made public can be uploaded

[128] http://figshare.com/terms
[129] http://figshare.com/contact

without a space limit. The upload process is very straight forward. The files can be dragged into the Web browser or selected by the native file browser. Information, such as title, type, authors, categories, tags and a description can be associated with the content.

12.3.2.5 Dissemination

The public data on FigShare can be browsed and accessed without an account. There are no restrictions and no groups can be defined to limit the dissemination of the data. However a collaboration module is currently under development, which might include such functionalities.

12.3.2.6 Evaluation

FigShare seems to be a practical choice to share and promote publications and research data. The metadata enables efficient and goal oriented search and the integration with social networks, such as twitter and facebook allows for an even broader spectrum of dissemination.

The disadvantage is the fact that data can only be shared entirely public, i.e. unrestricted access to anyone, or private, i.e. no access to anyone but the account holder. This eliminates the possibility of sharing sensitive data within closed research groups.

For long term preservation, however, FigShare would not be an obvious choice since the company "reserves the right, at its sole discretion, to modify or replace any of the Terms of Use, or change, suspend, or discontinue the Service (including without limitation, the availability of any feature, database, or content) at any time [...] Company may also impose limits on certain features and services or restrict your access to parts or all of the Service without notice or liability."

7 http://figshare.com/blog/Ensuring%20persistence%20on%20figshare/25

12.4Flickr

12.4.1 Organizational Framework

12.4.1.1 Purpose

Flickr states its main goals as 1) providing a photo sharing platform which allows users to share their pictures as openly or privately as they choose, and as hard- and software independently as possible; 2) “enabl[ing] new ways of organizing photos and video”, e.g. by allowing for (collaborative) tagging and commenting (http://www.flickr.com/about/). Thus, it strongly emphasizes “social” and Web 2.0 aspects rather than questions of (secure) file storage or preservation.

12.4.1.2 Legal and Regulatory Framework

Founded in 2004, Flickr was acquired by Yahoo! in 2005. To use Flickr, registration is required – either by creating a Yahoo! account, or by linking an existing Google or Facebook account to

Flickr. Signing up for Flickr means agreeing to comply with the following policies and regulations:

- Yahoo! Privacy policy (http://info.yahoo.com/privacy/us/yahoo/),
- Yahoo! Copyright and intellectual property policy (http://info.yahoo.com/copyright/us/details.html)
- Yahoo! Terms of Service (http://info.yahoo.com/legal/us/yahoo/utos/utos-173.html),
- Flickr Pro Additional Terms of Service http://www.flickr.com/atos/pro/ for subscription-based accounts
- Flickr Community Guidelines (http://www.flickr.com/help/guidelines/).

Flickr users give Yahoo! far-reaching rights to store and use their personal information (e.g. to provide targeted ads). This may include name, age, gender and geographical location, or information from one's computer and browser (including IP address, software and hardware attributes, requested pages). Yahoo! may share this information with business partners (e.g. for marketing purposes) and to "investigate, prevent, or take action regarding illegal activities, […] violations of Yahoo!'s terms of use, or as otherwise required by law" (Yahoo! Privacy Policy).

12.4.1.3 Licensing Options

For images uploaded, users can choose between the "All rights reserved" option (traditional copyright) or a Creative Commons license (http://creativecommons.org/). In addition, they can make their images available for licensing through the commercial provider Getty Images (see http://www.flickr.com/help/gettyimages/).

For any content that is made available on publicly accessible areas of Yahoo! Services (this includes Flickr), Yahoo! obtains non-exclusive rights "to use, distribute, reproduce, modify, adapt, publicly perform and publicly display" this content for as long as it remains in the publicly accessible areas (http://info.yahoo.com/legal/us/yahoo/utos/utos-173.html). This regulation is deemed "creator-friendly" by KeepYourCopyrights.org, a site developed by the Kernochan Center for Law, Media, and the Arts and the Program on Law & Technology at Columbia Law School (http://www.keepyourcopyrights.org/contracts/clauses/example/33), as creators keep their copyright.

12.4.1.4 Business Model

Flickr offers free services as well as subscription-based "Pro" accounts. Free Flickr accounts have limitations with regard to uploads, number of pictures displayed in the photostream, and quality of pictures (see table). In addition, the free version displays advertisements to users. Flickr Pro accounts are ad-free, offer additional functionality (e.g. statistics), and do not have the same restrictions with regard to storage, viewing, and downloading images as the free accounts.

Free accounts	Pro accounts
• 300 MB monthly photo upload (30MB per photo) • 2 video uploads each month (90 seconds max, 150MB per video) • photostream views limited to the 200 most recent images • post photos in up to 10 group pools • only smaller (resized) images accessible (originals are saved in case user upgrades later)	• Unlimited photo and video uploads (50MB per photo, 90 seconds max, 500MB per video) • ability to show HD Video • unlimited storage and bandwidth • archiving of high-resolution original images • ability to replace a photo • post photos/videos in up to 60 group pools • count and referrer statistics • limitation of maximum image size available to others • Ad-free browsing and sharing
Adapted from http://www.flickr.com/help/limits/#65	

Subscription fees range from $6.95 for 3 months to $44.95 for two years. Flickr generates further revenues through advertisements (http://advertising.yahoo.com/article/flickr.html) as well as through the Getty Images licensing scheme. See Hawk 2011; Kirkpatrick 2011 for a discussion of the main sources of Flickr's revenues.

