Data Service Infrastructure for the Social Sciences and Humanities

EC FP7

Grant Agreement Number: 283646

Deliverable Report

Deliverable: D4.3
Deliverable: List of Recommended Deposit Services for SSH
Nature: R

Responsible: DANS
Work Package Leader: NSD

Contributing Partners and Editors: Arjan Hogenaar (DANS), Paula Witkamp (DANS), Marika de Bruijne (CentERdata); Arnaud Wijnant (CentERdata); Erik Balster (CentERdata); Trond Kvamme (NSD), Vigdis Kvalheim (NSD), Astrid Recker (GESIS), Johan Fihn (SND), Torbjörn Berglund, (SND), Birger Jerlehag (SND), Anje Müller Gjesdal (UiB); Carla Parra (UiB); Bamba Dione (UiB); Koenraad De Smedt (UiB); Claudia Engelhardt (UGOE); Jens Ludwig (UGOE); Przemyslaw Lenkiewicz (MPI/TLA).

Executive Summary

This report was produced in the context of the project Data Service Infrastructure for the Social Sciences and Humanities (DASISH) work package 4.3 Convergence of Data Services. The goal has been to allow the selection and promotion of high-quality deposit services for researchers in the Social Sciences and Humanities (SSH) and to make suggestions for service improvements.

A survey was sent to 89 persons working at existing and developing data archives services (DASs) in Europe. With one exception, all these DASs have a scope related to at least one of the following infrastructures: CESSDA, CLARIN, DARIAH, ESS, and SHARE[1]. The survey had a response rate of 61%. The respondents are from 54 organisations, 42 of which have a fully functioning DAS, 9 a DAS under development and 3 no plans to set up a DAS.

The survey results reveal that CLARIN and DARIAH are relatively often interconnected, whereas ESS and SHARE are infrastructures with a strong basis in the Social Sciences and therefore are more often related to CESSDA. About one third of the DASs have relationships with two or more ESFRIs.

The maturity of the 42 existing DASs is related to the availability of a mission statement, deposit agreement, code of conduct, and a preservation policy. It appears that in southern Europe the maturity is somewhat lower than in other parts of Europe. The survey results indicate that the DASs from North-Western Europe have generally spoken reached a higher trust level than the ones from Eastern and Southern Europe (percentage of trust level 1 and above approximately 60% resp. 25%). The also indicate that the DASs within CESSDA DASs the maturity rate is somewhat higher in comparison to CLARIN and DARIAH.

The availability of a mission statement within a DAS is strongly correlated to trustworthy activities. Only about half of the respondents mention the existence of a preservation policy. This policy is not always accessible online or available in English. Further, it appears that the majority of the DASs has already implemented deposit and user agreements. A majority of the archives has a long-term preservation strategy, in most cases migration. About half of the DASs is involved in (self-)audit or certification activities intended to increase trustworthiness in the services. The Data Seal of Approval (DSA) is the most usual instrument for certification. The majority of services is publicly funded and in nearly all cases the cost for deposit or for access is borne by the DASs.

To gain more insight in the policies and views of DASs, in-depth interviews have been conducted with representatives of six different archives, related to the ESFRI's CESSDA, CLARIN, and DARIAH. Based on the information gathered during this task, a list of high-quality and promising DASs has been composed and suggestions for further improvements of existing DASs have been made.

[1] CESSDA: Council of European Social Science Data Archives. www.cessda.org
CLARIN: Common Language Resources and Technology Infrastructure. www.clarin.eu
DARIAH: Digital Research Infrastructure for the Arts and Humanities. www.dariah.eu
ESS: European Social Survey. http://www.europeansocialsurvey.org/
SHARE: Survey of Health, Ageing and Retirement in Europe. http://www.share-project.org

Table of Contents

1 Scope and Characteristics of Data Archive Services within the DASISH Communities

1.1 Introduction

Following the “Description of Work”, Annex 1 to the grant agreement for the project "Data Service Infrastructure for the Social Sciences and Humanities (DASISH)" task 4.3 (Convergence of Data Services) will lead to selection and promotion of high-quality deposit services for SSH researchers and to concrete suggestions for service improvements. The task relates to tasks 4.1 and 4.2 in so far as it will describe and analyse a selection of existing and emerging institutional and academic deposit services within the Social Sciences and Humanities (SSH) based on the framework and guidelines that were created in tasks 1 and 2.

This report focuses on existing or planned data archive services (DASs) within the Social Science and Humanities (SSH), ESFRI’s CESSDA, CLARIN, DARIAH, ESS and SHARE[2]. In this respect a Data Archive Service (DAS) may be defined as any service that accepts research data and accompanying documentation for the purpose of curating and preserving this data for a given timespan, and to make it available for access and re-use by a user community.

It investigates how archiving services and data dissemination are realised by the respective DASs. The report aim to help understand similarities and differences between the services, in particular with regard to functionalities and the levels of quality and trustworthiness achieved.

Making suggestions regarding a possible convergence is only feasible after having made an inventory of existing and developing data archive services (DASs) with their scopes and characteristics. To be clear, convergence relates to implementing similar baseline standards in processes like ingest, archiving, and dissemination and in using similar licence and usage agreements in the different ESFRIs represented within the DASISH project, however without losing sight of the specific needs of a particular DAS.

Accordingly, it was decided to set up a survey with the aim to gain broader insight about the organisation and general state of the DASs across Europe. Design of the survey was based on the report of WP4.2 describing existing DASs and on the Data Archive Description Sheet (DADS) developed to summarise these findings. The results, as presented in this report combined with the two previous reports within WP4, will be beneficial both to policy advisors and to researchers within the SSH-fields as to the DASs themselves. This is especially the case for DASs under development, as they can learn from the organisation of other, already existing, DASs.

The questionnaire was sent to 89 persons working in existing or developing data archive services. 46 respondents answered all questions in the questionnaire. Considering this number, differences in region or ESFRI in the outcomes of the questionnaire give only an indication of the variations that may exist. The outcomes of the survey discussed in this report, together with the reports of DASISH tasks 4.1 and 4.2, have formed the basis for in-depth interviews with selected

[2] CESSDA: Council of European Social Science Data Archives. www.cessda.org
CLARIN: Common Language Resources and Technology Infrastructure. www.clarin.eu
DARIAH: Digital Research Infrastructure for the Arts and Humanities. www.dariah.eu
ESS: European Social Survey. http://www.europeansocialsurvey.org/
SHARE: Survey of Health, Ageing and Retirement in Europe. http://www.share-project.org

data archive services and will be used to initiate discussions with promising candidate organisations to further improve their services.

1.2 Creating the Survey

The survey instrument was designed based on the results of DASISH task 4.2 and on the Data Archive Description Sheet (DADS published as an appendix to the WP4.2 report (DASISH_4 2_appendix-DADS.pdf[3]) for DASs included in the report.

The original DADS was rewritten and extended to create a web questionnaire that was adapted after ample discussion amongst the 4.3 task members (see annex 1). CentERdata was responsible for implementing the survey instrument.

1.3 Questionnaire procedure

European organisations with a relation to the five DASISH infrastructures were invited to participate ihe survey which was fielded in the period between 20 September and 4 November 2013. All these organisations are supposed to offer or plan to offer a Data Archive Service within the fields of Social Sciences and Humanities (SSH).

Eventually the questionnaire were sent to 89 persons active in 73 DAS, 7 from Southern Europe (Spain, Portugal, Italy and Greece), 15 from Eastern Europe and 67 from the other European countries (North-Western Europe; see Figure 1). Especially Southern Europe was under-represented. In a few cases several contact persons from the same organisations were invited. Sometimes because it was not clear who was the right person to contact, in other cases because the organisation was active in more than one ESFRI having a contact person for each ESFRI.

Of the invited persons, 27 persons were primarily related to CESSDA, 25 to CLARIN, 23 to DARIAH, 2 to ESS, and 3 to SHARE.

[3] https://theuniversityofgothenburg.basecamphq.com/projects/8978596/file/169781235/DASISH_4%202_Appendix-DADS.pdf

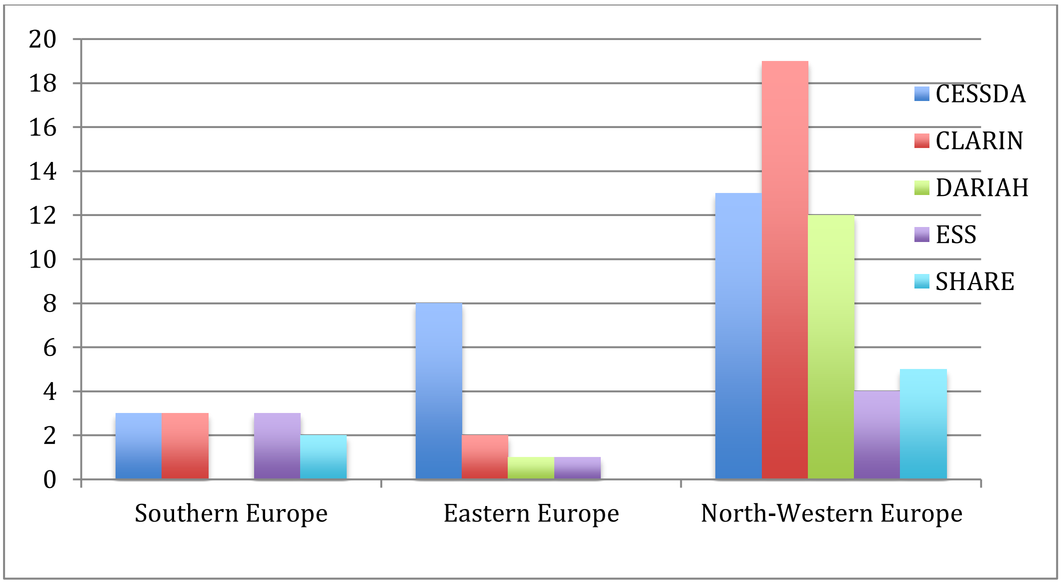

Figure 1: Division of invited persons over region and ESFRI (n=89)

During the fieldwork period of the survey, task members made several attempts to reach non-respondents to ask for the reason for their non-response and to encourage them to take part in the survey. If it appeared that other persons within an organisation were more suitable to contact, these persons were invited to participate.

1.4 Overview of the response

Of the 89 selected persons (from 73 different organisations), 33 (37%) are non-responders. From two organisations duplicate answers were received. The 54 respondents (response rate 61%) are from 54 individual organisations, 42 of which have a functional DAS, 9 a DAS under development and 3 no DAS at all. The 3 respondents from organisations that have no intention to launch a DAS only answered two introductory questions of the survey. From the respondents with a functional DAS or DAS under development, 46 answered the questionnaire completely, whereas 5 respondents submitted incomplete questionnaires.

Of the 42 established DASs, 30 come from North-Western Europe, 8 from Eastern Europe and 4 from Southern Europe. From the nine planned DASs, four come from North-Western Europe, four from Eastern Europe and one from southern Europe.

Of the established DASs, 22 have a scope related to CESSDA, 17 related to CLARIN, 9 related to DARIAH, 8 related to ESS, and 4 related to SHARE (multiple answers were possible in the survey). For the planned DASs 2 are related to CESSDA, 4 to CLARIN, 2 to DARIAH, 4 to SHARE and 1 to none of these 5 ESFRIs.

1.5 Results of the survey

1.5.1 Organisational Context

As the task of DASISH 4.3 is to study the possibilities of a convergence of data archive services within the SSH ESFRIs, it is important to know how the 5 ESFRIs were represented in this survey. The survey revealed that only one respondent had indicated that the scope of their organisation was not related to any of the five SSH-ESFRIs. The other 53 respondents answer that their organisations have a scope related to one or more of the ESFRIs (multiple answers allowed; see Figure 2).

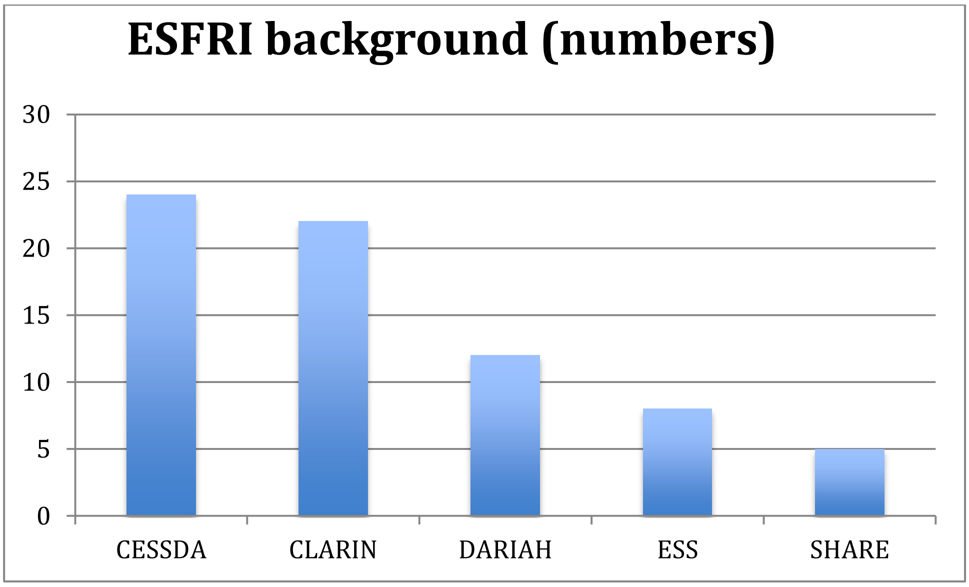

Figure 2: Distribution of the respondents over the 5 ESFRIs within DASISH (n=53, multiple answers possible)

The number of respondents with a scope related to CESSDA or CLARIN is comparable, but the number of respondents having a scope related to DARIAH is lower.

Only a few respondents have a scope exclusively related to DARIAH. There is only one respondent having a scope related to CESSDA, DARIAH, and CLARIN.
As was to be expected from their proximity related to discipline, there is a relatively high number of co-relationships between CLARIN and DARIAH (both humanities ESFRIs), while the respondents with a CESSDA-background in most cases – if a co-relationship exists – have a co-relationship with the ESS and SHARE ESFRIs (all three with a strong social science character).

Nearly all organisations involved in data archive services offer other services as well. This is a broad range of services. Most frequently mentioned are training/teaching/consultancy. Other services mentioned are the offering of web services and of web tools.

Regional background of respondents

Of the respondents 11% come from Southern Europe, 22% from Eastern Europe, and 67% from North-Western European countries**.** This distribution is quite comparable with the distribution of invitations (9% southern Europe, 19% Eastern Europe and 72% North-Western Europe).