12.4.2 Technological Environment

12.4.2.1 IT Architecture

Some (possibly outdated) information on Flickr's web architecture can be gathered from a 2007 presentation by Cal Henderson, former chief software architect at Flickr (see Henderson 2007). This does not include information about how data is stored and backed up.

12.4.2.2 Data Storage, Security and Risk Management

While Yahoo!/Flickr provides information about how it keeps users' personal data safe (i.e. how they are protected against unauthorized access; see the Yahoo! Privacy policy in particular), this information does not extend to issues such as data storage or disaster preparedness. It seems safe to assume that Yahoo! has implemented sufficient measures to back-up data and protect itself (and its users) against data loss. However, using this service as (the only) back-up for one's pictures or videos is not recommended as users ultimately have no control over what happens to their data. According to Yahoo!'s Terms of Service, data may be deleted along with a user's account, e.g. if infringements of copyright or IPR occur (note that this may also happen accidentally as in the case described by Zhang 2011).

12.4.2.3 Standards and Formats

Flickr restricts the file formats that can be uploaded. For images it supports JPEG, GIF (non-animated), and PNG format. In addition, it is possible to upload files in other image formats (e.g. TIFF) – these will automatically be converted to JPEG, however (see (http://www.flickr.com/help/with/uploading/). Accepted video formats are AVI, WMV, MOV, MPEG, 3gp, M2TS, OGG, and OGV (see http://www.flickr.com/help/video/).

During upload, a form of format identification/characterization appears to take place, as Flickr will not accept files whose extension has simply been changed (e.g., changing the file extension DOC to GIF will still not enable you to upload a word file). Also, Flickr will compress and resize photos to a set of standard sizes ranging from 75x75 pixels square to 2048 pixels.

12.4.2.4 Ingest

Flickr allows users to upload files by means of the following tools:

- Flickr web uploader (select files from directory or use drag and drop)
- Desktop uploader (work offline; select files from directory or use drag and drop)
- Upload by email, mobile app, or web page for mobile
- Third-party apps.

It thus offers a considerable range of channels through which users can upload content to Flickr.

During upload, Flickr will automatically extract and store Exif metadata with technical information about the image and the camera that it was taken with. The following metadata can be added to describe the content and context of the image or video:

- image name (free text)
- description (free text)
- tags (free text)
- persons in picture/video (select from contacts or enter email address)
- notes (free text)
- geo tags (place images on a map)

To organize images/videos further, they can be arranged into sets according to freely assignable categories (e.g. “mountains”, “black-and-white”, etc.) and/or added to groups of which the user is a member.

It should be noted that Yahoo! – and Flickr accordingly) limit the kind of content that can be uploaded to its services. Thus, for example, content that is deemed “vulgar” may not be uploaded to Flickr (Yahoo! Terms of Service). While the reasons for such restrictions are clear, it might mean that images/videos which are the subject of research cannot be uploaded (consider, for example, [images of] paintings relevant to art history research).

12.4.2.5 Dissemination

Users have control over whom they give permission to access images/videos they upload. Thus, in addition to sharing their images/videos with the public, they may make pictures completely private, visible only to persons marked “friends” or “family”, or to someone invited through a “guest pass”.

Public images are searchable and browsable through the Flickr webpage (viewing them does not require registration or log-in), as well as the Getty Images Flickr collection. Images and videos can further be shared via email and by pushing them to social media sites such as Facebook, Twitter, or blogs. As images/videos receive no persistent identifiers, however, their long-term availability and accessibility is uncertain and depends on a user’s decision to maintain or delete an image.

12.4.2.6 Evaluation

The strength of Flickr is that it makes it comparably easy to share images and to work with them collaboratively, e.g. by discussing them, commenting on them or tagging them. These functions can be useful in teaching environments or collaborative research centering on visual information (although one has to keep in mind the drawback that Flickr/Yahoo! being a commercial provider, one’s personal data will be used for business purposes). In this sense, Flickr is truly an invention of the “social web,” making it easy to share content and get others involved. But it is not a place to safely back-up one’s data for the reasons outlined above. Losing one’s data will also mean losing the added-value generated through discussions and commenting. Thus, while a number of back-up apps are available which allow users to download their images and some metadata from Flickr, these will typically not include comments, discussions, or notes.

Another drawback of Flickr is that with the free accounts, users do not have access to the original files they uploaded but only to the compressed and resized copies. Again, most often this will certainly be sufficient where sharing images/videos with others is the main purpose. Depending on how important image quality is in a given context and to a given user, however, it may not be advisable to back them up using a service which converts and compresses them without the user being able to control this process.

With respect to long term preservation, it must be noted that the user accounts are personal and non-transferable, that rights to materials will be canceled and materials can be permanently deleted after the user's death.

12.5 Youtube

12.5.1 Organizational Framework

12.5.1.1 Purpose and Requirements

YouTube is a commercial video content sharing platform, created by former Paypal employees in 2005. It was sold to Google in 2006 and has been running as a subsidiary of Google since then [1]. Youtube is not inherently directed towards a specific kind of content and the genre of uploaded content varies from vacation videos of private users to university lecture recordings. YouTube allows the upload of any video content by any registered user. Naturally, since anybody is able to upload any content, there are issues with potentially offensive content, which however is locked, or at least protected, i.e. only viewable to registered users who are older than 18 years of age.