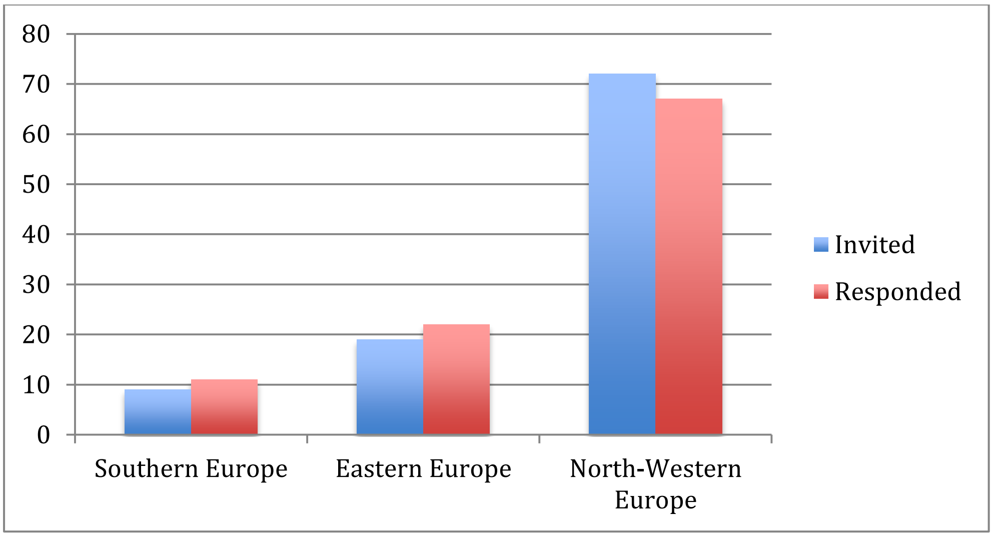

Figure 3: Regional background (in percentage); invited n=89; responded n=54

In Eastern Europe we see a relatively low number of CLARIN and DARIAH respondents (9%, resp. 8% of the respondents). Southern Europe is not represented in the DARIAH-respondents (0%).

Southern Europe is somewhat overrepresented in the ESS and SHARE respondents (38% resp. 29%), whereas Eastern Europe is underrepresented respectively not represented in ESS and SHARE (13% resp. 0%). Finally, North-Western Europe is overrepresented in DARIAH (92%) and CLARIN (83%).

Start population

As the questionnaire is part of the study into data archive service convergence, it was necessary to find out whether the respondents have a data archive service – or have plans to launch one in the near future. Three of the 54 respondents (one from Germany (SHARE-related), one from Italy (multidisciplinary), one from Denmark, DARIAH-related) appear to have no plans in that direction. For them this negative answer automatically led to the end of the survey. Therefore, all other questions have a start population of 51 respondents from 51 organisations with a (planned) DAS.

Nine respondents indicate to plan a launch of a data archive service, six of them with a launch date in 2014.

Funding

Data Archive Services play an important role in the (inter-) national sharing of data. The type of funding is important with regard to stability/sustainability of the services. From the 51 organisations, 48 have some kind of public funding (totally or partly). Other sources mentioned are third party funding[4] (20 respondents) or revenues (10 respondents). Third party funding is mentioned once as the only source of funding. Funding from membership fees is not mentioned and funding from public-private partnerships is mentioned only once.

The same is true with respect to the funding from other sources. Out of the six positive answers, two respondents answer to get funded by the mother organisation (implicit a kind of public funding). The other four respondents answer that funding of the service comes from projects or grants, indicating a form of third-party funding. Two of these respondents are from

[4] In case of third party funding this may also relate to indirect public funding (for instance via a national research funder).

organisations planning to launch a DAS. The other two are from North-Western European organisations acting on a national scale in CLARIN-related disciplines.

Designated communities

According to the OAIS model[5], one of the responsibilities archives has to fulfil is the definition of their designated community. The latter is "[a]n identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time" (CCSDS 2012, p.1-11). The designated communities included in the survey are social scientists, linguists and historians as possible responses, with the option for the respondent to enter any other community.

'Social scientists' are mentioned by 21 Data Archive Services (41%), followed by 'linguists' (9 respondents, 18%) and 'historians' (3 respondents, 6%). The category 'other' yielded 18 answers (35%; see Figure 4).

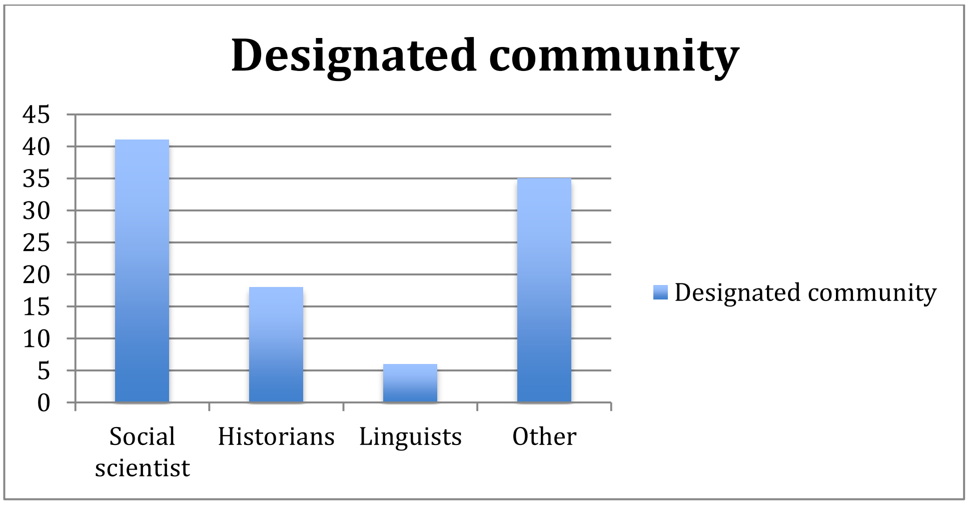

Figure 4: Designated communities indicated by the respondents (in percentage)

Thirteen of the DASs who selected the option 'other' in this question appear to have relationships with several designated communities. They have a broader scope, focusing on multiple disciplines within the SSH, and are in most cases members of the CLARIN-community. Thus it seems that it is not possible to limit the designated community of CLARIN-related DASs purely to linguists. Three DASs focus primarily on archaeologists, one on philologists, and one DAS explicitly mentions broadcast media organisations as its designated community.

Mission statement

Almost two-thirds (65%) of the respondents were capable to refer to a mission statement. A comparison of the contents of these mission statements revealed that the available statements referred directly (or indirectly, for instance via the CLARIN mission statement) to concepts like data storage, data archiving, data preservation and data dissemination.

Of the 18 organisations that did not have a mission statement, 9 were identical with the ones that intend to launch a data archive service in the years to come. Of the organisations with an

[5] Reference model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2, Magenta Book, June 2012, http://public.ccsds.org/publications/archive/650x0m2.pdf

established DAS, six organisations without a mission statement were from North-Western Europe, one from Eastern, and one from Southern Europe. Seven of these eight DASs were CESSDA-related. One organisation gave incomplete answers.

Having a mission statement seems to be correlated to the availability of deposit agreements. Only 46% of the DASs without deposit agreements have a mission statement, but 76% of DASs with a deposit agreement also had a mission statement. An even stronger correlation exists between the level of trust of a DAS (see below) and the availability of a mission statement. Over 80% of the DASs with activities in the field of trustworthiness have mission statements. For DASs without trust-related activities this percentage is only 41%.

Deposit agreements

The preservation and dissemination of data has a number of legal implications relating, among others, to questions of ownership / intellectual property rights or data protection. License or deposit agreements can help to clarify the legal situation and serve to protect the archive service preserving and disseminating the data. 38 of the 51 respondents (75%) use such an agreement. Of the 13 other respondents, 5 are future data archives, indicating that 8 established data archive services currently do not employ license or deposit agreements.