12.5.1.2 Legal and Regulatory Framework

As YouTube was acquired by Google, users with a valid Google account can directly access all YouTube functionalities by entering their Google account information. The privacy regulation of YouTube is directly derived from Google's statement, viewable at: http://www.google.com/intl/en/policies/privacy/

The Copyrights, which are enforced by YouTube are described here: http://www.youtube.com/t/copyright_center

This information covers the copyrights from the view of content owners, as well as content viewers. General terms of service can be found at: http://www.youtube.com/t/terms

YouTube’s general statement is that respecting copyrights, privacy and other legal issues the responsibility of the community.

12.5.1.3 Licensing

YouTube offers two choices for uploaded content. The “Standard YouTube License” and the “Creative Commons Attribution”.

12.5.2 Technological Environment

12.5.2.1 IT Architecture

The video content on YouTube is accessible by any computer with a browser and the Adobe Flash Player plug-in. The Adobe Flash player plugin is one of the most common installed pieces of software. Adobe Flash videos account for almost 75% of online video material [2]. Starting in 2010, YouTube is experimenting with presenting the content through HTML5 technology, which would increase the coverage even further, since HTML5 can be interpreted by any modern browser. This functionality is already implemented as an option, though still in a beta phase. The upload process is also entirely browser-driven and no other software is required.

12.5.3 Data Storage, Security and Risk Management

12.5.3.1 Standards and Formats

The formats used by YouTube is restricted to most common video file standards. These include the MP4, WMV, FLV, 3GP and WebM, amongst others. The quality of the video depends on the encoding of the video stream (see table at http://en.wikipedia.org/wiki/YouTube)

12.5.3.2 Ingest

The content (video files, such as mp4's, wmv's, etc.) can be uploaded directly via an interface, on which the data provider can set various setting, described in this document (see Licensing and Dissemination). Videos up to 15 minutes in length can be uploaded. Longer conctent, up to 12 hours can be uploaded upon verification, for example through a mobile phone.

12.5.3.3 Dissemination

In order to share data through YouTube a registration is needed. During the registration process and any time after that, an existing Google account can be added, giving several option of combining activities related to Google services.

During the upload process of a video file, data creators have the choice of availability of their contribution. The following choices are included:

1. Public: The content is viewable by anyone and is listed in the YouTube search engine.

2. Unlisted: The content is viewable by anyone, however it is not listed in the search engine, hence it is not easily found. Regular users will only be able to find the content if they receive an URL pointing to the content.

3. Private: The video is only viewable by other registered YouTube users who are explicitly defined by the data provider.

If the content is set to public or unlisted YouTube does not restrict users in terms of viewing uploaded material, except for potentially offensive content (login required).

Another property set during the upload process is the category. The list of categories includes 15 items and ranges from “Entertainment” to “Science and Technology.”

12.5.3.4 Evaluation

The big advantage of YouTube is its ease of use. In most cases, no software needs to be installed on the computer, as most operating systems come with a browser and Adobe Flash installed. The upload process, as well as the search and view process are very intuitive. Further, data providers can restrict content to selected users. A useful feature is the view recommendations, while a video is selected, YouTube (in combination with Google) tries to find related content and makes several suggestions, which could be potentially useful for research purposes. Recent *social Web* integration with "like" buttons enable the extension of collaborative work. As a data consumer it is possible to add videos to a personal play list, according to custom categories, allowing for a structured collection of personal content.

The drawback of YouTube is that there is no guarantee that the data stays available at any time in the future. Though it is very unlikely that any other content, other than material deemed illegal, will be taken out of the YouTube database, there exists no written assurance from Google as to the long term preservation of the content. "YouTube reserves the right to discontinue any aspect of the Service at any time."

Sources

[1] Hopkins, Jim (October 11, 2006). "Surprise! There's a third YouTube co-founder". *USA Today*. Retrieved November 29, 2008.

[2] Fildes, Jonathan (October 5, 2009). "Flash moves on to smart phones". BBC. Retrieved November 30, 2009.

http://www.youtube.com/t/about_youtube

12.6 Conclusion

The investigated commercial services are very accessible and user friendly. Some are more aimed at the research community while others are aimed at a large general audience. In general, users retain the copyright of their materials, but some services also reserve extensive rights. These commercial services are backed by storage systems which generally have a high degree of reliability. However, none of the services offers long term guarantees that the data will be preserved and will remain accessible. Thus, these services are no replacements for libraries and data archives, but must be seen in the first place as short term dissemination channels.

Sources

Hawk, Thomas. 2011. "How Much Revenue Does Flickr Make From Paid Pro Accounts. My Guess? $50 Million Per Year." *Thomas Hawk's Digital Connection*. http://thomashawk.com/2011/01/how-much-revenue-does-flickr-make-from-paid-pro-accounts-my-guess-50-million-per-year.html.

Henderson, Cal. 2007. “Flickr Architecture (Slide).” *Slideshare.net*. http://de.slideshare.net/techdude/scalable-web-architectures-common-patterns-and-approaches/138.

Kirkpatrick, Marshall. 2011. “How Much Is Flickr Worth to Yahoo? Not Very Much (Updated).” *Readwrite.com*. http://readwrite.com/2011/01/13/how_much_is_flickr_worth_to_yahoo_not_very_much.

Zhang, Michael. 2011. “Flickr Accidentally Deletes the Wrong Account, Vaporizing 4,000 Photos.” *PetaPixel.com*. http://www.petapixel.com/2011/02/02/flickr-accidentally-deletes-the-wrong-account-vaporizing-4000-photos/.

Appendix: Data Archive Description Sheets (DADS)

UK Data Archive

Functionalities	Short description	Reference
Administrative context		
Funding	Largely funded by the ESRC, the JISC and the University of Essex.	http://data-archive.ac.uk/about/archive
Depositor Agreements	License Agreement: Legal agreement for depositing standard End User Licensee data in the collection.	http://www.esds.ac.uk/aandp/create/licenceForm.pdf
Usage Agreements , Code of Conduct to be signed	Same as above, includes depositor signature	http://www.esds.ac.uk/aandp/create/licenceForm.pdf
Policies in place	Depositor Signature, Copyright, Access Conditions and Acceptance of Terms and Conditions are included in the License Agreement	http://www.esds.ac.uk/aandp/create/licenceForm.pdf
	Collections Development Policy	http://data-archive.ac.uk/media/54773/ukda067-rms-collectionsdevelopmentpolicy.pdf
	Preservation Policy	http://data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf
Rights on data claimed by the archive	Defined in the License Agreement.	http://www.esds.ac.uk/aandp/create/licenceForm.pdf
	..and the Preservation Policy	http://data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf
Data Curation strategy	Collections Development Policy	http://data-archive.ac.uk/media/54773/ukda067-rms-collectionsdevelopmentpolicy.pdf
	Preservation Policy	http://data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf
Pre-Ingest		
Primary community in focus for deposits	Social science data users within higher education (HE) and further education (FE) in the UK, though best efforts are made for all users	http://data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf
Secondary communities accepted for deposits	See above	
Ingest		
Formats accepted and curated	File formats table	http://data-archive.ac.uk/create-manage/format/formats-table
	File formats also mentioned in 'Managing and Sharing Data - Best Practice for Researchers'	http://data-archive.ac.uk/media/2894/managingsharing.pdf
Formats accepted and not curated	See above	
Metadata formats accepted	See above	
User-based ingest	Data deposit	http://www.esds.ac.uk/aandp/create/ukdadeposit.asp
	Data collection deposit forms	http://www.esds.ac.uk/aandp/create/depform.asp
Archival storage and preservation		
Size of current archive in TB	-	
Size of current archive in other means	-	
Maximal deposit size in TB	-	
Long term guarantees / standards of trust	Data Seal of Approval	http://data-archive.ac.uk/curate/trusted-digital-repositories/standards-of-trust?index=1
Checks on quality / quality control	Quality Control	http://data-archive.ac.uk/curate/archive-quality
	Data Processing Standards	http://data-archive.ac.uk/media/54782/ukda079-ds-dataprocessingstandards.pdf
	Catalogue and Study Quality Control Procedures	http://data-archive.ac.uk/media/54779/ukda084-ds-cataloguequalitycontrolprocedures.pdf
	Data Processing Quick Reference	http://data-archive.ac.uk/media/54764/ukda080-ds-processingquickreference.pdf
	Quantitative Data Processing Procedures	http://data-archive.ac.uk/media/54770/ukda081-ds-quantitativedataprocessingprocedures.pdf
	Documentation Processing Procedures	http://data-archive.ac.uk/media/54785/ukda078-ds-documentationprocessingprocedures.pdf
Dissemination		
Costs / Conditions for Access	License Agreement	http://www.esds.ac.uk/aandp/create/licenceForm.pdf
Tools / Interfaces used for Access	ESDS Nesstar Catalogue	http://nesstar.esds.ac.uk/webview/index.jsp
	ESDS Qualidata Online	http://www.esds.ac.uk/qualidata/online/about/introduction.asp
	ESDS Online Data Search	http://www.esds.ac.uk/Lucene/Search.aspx