Looking at differences per ESFRI, it seems that the implementation of deposit agreements is somewhat less common in DARIAH-related organisations (42%, n=12) in comparison to CESSDA (83% n=24) or CLARIN (73%, n=22). For 8 of the 38 archives with licence/deposit agreement these agreements are not available online (from which 3 indicate that they have plans to do so). 7 agreements are not available in English. One archive works with submission forms in which the depositor describes the access conditions and possible embargo period.

Rights

Preserving digital data for the long term is impossible without creating copies (for back-up and dissemination) and without altering the digital object in question, e.g. by converting it to a different file format when the original format threatens to become obsolete. These actions can mean an infringement of the data producer's intellectual property rights – it is important, therefore, that the data archive service is granted the rights necessary to preserve the digital assets by the rights holder.
Three respondents to the survey state that the data archive service does not obtain any rights from depositors/data producers. 25% (number) of the respondents answer that in their data archive service the depositor retains all rights. The rights-policy most often mentioned is the transfer of non-exclusive rights to the archives. 23 archives (45%) work with this concept. Analysis of the category 'other' revealed that there are 3 other archives working with non-exclusive rights transfer, bringing its total to 26 archives (more than 50%). There were no respondents mentioning that their archives become the exclusive rights owners of the data.

Usage agreements

For a DAS it is important that roles and responsibilities are clear for all parties involved (depositor, archive, and user). Archived datasets may only be accessed and re-used under certain conditions, described in so-called user agreements.

35 of the 51 respondents (69%) have a usage agreement (see for example the CLARIN data user agreement[6] and the SHARE user agreement[7] available on the Internet). Of the other 16

[6] See: http://repos.ids-mannheim.de/resources/DataUserAgreement.pdf

respondents, 5 are from future data archive services and 1 from a respondent who has not answered all the questions in the survey, making it impossible to say something on the character of this specific DAS. Therefore, 10 respondents from existing data archive services answer not to have a usage agreement. Here too, the percentage of DASs with a usage agreement within DARIAH seems to be a little bit lower (58%, n=12) than in CESSDA (75%, n=24) or CLARIN (68%, n=22).

An analysis of the different usage agreements reveals that many variations exist, all describing different, specific conditions. Important common conditions in the licences are preserving confidentiality (if relevant), citing data correctly, and informing the data archive service if data is used for publications. If permitted at all, commercial re-use is only allowed after prior permission from the right holder(s). Sometimes a data archive service chooses to offer tailor-made user agreements, based on the needs of the very user.

Preservation policy

A preservation policy is an important document demonstrating an organisation's commitment to the preservation of its digital collections. It can be defined as a "[w]ritten statement, authorized by the repository management, that describes the approach to be taken by the repository for the preservation of objects" (CCSDS 2011, p. 1-4). 26 (about half of the respondents) indicate that they have such a preservation policy. In 12 cases the policy is not online available (yet), in 6 cases it was written in languages other than English.

Further analysis of the answers revealed that for most services an elaboration of this preservation policy is needed, as details are not often described in the documents respondents refer to.

1.5.2 Deposit and Ingest

Accepted and preferred file formats

For researchers who want to deposit data it has to be clear which file formats are accepted and curated by the data archive service. 17 respondents (one third) do not mention a restriction in format for data and documentation in their DAS. On the other hand, five respondents (10%) state that their services only accept data in specific formats. Three of these services are related to CLARIN which has formulated standard recommendations for the Language Resources and Technology domain[8].

Many respondents (19, i.e. 38%) indicate that their DAS works with a list of preferred formats in order to have some control over the formats of the ingested data without introducing possible obstacles that hinder the deposit process. Eight respondents describe a somewhat more flexible practice: the data archive services in question will start negotiations with data depositors, giving them advice regarding the file format and – if needed – carrying out a normalisation process by converting less suitable formats into more accepted ones.

The survey also asked respondents for the file formats most often curated by their data archive services. The file formats mentioned most often are formats associated with statistical packages

[7] See: http://www.share-project.org/data-access-documentation/research-data-center-access.html

[8] See: http://www.clarin.eu/node/2230

such as SPSS, STATA and SAS (14 respondents); pdf (11 respondents); Excel/csv (10 respondents) and xml (7 respondents).

Metadata

In order to administer, preserve and retrieve ingested data, the use of metadata is indispensable. Ideally, this metadata should be standardized (e.g. following international or community standards).

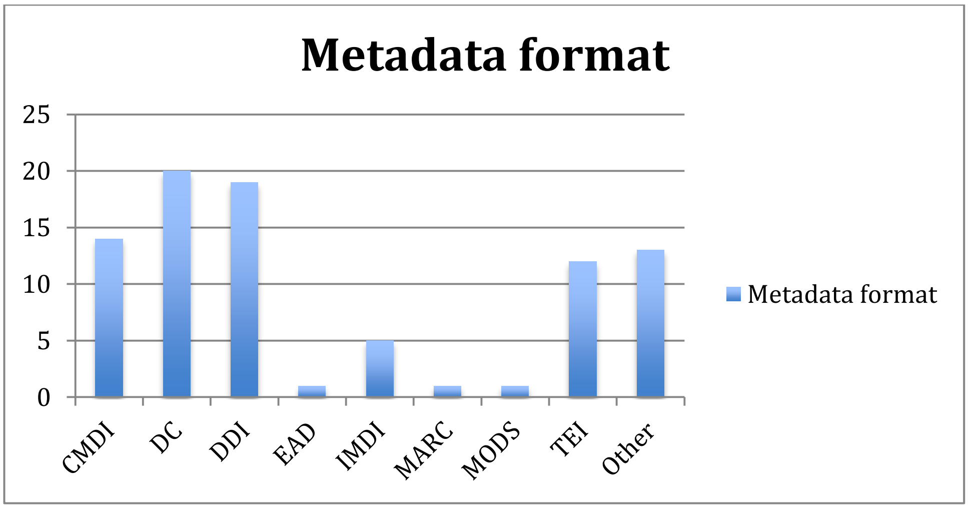

Figure 5: Metadata formats used by data archive services (multiple answers possible; n=88). Respondents indicated as other formats Metashare, OLAC, and LRMI.

In this respect, it has to be stated that DDI may not only be considered as a general metadata standard. DDI has its background in the social sciences and therefore it is also a discipline-specific metadata standard.

Not surprisingly, among the ten respondents that indicate that their services use discipline-specific metadata standards, besides specific adaptations of Dublin Core, Metashare and MPEG7, DDI, DDI2, and DDI3 are mentioned.

Routes of ingest

The usability and acceptance of a data archive service also depends on how accessible and convenient the submission process is. Occasionally, depositors bring the data personally (one respondent). More often, data will be sent to the archive. Archives normally offer several routes to deliver the data.

16 respondents mention the use of online deposit forms by their services. Out of these 16

- 4 also offer the options to send data by e-mail and/or by CD/DVD
- 6 also offer to option to send by CD/DVD and
- 4 also offer other delivery options (SWORD; file sharing; SOAP; secure FTP; harvesting of the data by the archive).