GESIS Data Archive for the Social Sciences

Functionalities	Short description	Reference
Administrative context		
Funding	GESIS is sponsored jointly by the federal government and the federal states.	Article 91b of the German Federal Constitution
Depositor Agreements	The depositor agreement (“Archivierungsvertrag”) is sent to depositors during the pre-ingest process. A sample agreement will be available on the webpage soon.	GESIS Datenarchiv für Sozialwissenschaften - Archivierungsvertrag
Usage Agreements, Code of Conduct to be signed	The Usage Regulations and information on fees are published on the Internet. These regulations outline access conditions and categories, request and provision of material, measure to take after completion of project, obligation to quote, specimen copy, and fees (see Costs below) among other things.	http://www.gesis.org/en/services/data-analysis/data-archive-service/usage-regulations/
Policies in place	License policy: see usage regulations; preservation policy; various policies governing IT security and the processing and archiving of data; rules for scholarly conduct & integrity of research Digital Preservation Policy: Grundsätze der digitalen Langzeitarchivierung am Datenarchiv für Sozialwissenschaften (to be published)	http://www.gesis.org/en/services/data-analysis/data-archive-service/usage-regulations/, http://www.gesis.org/fileadmin/upload/institut/leitbild/Gute_Praxis_GESIS-Regeln.pdf
Rights on data claimed by the archive	Defined in depositor agreement. All rights associated with the data stay with the data producer. The archive only requests non-exclusive rights necessary for active data preservation and dissemination.	
Data Curation strategy	Data quality checks, documentation and enhancement undertaken according to internal guidelines; versioning and assignment of DOIs; preservation strategy: migration	Internal documents; Preservation Policy (to be published)
Pre-Ingest		
Primary community in focus for deposits	Empirical social research, in particular in sociology, social sciences, political sciences	
Secondary communities accepted for deposits	Neighboring disciplines if no other suitable archive exists	
Ingest		
Formats accepted and curated	No formats are per se excluded; the acceptance and curation of formats that are not standard social sciences formats is discussed with depositors during the pre-ingest phase.	A list of recommended standards is currently under preparation.
Formats accepted and not curated	See “Formats accepted and curated”.	
Metadata formats accepted	Preferred standard: DDI. However, no formats are per se excluded. If structured metadata in a different format exist, its submission is discussed with depositors during the pre-ingest phase.	
User-based ingest	In preparation (Datorium). Currently, data are mainly submitted by email, download, or on other media (e.g. CD, DVD).	
Archival storage and preservation		
Size of current archive in TB	2	
Size of current archive in other means	ca. 5,100 studies; ca. 600,000 files (as of Dec 2012)	
Maximal deposit size in TB	No general limitation. Discussion with depositors during pre-ingest phase.	
Long-term guarantees / standards of trust	Unlimited retention period (deposit agreements are of unlimited duration); Data Seal of Approval (currently under review)	
Checks on quality / quality control	Intellectual quality checks performed during ingest according to internal quality guidelines.	Internal documents
Dissemination		
Costs / Conditions for Access	See Usage Regulations. Access to data is usually free but fees exist for data requested on CD or DVD, or for customized data, for example.	http://www.gesis.org/en/services/data-analysis/data-archive-service/charges/
Tools / Interfaces used for Access	Data Catalog DBK, ZACAT Online Study Catalog, Codebook Explorer (offline tool), Qbase, Histat (historical studies)	http://www.gesis.org/en/services/research/data-catalogue/ http://www.gesis.org/en/services/research/zacat-online-study-catalogue/ http://www.gesis.org/unser-angebot/recherchieren/codebookexplorer/ http://www.gesis.org/en/services/research/german-question-text/ http://www.gesis.org/en/services/research/english-question-text/ http://www.gesis.org/histat/en/index

SHARE

Functionalities	Short description	Reference
Administrative context		
Funding	National and international funding	http://www.share-project.org/contact-organisation/funding.html
Depositor Agreements	Complies to Criteria of the German council for Social and Economic Data	http://www.ratswd.de/download/publikationen_rat/RatSWD_FDZCriteria.pdf
Usage Agreements , Code of Conduct to be signed		http://www.share-project.org/data-access-documentation/research-data-center-data-access.html
Policies in place	See the usage agreements	http://www.share-project.org/data-access-documentation/research-data-center-data-access.html
Rights on data claimed by the archive	See the usage agreements	http://www.share-project.org/data-access-documentation/research-data-center-data-access.html
Data Curation strategy	Via GESIS Data Archive for the Social Sciences; doi-registration via da|ra ongoing. SHARE also holds a local archive. Data curation is accomplished here by keeping backup copies of all the data at the Dutch National Computing Centre.	
Pre-Ingest		
Primary community in focus for deposits	The primary community (researchers in the field economics, health and social sciences) for SHARE is described in the SHARE Brochure	http://www.share-project.org/fileadmin/SHARE_Brochure/share_broschuere_web_final.pdf
Secondary communities accepted for deposits	Not applicable	
Ingest		
Formats accepted and curated	SPSS/STATA/Logfiles/Blaise Databases	
Formats accepted and not curated	-	
Metadata formats accepted	HTML	
User-based ingest	Questions to user support and suggestions by user monitoring board are taken into account	
Archival storage and preservation		
Size of current archive in TB	1 TB	
Size of current archive in other means (collections, files, etc.)	In 2013 4 waves of SHARE data are released, round about 120 modules for 20 countries	
Maximal deposit size in TB	Not applicable since only SHARE data is archived	
Long term guarantees / standards of trust	Via GESIS Data Archive for the Social Sciences, da|ra (doi registration)	
Checks on quality / quality control	Various quality checks are done by CentERdata, MEA , university of Venice, all country teams and survey agencies see	http://www.share-project.org/contact-organisation/project-coordination.html
Dissemination		
Costs / Conditions for Access	Free for scientific usage	http://www.share-project.org/data-access-documentation/research-data-center-data-access.html
Tools / Interfaces used for Access		http://www.share-project.org/data-access-documentation/research-data-center-data-access.html

NSD

Functionalities	Short description	Reference
Administrative context		
Funding	Directors' report for 2011: "The main allocations come from the Research Council of Norway, the ministries, the university and university college sector and the EU.	http://www.nsd.uib.no/nsd/doc/nsd_annualreport2011.pdf
Depositor Agreements	Web-based "Deposit Form" (the form is generated for each deposit)	-
Usage Agreements , Code of Conduct to be signed	Within a week after the application is submitted, the user will receive a reply from the NSD with the necessary information, confidentiality and any supervisor's declaration and report forms.	http://www.nsd.uib.no/nsd/english/order.html
Policies in place	Copyright, Access Conditions and Acceptance of Terms and Conditions are included in the License Agreement	-
Rights on data claimed by the archive	Defined in "Pledge of secrecy", in "License agreement" and in "Supervisor's Declaration"	Not online
		-
Data Curation strategy	"Data in - filing and storage - data out. Routines for handling data: a handbook" "Guide for study documentation in Nesstar"	Internal documents
Pre-Ingest		
Primary community in focus for deposits	Preservation policy: "The work is a continuation of the responsibility that NSD has had for almost 40 years to archive data from projects that receive funding from the Research Council of Norway. NSD will archive data from projects conducted by researchers and students at universities, specialized and state university colleges, institutes and other research institutions."	http://www.nsd.uib.no/nsd/doc/nsd_annualreport2010.pdf http://www.nsd.uib.no/nsd/english/index.html
Secondary communities accepted for deposits	See above	
Ingest		
Formats accepted and curated	All dynamic matrix/spreadsheet formats accepted	-
Formats accepted and not curated	All other formats	
Metadata formats accepted	DDI, Dublin Core	
User-based ingest	Data deposit form	http://www.nsd.uib.no/nsddata/arkivere_data.html
Archival storage and preservation		
Size of current archive in TB	Approx. 1 TB	
Size of current archive in other means (collections, files, etc.)	Including administrative data: approx. 3 TB	
Maximal deposit size in TB	Not specified, evaluated for each new depositor	
Long term guarantees / standards of trust	Data Seal of Approval (ongoing process)	
Checks on quality / quality control	Manual checks and quality controls. "Data in - filing and storage - data out. Routines for handling data: a handbook"	Internal document
Dissemination		
Costs / Conditions for Access	Free. Institutional. License Agreement	http://www.nsd.uib.no/nsd/english/order.html
Tools / Interfaces used for Access	Nesstar/SPSS	http://www.nsd.uib.no/solr/nsu

DANS

Functionalities	Short description	Reference
Administrative context		
Funding	KNAW and NWO	
Depositor Agreements	Licensee agreement generated and accepted upon submission of a dataset	http://www.dans.knaw.nl/en/content/dans-licence-agreement-deposited-data http://www.dans.knaw.nl/sites/default/files/file/archief/Licence_agreement_DANS_UK.doc
Usage Agreements , Code of Conduct to be signed	The DANS General Conditions of Use are in place	http://www.dans.knaw.nl/sites/default/files/file/archief/DANS_General_Conditions.pdf
Policies in place	Deposited datasets are processed within one week according to a established protocol. Hereafter, these datasets will be published.	http://www.dans.knaw.nl/sites/default/files/Provenance_document_DEF.pdf
Rights on data claimed by the archive	The license agreement is 'non-exclusive'; the owner of the data is at liberty to deposit and/or make available these data in other places as well. Copyright is not waived when data are deposited: it continues to rest with the researcher. The license entitles DANS to include the dataset in the archive and to make it available under the conditions stipulated by the project leader when it is deposited.	http://www.dans.knaw.nl/en/content/dans-licence-agreement-deposited-data
Data Curation strategy	DANS archivists work according to a standard protocol in order to provide long-term preservation, findability and accessibility of the data, as well as checking for and anonymising privacy-sensitive data.	
Pre-Ingest		
Primary community in focus for deposits	Researchers of all kinds of scientific disciplines	
Secondary communities accepted for deposits	N/A	
Ingest		
Formats accepted and curated	PDF/A; Unicode TXT; Comma Separated Values; ANSI; SQL; SPSS Portable; SAS Transport; STATA; JPEG; TIFF; Scalable Vector Graphs; MPEG-2; MPEG-4 H264; Lossless AVI; QuickTime; WAVE; AutoCAD DXF version R12; Mapinfo Interchange Format.	
Formats accepted and not curated	All other formats (in principle)	
Metadata formats accepted	Dublin Core; Qualified Dublin Core; file-descriptive metadata usinf DDI attributes with optional additions from FGDC or non-standardised metadata.	
User-based ingest	*Self-archiving,* deposits in the EASY archiving system	
Archival storage and preservation		
Size of current archive in TB	4.3	
Size of current archive in other means	25,000 datasets and 2,000,000 data files	
Maximal deposit size in TB	Not specified, but the costs of data archiving need to be covered; i.e. included in the budget of the data collection project.	
Long term guarantees / standards of trust	The DANS Preferred Formats have, according to our best estimation, a high chance of remaining usable in the far future.	
Checks on quality / quality control	Upload of datasets into the archive is automated by EASY. The process of data cleaning and metadata enrichment is mostly done manually, partially supported by existing tools and scripts.	