Deposit fees

Research into the cost of digital preservation suggests that ingest is one of the most expensive activities in this context (see, for example, LIFE and Keeping Research Data Safe)[9]. At the same time, the amount of digital information produced grows exponentially and to keep digital

[9] http://www.life.ac.uk/, http://www.beagrie.com/krds.php

preservation sustainable it may be necessary in the future to consider charging depositors for cost-intensive preservation activities. Most of the respondents (42 out of 50) indicate that their services do not charge the depositors at the moment. 2 respondents answer that plans exist to introduce fees for data deposit and 6 respondents state that their archives already charge depositors, although they distinguish depositor categories (for instance: researchers have the right to deposit their data at no cost). One DAS is considering the introduction of a complex deposit fee regime, based on type of material, size, complexity, and format.

Access fees and access conditions

41 of the existing data archive services state that no costs are involved for accessing the data. 5 respondents indicate that their services have some type of charging. This may depend on the user category (commercial or non-commercial), on the dataset type or on special operations to be executed by the data archive (like burning DVDs).

Apart from any cost restrictions, other restrictions may be implemented in accessing the data. Examples of such conditions are access only to academia, access only for registered users or depositor-defined access. Of the 46 respondents that answered this question, 36 indicate that their services involve access conditions. The questionnaire offered no possibility to elaborate this answer.

1.5.3 Archival Storage and Preservation

Size of data archives

Across the ESFRIs the size of data archives (in TB) varies to a great extent. The smallest size mentioned is 0,01 TB. The biggest size is 7000 TB, given by a data repository specialised in archiving broadcasted information. For the 42 already established archives, in 8 cases the respondents declared they were not able to give an answer, in 2 cases no answer was given at all.

Not surprisingly, it was even more difficult to give an indication of the size of the archive in terms of 'number of data sets' or 'number of data files'. According to the respondents, the size of existing data archives varies between 10 and 25,000 datasets and 35 to millions of data files. This means that it could be interesting to realise a mapping of archives based on size, apart from a mapping based on region or ESFRI.

As the size of a dataset may vary, it is important to know if a data archive will only accept datasets below a maximum size. Only 2 of the 50 respondents indicate that their services have adopted a policy to limit the size of deposited datasets. Unfortunately, they were not able to give the volume of the maximum size.

Retention period

How the preservation process is shaped in a particular data archive service strongly depends on the retention period aimed at and guaranteed by the archive. At the same time, depositors and users of data archive services need to know for how long a resource they deposit or re-use will be available. This may impact their decision to use a particular DAS. For the 42 established data archives services, 34 respondents answer that their services aim to preserve the research data indefinitely. In one case a guarantee is given by the data archive service that deposited data will be preserved for 50 years.

Four respondents declare that the guarantee of the service depends on the continuation of the funding. One respondent indicates the service has different preservation levels relating to the type of material deposited and depending on the arrangements made with the depositors.

Preservation strategies

Preservation strategies are an important instrument to maintain the accessibility of the digital assets and to counter the risk of technological obsolescence. Not a preservation strategy in its own right, bitstream preservation, aimed at maintaining the bits of the digital object unchanged and uncorrupted, forms the basis for strategies aimed at maintaining the usability of digital objects: emulation and migration.

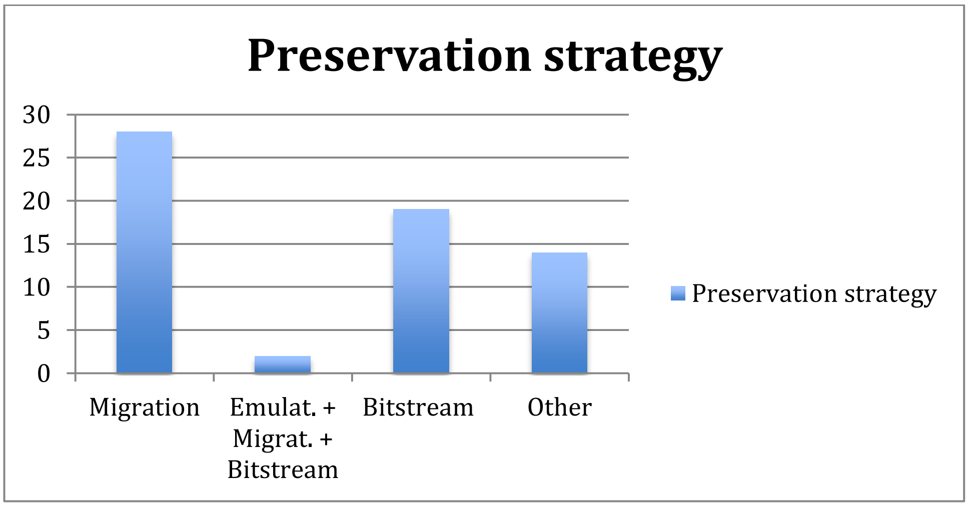

Figure 6: Preservation strategies used by data archive services (multiple answers possible; n=64)

28 respondents employ migration, often in combination with bitstream preservation (14). Only two respondents mention emulation as a preservation strategy of their service, and only in combination with migration and bitstream preservation. Due to a flaw in the survey, it was unfortunately not possible to trace the content of the other preservation strategies, mentioned in 14 cases. Of the DASs with an indefinite preservation period (38), 26 have chosen for migration and 2 for emulation according to the respondents. 3 of these DASs have chosen for bitstream preservation only and 7 for 'other preservation strategies'.

Trust

As stated before, most of the respondents indicate that their services intend to preserve the ingested data indefinitely. Third party assessments can give some indication of the trustworthiness of the data archive services[10]. 24 respondents of 46 existing data archive services and 1 of a future data archive service indicate that their services have undertaken activities to determine the level of their trustworthiness. At the same time, 15 respondents from existing data archive services indicate that these services have not undertaken any action in this respect yet.

[10] Report about the Preservation Service Offers. Deliverable 4.2 of the DASISH project. Available via http://dasish.eu/publications/projectreports/D4.2_-_Report_about_Preservation_Service_Offers.pdf/

Eight services underwent a DSA/DRAMBORA self-assessment (i.e. trust maturity level 2), nine services underwent a peer-reviewed DSA self-assessment (trust maturity level 3) and six services are in a preparatory phase to reach a DSA-self assessment (trust maturity level 1). This last group of respondents answered this question with “other”. Only one respondent declared that their service reached the level of a peer-reviewed ISO 16363 self-assessment (trust maturity level 4). None of the data archive services has reached trust maturity level 5 (external audit based on DIN 31644 or ISO 16363) yet.

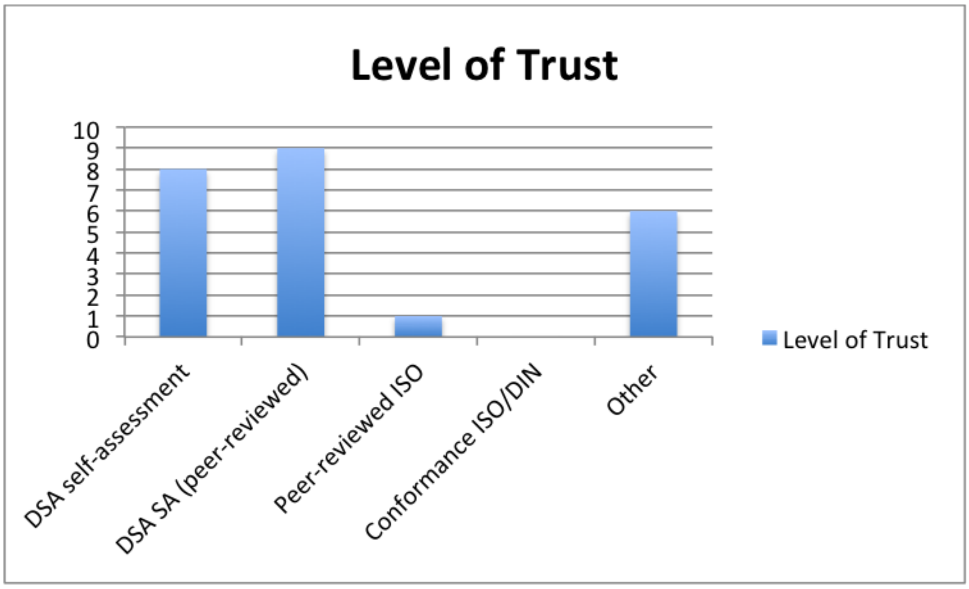

Figure 7: Level of Trust of data archive services (n=25)

The ten DASs involved in trustworthiness activities above level 1 are all situated in North-Western Europe, (see Figure 8: Level of trust per region). However, about a quarter of the DASs in Southern and Eastern Europe has started to improve their trustworthiness (4 respondents).