	All sorts of security aspects are continuously being monitored by the DANS security officer as well as the security officer of the KNAW ICT service I&A.	http://www.dans.knaw.nl/sites/default/files/Provenance_document_DEF.pdf
	Dissemination	
Costs / Conditions for Access	User definable access rights; Open Access (direct access for registered users); Restricted Group (access to users registerd to a specific user group; Restricted-Permission Request (access after granted permission by the data depositor); Other Access (data available elsewhere). Costs for Deposits or for Access are in implementation. Based on the Activity Based Cost Model, in combination with the Balanced Score Card Method.	
Tools / Interfaces used for Access	EASY	https://easy.dans.knaw.nl/

The Language Archive

Functionalities	Short description	Reference
	Administrative context	
Funding	The Max Planck Society (MPG); The Berlin-Brandenburg Academy of Sciences (BBAW); The Royal Netherlands Academy of Sciences (KNAW); External projects funded by the EC, BMBF, DFG and others.	
Depositor Agreements	A template depositor agreement is used which can be adapted to the specific needs for given deposits.	
Usage Agreements , Code of Conduct to be signed	Different usage agreements and licenses exist for different deposits, e.g. DBD conditions of use, DOBES Code of Conduct, GNU GPL and Creative Commons Attribution 3.0.	
Policies in place	TLA will undertake measures to maintain and protect the data for scientific purposes to the best of its ability while protecting the rights of consultants and communities. In principle, all data should be available to the entire scientific community, except for ethically sensitive data or data collected for a PhD program research.	
Rights on data claimed by the archive	The archive claims the right to archive the data and to make as many duplicates as it deems necessary. The original owner of the data remains the owner and determines the access conditions.	
Data Curation strategy	Immediate conversion policy where possible to avoid future curation costs. Local storage system, which stores two copies of all resources immediately at upload. Dynamic copies are created at two large computer centers in Germany, which also have agreements with another computer center about long-term archiving. Thus all archived objects are available in 6 copies spread geographically.	
	Pre-Ingest	
Primary community in focus for deposits	Language researchers affiliated with the main funders of the archive as well as partners in projects with TLA involvement such as DOBES and CLARIN	
Secondary communities accepted for deposits	Any interesting and relevant research data from the humanities, e.g. linguistics, psycholinguistics, psychology, gesture studies.	
	Ingest	
Formats accepted and curated	Audio: wav, m4a Video: mpeg1, mpeg2, mpeg4 Image: jpg, png, tiff, svg Document: pdf, html Transcription/Annotation: EAF, Toolbox, TextGrid, CHAT, plain text	Complete list: LAMUS user manual, Appendix A: http://www.mpi.nl/corpus/manuals/manual-lamus.pdf
Formats accepted and not curated	-	
Metadata formats accepted	IMDI, CMDI	
User-based ingest	Deposit is performed by the user with LAMUS tool.	
	Archival storage and preservation	
Size of current archive in TB	80	
Size of current archive in other means (collections, files, etc.)	11.000 hours of audio recordings; 10.000 hours video recordings; 168.000 metadata described sessions; 125.000 annotation files (30k eaf, 50k cha); 5 million annotated segments; 85 lexica. (Approximately).	
Maximal deposit size in TB	Not defined, evaluated for each new depositor	
Long term guarantees / standards of trust	50 years institutional guarantee for all resources by the president of the Max Planck Society.	
Checks on quality / quality control	The Data Seal of Approval assessment.	
	Dissemination	
Costs / Conditions for Access	No costs for academic and non-commercial use.	
	Four possible access levels for each resource: 1 – Open for everyone	

	2 – Creating an account and login required 3 – Asking the owner for access permission required 4 – Closed, only available for the owner	
Tools / Interfaces used for Access	The LAT software package is used for the complete life-cycle of the data. It has been fully developed at TLA.	http://tla.mpi.nl/tools/tla-tools/