A similar comparison of trust level can be made for the different ESFRIs. The only statement that can be made is the fact that almost half of the DASs within the ESFRIs have not undertaken trustworthiness activities yet.

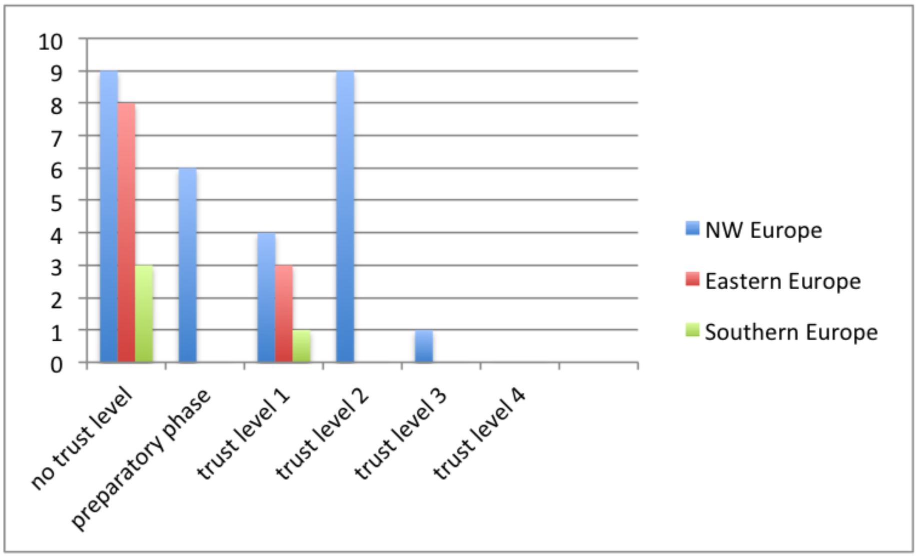

Figure 8: Level of trust per region

Data authenticity and integrity

An important objective of digital preservation is to maintain the authenticity and integrity of the digital objects. Thus, data archives have to take measures to make sure that the digital object remains complete, uncorrupted, accessible and reliable (“trustworthy”). This can be achieved with different technical and administrative measures.

38 respondents stated that they employ access controls to protect the integrity of resources. Of the 38,

- 9 also use check sums,
- 1 also uses audit trails,
- 10 also use check sums and audit trails,
- 2 respondents use check sums only,
- 6 respondents have not implemented any measures yet.

The remainder reports that measures implemented in the services are the registration of provenance data (strongly related to audit trails), backups, and workflow registration (a protocol-based authenticity and integrity policy).

1.5.4 Dissemination

A data archive service will ingest, archive and curate data with the goal to make these data accessible and re-usable by other users (researchers). Giving access to the data is possible by many ways. Most mentioned by the respondents in this survey are websites (29 services), online catalogues (24 services) and special interfaces (7 services).

1.5.5 Future developments

All survey results discussed until now were related to the existing situation. However, DASISH has a focus on the future: convergence of the data archive services within the SSH. Therefore, the last questions in the survey are related to future developments.

Maturity level of Data Archive Services

We asked the respondents if they are satisfied with the maturity level of their data archive service and in which areas they were planning to make improvements. We split this item into 5 sub-items:

a. data archive administration
b. ingest facilities
c. archival storage & preservation
d. dissemination facilities
e. other measures

Respondents had the option of providing information on as many or as few of these sub-items as relevant to their DAS.

The maturity level of DASs is strongly related to the level of trust. Thus a higher maturity level is likely to correlate with a higher trust level (see DASISH report delivered by task 4.1). The criteria for the different trust maturity levels (see below) form an indication of the status of all preservation-related processes within a specific DAS. So, the answers given by the respondents from the different organisations give an indication of the trust level reached by them and so – implicitly – what is needed for them to reach a higher trust level.

Data archive administration

21 of the 46 respondents answer that improvements in the field of administration are needed. Looking at the elaborations of these answers, we see differences between the ESFRIs[11]. CESSDA-respondents tend to be more focused on formalisation of procedures and improvement of the systems in general. The CLARIN- and DARIAH-respondents tend to have a focus on policy: development of policy papers, preservation plans, rights and access management, and documentation of the processes.

Ingest facilities

25 of the 46 respondents indicate that measures in the field of ingest facilities are needed. Here too, some differences between the ESFRIs exist. CESSDA-respondents stress the need for automated upload and ingest and easy deposit procedures in general. The CLARIN-respondents, however, see the support of additional metadata formats and insight in file formats as areas where additional developments are required. Finally, the DARIAH-respondents give similar answers as the ones from CLARIN, but emphasise the import of descriptive metadata.

Archival storage and preservation

In the field of storage and preservation 20 respondents saw possibilities for improvement. Among the CESSDA-respondents there is a need for Persistent Identifiers (PIDs) and collection of provenance information. In addition to this, CLARIN-respondents want more attention to be paid to data integrity and authenticity. DARIAH-respondents would like to have a stronger focus on the software used for archival storage and preservation.

Dissemination facilities

The highest numbers of respondents (26 out of 46) see improvement possibilities regarding the dissemination of digital data in their data archive services.

The CESSDA-respondents see a need for improved searching and browsing facilities and for flexible access systems. In addition to this, CLARIN- and DARIAH-respondents ask for the implementation of Single-Sign-On (SSO) procedures and bulk download opportunities.

Other aspects

Respondents were given the opportunity to describe possible improvements for their data archive services not directly related to the four sub-items described above. Relevant remarks concern further certification; implementation of annotation tools; creating a network of data archives (in the humanities); giving support in data management; improved documentation; improved Security Assertion Markup Language (SAML)-based authentication for users/ AAI-Shibboleth (Authentication and Authorization Infrastructure).

No measures

Six of the 46 responders (13%) appear to be satisfied with the maturity level reached and indicate that in their opinion no measures are currently needed. Three of these come from Eastern Europe, one from Southern Europe. Two of these respondents are from data archives that have not been launched yet.

[11] In these paragraphs on future development the focus will be on differences between CESSDA, CLARIN, and DARIAH as ESS and SHARE have DASs that are strongly related to the ones in the ESFRI CESSDA

Future cooperation

All respondents indicate to be interested in further cooperation within one or more of the 5 ESFRIs.