ESS

Functionalities	Short description	Reference
Administrative context		
Funding	Central Coordination grants provided by the European Commission under its various framework programmes. Future funding will be secured via the ESS-ERIC.	Round 6 Project Specification for participating countries (section 1.3) Round 7 Project Specification for participating countries (section 1.1)
Depositor Agreements	The ESS Archive has a license from the Norwegian Data Inspectorate, which defines NSD as the Data controller of ESS data at the *International* level. The license also defines the organisations responsible for the collection of data in each country as the data controllers at the *national* level. An additional agreement between NSD and the national organisations regulates the transfer, storage and dissemination of data that could possibly indirectly identify individuals (e.g. non-anonymous raw data and sample design data). All EU and EEA countries are covered by the license.	Round 6 Project Specification for participating countries (section 8) Kolsrud, K., Kalgraff Skjåk, K. and Henrichsen, B. (2007) Free and immediate access to data. In: Measuring Attitudes Cross-nationally: Lessons from the European Social Survey, R. Jowell, C. Roberts, R. Fitzgerald and G. Eva. London: Sage Publications, pp.139-156.
Usage Agreements , Code of Conduct to be signed	Data that could indirectly identify individuals is not publicly released, but stored securely by the ESS Archive in accordance with NSD's license (see section 2). Registered ESS data users in countries that have implemented EU Directive 95/46 "On the Protection of individuals with regard to the Processing of Personal Data" or that have bilateral agreements with the EU, can access indirectly identifiable (non-anonymous) data according to a special license from NSD. Researchers from elsewhere can get access to the data at NSD. All other ESS data and documentation is publicly available online for free without restriction, for not-for-profit purposes.	ESS Archive: http://ess.nsd.uib.no/
Policies in place	See section Usage Agreements, above	
Rights on data claimed by the archive	NSD acts as the Data Controller of ESS data at the International level (see section 2 above)	
Data Curation strategy	Deposit of data by National teams according to same specifications (outlined in ESS Data Protocol); deposit of documentation by National teams and completion of National Technical Summary form (NTS) which documents how the ESS was implemented in a particular country. Data checking, data edit, edit control and data approval activities carried out by archive team according to NSD procedures and internal guidelines.	ESS Round 5 Data Protocol ESS Round 5 Documentation Report Kolsrud, K., Midtsæter, H., Orten, H., Kalgraff Skjåk, K., and Øvrebø, O. (2010) Processing, Archiving and Dissemination of ESS data. The Work of the Norwegian Social Science Data Services ASK Research Methods 19, 1 51–92.
Pre-Ingest		
Primary community in focus for deposits	ESS National Coordination teams and Survey agencies are responsible for depositing ESS data and documentation to the ESS Archive.	
	The academic community (including undergraduate and postgraduate students, faculty and researchers) from across the world at the primary users.	
Secondary communities accepted for deposits	Not applicable	
Ingest		
Formats accepted and curated	SPSS/PASW, Stata, SAS, NSDstat, Statistica, DIF, Dbase, Text delimited, WORD, PDF, ASCII text (if text is the only possible output from the CAPI system)	ESS Round 5 Data Protocol European Social Survey, Data Team. (2012) ESS Archive

		Architecture Report to DASISH WP2, Task 2.1 "State of the Architectures Report" Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services.
Formats accepted and not curated	Not applicable	
Metadata formats accepted	ESS metadata are stored in a 'question database' and a 'documentation database'. Both use an MS SQL server and Java Programming. Java allows various parts of the databases to be outputted to PDF reports for end users & XML files and rtf files for further editing.	Kolsrud, K., Midtsæter, H., Orten, H., Kalgraff Skjåk, K., and Øvrebø, O. (2010) Processing, Archiving and Dissemination of ESS data. The Work of the Norwegian Social Science Data Services ASK Research Methods 19, 1 51–92.
User-based ingest	See Section Checks on quality, below.	
Archival storage and preservation		
Size of current archive in TB	Not known	
Size of current archive in other means (collections, files, etc.)	Large-scale data-sets containing responses to survey questions from individual respondents in each participating country are available as integrated data files (and country-specific files) for each round. In total, this consists of 248,104 cases (approx 45,000 per round). Country-specific files recording parents occupation and sample data files are also available as well as a wide range of methodological data, including tests of reliability, call records, data on interview settings and event data.	ESS Archive: http://ess.nsd.uib.no/ Data available for the latest round (2010): http://ess.nsd.uib.no/ess/round5/download.html
Maximal deposit size in TB	Not known	
Long term guarantees / standards of trust	Via Norwegian Social Science Data Services (NSD)	
Checks on quality / quality control	Rigorous quality checks are carried out as part of routine data processing for each country's data and documentation. Files are not released with known errors unless it is not possible to correct these. Any errors that are detected between data releases are released as separate corrected files (if possible), or users are made aware of the error in an 'alert' section of the archive. Errors/data deviations are usually detected through secondary analysis by data users and reported back to the ESS Archive Team. The errors detected vary and solutions are dependent on the nature of the error detected and the availability of information to rectify these.	Kolsrud, K., Midtsæter, H., Orten, H., Kalgraff Skjåk, K., and Øvrebø, O. (2010) Processing, Archiving and Dissemination of ESS data. The Work of the Norwegian Social Science Data Services ASK Research Methods 19, 1 51–92.
Dissemination		
Costs / Conditions for Access	To access data files, individuals have to register as an ESS data user – supplying a few personal details and a valid email address. Once registered, users have immediate access to the data online, and are also able to download files. There are five 'conditions of use' that apply to those wishing to use the ESS data and documentation.	ESS Archive: http://ess.nsd.uib.no/ Conditions of use: http://ess.nsd.uib.no/ess/conditions.html
Tools / Interfaces used for Access	Dissemination formats for data: SPSS/PASW, Stata, SAS, NSDstat, Statistica, DIF, Dbase, Text delimited; Dissemination formats for metadata: PDF, HTML, MS Office. ESS Data Archive website (including Cumulative Data Wizard Multi-level data resource); Online analysis and distribution tool Nesstar (built on the standardised tag library Document Type Definition (DTD), developed by the Data Documentation Initiative (DDI)); ESS teaching resource ESS EduNet; ESS Bibliography	http://ess.nsd.uib.no/ http://nesstar.ess.nsd.uib.no/webview/ http:/essedunet.nsd.uib.no/ http://ess.nsd.uib.no/bibliography/ Kolsrud, K., Midtsæter, H., Orten, H., Kalgraff Skjåk, K., and Øvrebø, O. (2010) Processing, Archiving and Dissemination of ESS data. The Work of the Norwegian Social Science Data Services ASK Research Methods 19, 1 51–92.