Final remarks

The respondents had the opportunity to place remarks/comments at the end of the survey. Nine respondents have commented, two from future data archive services, indicating that their answers are preliminary. Two comments relate to the time needed to answer the questionnaire: more than 10 minutes needed. One remark was on q25. This question had no room for elaboration (presumably q26 is meant). Indeed, a mistake in the survey. The same is true for the explanation of q7 and q8. In the earliest online versions the examples given next to these questions have been mixed up. So, quick responders were confronted with some inclarity here. Finally, the suggestion was made that yes/no were not always the right options to answer a question. Many aspects of the data archive services are still under development and a third option 'under development' would have made it easier to answer the questions.

2 In-depth Interviews with Promising Data Archive Service

2.1 Selection of interview candidates and goals of the interviews

The survey conducted winter 2013/2014 amongst existing DASs and DASs under development revealed already a lot of interesting information on how DASs have been organised within the various Infrastructures and European regions.

In order to get a more detailed picture, it was decided to conduct in-depth interviews with representatives of a limited number (six) of *promising* DASs. "Promising DASs' are defined here as DASs that do not belong to already well-known existing DASs like DANS, GESIS or NSD, but have the potential to develop into high-quality deposit services.

In the selection procedure we've tried to realise an even representation of the different ESFRIs. It was agreed to conduct interviews with the following organisations:
BAS; UFAL/LINDAT; TextGrid; DRI; DDA, and ADF.

DAS	ESFRI	Country	Region	Interviewee	Function	Interviewer
BAS	CLARIN/ SHARE	Germany	NW	Florian Schiel	Co-director	Arjan Hogenaar
UFAL/ Lindat	CLARIN	Czech Republic	E	Jozef Misutka	CTO; repository manager	Bamba Dione
TextGrid	DARIAH	Germany	NW	Sibylle Söring; Stefan E. Funk	Consortium manager; developer	Claudia Engelhardt
DRI	DARIAH	Ireland	NW	Aileen O'Connor	Policy manager	Bamba Dione
DDA	CESSDA	Denmark	NW	Anne Sofie Fink	Head of Section	Birger Jerlehag
ADF	CESSDA	Slovenia	E	Irene Vipavc	Senior administrative officer	Birger Jerlehag

2.2 Questions asked during the in-depth interviews

From the outcomes of the survey we know that is it impossible to compare DASs on all relevant aspects of a data archive service. Therefore we have decided to split the interview into two parts: the biggest part consisted of 10 broad questions to be asked to every single interviewee. The other part consisted of 4 smaller questions, only asked if relevant (relevancy based on answers given in the first part of the interview). The interview questions were developed in cooperation with the Task 4.3 members. Interviewers came from UiB, SND, UGOE and DANS. The interviews were conducted in June and July 2014.

2.3 Outcomes of the interviews

In an appendix detailed descriptions of the interviews are given. This section gives the main trends that can be distilled from of the interviews

The goal of the interviews was to get a more detailed picture of six Data Archive Services, coming for the ESFRIs CESSDA, CLARIN, and DARIAH. We have selected promising DASs, as the well-known organisations are familiar with presenting their views to a big audience. The interviewees are mostly from organisations that are already well organised, but are a little bit less pronounced in the international discourse. Looking at the more technical issues to start with, it is remarkable to see that there certainly is a wish to use PIDs. All three major PID-systems (EPIC/Handle; DataCite/DOI; URN:NBN) are being used.

We think an important conclusion is that these DASs see the importance of PIDs for the tracing of deposited datasets. Implementing a PID system is more important than the choice for a specific one, under the condition that interoperability between the different PID-systems needs to be realised. The members of WP7 of DASISH will organise a workshop on this topic in November 2014, indicating its importance within the DASISH community.
Every DAS is free in the way it set ups its data repository. This is exactly what we see in the interviews: technically spoken, there are a lot of differences between the DASs. That is no problem as the functionality offered by them is more or less the same.

Dealing with authentication and authorisation in relation to access to, especially sensitive, information is also a major issue. Shibboleth and EDUGAIN/Edugate are mentioned, but DASs see also advantages of the usage of AAI-systems developed in their own ESFRI (DARIAH AAI or CLARIN IdP). Within CESSDA, AAI realisation is under development. A DASISH Strategic Workshop relevant to this theme was organised earlier in 2014. The overall conclusion was to aim at the development of Federated Identity Management (FIM), allowing simple access for members from one ESFRI to the other ones. We support the conclusion of this workshop as it leaves a specific DAS free in its choice for a certain AAI-system, without hindering the users of the different DASs thanks to the Federated Identity Management.

The sustainability of DASs relates to the financial sustainability and to the way back-ups are stored. Three of the six interviewees give a description of a financial sustainability under development. The financial construction for these mostly rather young DASs is clear until 2015, but new financing models have to be implemented after that time. The interviewees are nevertheless optimistic on the future. The other interviewed DASs receive direct financial support from a university.

Regarding the sustainability of data, all DASs have some kind of external storage in order to make this storage less dependent of the viability of their own organisation.

One interviewee came with the suggestion to have a closer look on the Data-PASS model of the Data Preservation Alliance for the Social Sciences. Data-PASS[12] is a voluntary partnership of organisations for archiving, cataloguing, and preserving data. The current Data-PASS model is only valid for datasets produced in the social sciences in the USA. The conceptual model may be applied to the Social Sciences and the Humanities (SSH) in Europe. Similarly, as in the Data-PASS model, a common European Digital Stewardship Alliance should be created, responsible, amongst others, for long-term preservation of data and the organisation of access of data from the different partners within the Alliance[13].

The interviewees have given an indication that they don't give the highest priority to cataloguing, using classification schemas.

Looking at the organisational aspects of the DASs, we see answers that are a little bit different compared to the responses in our survey. In the OAIS-model the designated primary community

[12] http://www.data-pass.org/

[13] In a way, Data-PASS functions like LOCKSS (Lots Of Copies Keeps Stuff Safe): copies of datasets will be stored on many different places.

is an important aspect of a DAS. The six interviewees say that it is hard to distinguish a sharp defined community. The trend seems to be that datasets originally set up for a specific discipline within the SSH are being used by users from all kinds of disciplines within the SSH or even from outside these scientific fields (natural sciences and medicine). For the deposit-aspects it seems to be easier to identity the primary designated community.

Workshops and training courses are being organised both for depositors and users, although not always in cooperation with the ESFRI. Here is a chance that CESSDA, CLARIN, DARIAH, ESS and SHARE – as an umbrella organisations – promote the development and usage of training tools.

For users of a DAS it is important to see adequate descriptions of the datasets deposited in it. Some kind of uniformity is practical. All interviewees therefore say that a minimum set of metadata should be present for every deposited dataset**.** In most instances, this minimum set of metadata has been implemented into the deposit procedure. Within this process, each individual DAS seems to have developed his own rules. Sometimes one or just a few metadata formats may be used; in other cases the DAS offers the depositor many different metadata formats.

The interviewees know the possibilities to get certified, for instance through the DSA. DSA is popular, all interviewed DASs have obtained the DSA or has started the process to obtain in. For CLARIN centres, the DSA is part of the process to obtain the CLARIN-B status and all CESSDA members are required to obtain the DSA by the end of 2015. Trust is an important aspect of the data repositories. Certification on a basic level is supported, but certification on higher levels (DIN 31644 or ISO 16363) does not seem to be considered. Only the DRI states that it will develop its data archive further according to the recommendations described in ISO16363. This illustrates the advantage of setting up a DAS after the development of certification tools. In a way, these certification tools may be regarded as the building blocks of the creation of a high-quality level DAS.

It is theoretically possible to set up a large-scale interconnection of European data archives. The interviewees agree with that, but prefer at this time the further development of the individual ESFRI-infrastructures, leading to a clear profile of the services within a specific ESFRI. No uniformity exists in the deposit agreements used. Some interviewees recommend the usage of CC licences, but without a strict obligation. In other situations depositors are more or less free in choosing the deposit agreement. It may be worthwhile to investigate the possibility of harmonising deposit agreements within the five SSH ESFRIs.

The most complex question asked to the interviewees relates to the policy framework to govern the total of procedures and guidelines for data management, archiving and sharing. It appears that information on this subject is often only available in internal documents and/or in documents written in the national language. It seems that it is rather difficult to develop clear procedures and guidelines in this respect, leading to delays in publication of this information. It would therefore be advisable to tackle this issue in a broader context within CLARIN, DARIAH, CESSDA, ESS and SHARE or in cooperation between the five.

3 General discussion

The results of the survey and the interviews with representatives of the six DASs clearly indicate that within the different SSH ESFRIs similar developments are taking place to create high-quality data archive services. ESS and SHARE have a different position, as data archiving is not really an explicit activity within these two ESFRIs. The actual archiving and data dissemination services are handed over to organisations that also are active in one or more of the other ESFRIs.

Data Archive Services, especially trusted services, have to be sustainable and play an important role in the (inter-) national sharing of data.

In this respect a problem may arise regarding depositor agreements. In about 25% of the DASs, these agreements do not exist. A high-level data archive must have the right to change deposited data for the matter of long-term sustainability. The absence of a depositor agreement may cause legal problems in the future. The interviewees indicate that recommendation of practical deposit agreements are often given (for instance: the CC licences), but a strict obligation to sign such an agreement is not always present**.** It is advisable to come to a (limited) set of acceptable deposit agreements within the five SSH ESFRIs.

Authentication and authorisation are really important issues in the access policy of a DAS. In the survey no AAI-related question was asked. The interviews reveal that DASs all work with some kind of AAI-tool (Shibboleth, EDUGAIN; EDUGATE). EDUGAIN, as part of the Géant services, is an important tool in the realisation of trustworthy exchange of information between federations represented within Géant.

Related to this problem is the fact that about one third of the DASs lack a usage agreement, especially outside North-Western Europe. Depositors may lose trust in a DAS when it is not clear what the (re-) users are allowed to do with the deposited data. In the light of convergence of data service, a focus on regulations within all ESFRIs regarding deposit and re-use would be a quick win. In this respect it is advisable to have a closer look at the DDA-model, with 5 different access levels.

Nevertheless, the conclusion has to be that a one fits all model will not work for the broad range of disciplines in the SSH. A list of recommended data archive services (and derived from that: a set of recommendations) will certainly help new DASs in the setting-up of their services, but it will depend on their 'customers' (both depositors and users) how they will deal with the technical and organisational details.

Introduction of preservation plans for all DASs will be important at a later stage and is closely related to the policy to increase their trust levels, using certification tools like DSA, DIN 31644 and ISO 16363.

Efforts should be spent on making DASs aware of the shortcomings of using bitstream preservation only as a preservation strategy. Bitstream preservation, which may lead in the end to inaccessible data, should be replaced by migration and or emulation strategies.

The interviewed DASs are rather young. Interesting to see is that they chose in most cases for bitstream preservation and certification on the DSA-level. More complex preservation strategies or certification at a higher level are not the issues relevant to repositories in development. But what is not important at this stage will certainly become important later on. Depositors and users would benefit from DASs with clear quality standards. In this respect the willingness of a DAS to obtain DIN31644 or ISO16363 would be an indication of its trustworthiness.

4 High-quality data archive services

The overall aim of the survey and interviews conducted as part of WP4.3 Data Archiving has been to come to - what is described in the DASISH-DoW - as “Selection and promotion of a number of high-quality deposit services for SSH researchers and concrete suggestions for service improvements”.

In WP4 of the DASISH project, the reports produced within tasks 4.1 and 4.2 contain important information that has been used within task 4.3. Task 4.1 has delivered a pragmatic overview of the different trust levels of a DAS, prescribing a 5-level Trust Model, whereas Task 4.2 has delivered a set of concrete recommendations (and a list of key requirements and descriptions of a number of DASs in so-called Data Archive Description Sheets).

In task 4.3 the survey and the interviews yielded background information that has been used to come to a selection of a number of high-quality deposit services.

The selection of these services is not an easy job and may be disputed. We did not expect to find DASs that meet all the requirements as formulated in the report from WP4 on Preservation Service Offers. We have to compare the weight of the different requirements to come to a selection of high-quality deposit services.

In our view, the most important requirements to be met by a DAS are

- Having met the requirements of minimal the DSA
- Availability of a Preservation Policy
- Clear deposit licences
- Clear usage licences
- Clear rights management
- Long preservation time
- Preservation strategy using migration and/or emulation

The last two requirements are relatively less important, as they deal with practical elaborations of the policy and philosophy of a specific DAS.

The DADS, the extended DADS (based on the survey, see appendix) and the interviews yielded the information to make a comparison based on the requirements mentioned above.

As part of Task 4.2 Data Archive Description Sheets (DADS) have been composed for 5 well-known DASs and the ESS- and SHARE-infrastructures. The five DASs have already a strong reputation regarding the handling of and the care for the data deposited there. Not surprisingly, all these DASs meet the important requirements described above.

These DASs are:

UKDA	(UK)	CESSDA
GESIS	(Germany)	CESSDA
NSD	(Norway)	CESSDA
DANS	(Netherlands)	CESSDA, CLARIN, DARIAH

TLA	(Netherlands)	CLARIN

Within Task 4.3 the aim was, based on the requirements and recommendations incorporated in the DADS, to map the European SSH landscape and find out if there are other DASs that may meet these requirements.

The answers in the survey were filled into the DADS and the completed DADS were sent to the respective DASs for validation purposes, making sure that every DAS had the possibility to check and if necessary correct their contents.

The result is an additional set of DASs, which unlike the DADS included in the first assessment and the Preservation Service Offers report, has big variations in characteristics: from mature services into services just started; from services with clear licences to services where the rules for depositing and (re-) using data are not very clear yet.

Nevertheless, this procedure revealed an additional set of DASs that meet the criteria:

ADS	(UK)	DARIAH
BAS	(Germany)	CLARIN; SHARE
CSDA	(Czech Republic)	CESSDA; ESS
Oxford Text Archive	(UK)	CLARIN
RODA	(Romania)	CESSDA
TextGrid	(Germany)	DARIAH
St. Beeld & Geluid	(Netherlands)	CLARIN
UFAL	(Czech Republic)	CLARIN
DDA	(Denmark)	CESSDA
DTARe	(Germany)	CLARIN

Besides, we see a set of promising DASs that have a good chance to be able to meet the criteria in 2015:

ADP	(Slovenia)	CESSDA	No DSA at the moment
DARIS	(Switzerland)	CESSDA/ESS/SHARE	No DSA; policies not clear yet
Réseau Quetelet	(France)	CESSDA/ESS/SHARE	No DSA
SLDR	(France)	CLARIN; DARIAH	No peer-reviewed DSA
Tarki	(Hungary)	CESSDA; ESS	No DSA

Not functioning yet, but very promising is also the Digital Repository of Ireland. The DRI has the advantage to start later than the other European DASs, so that it is able to build its services according to the description given in ISO 16363. When DRI starts in 2015, it will automatically meet the high ISO 16363 criteria.

The lists above give an indication that in the Southern European countries very few, if any, trustworthy DASs exist. It is recommended to promote the creation of high-level DASs in that region. The situation is better in Eastern Europe, where several DASs have made progress in improving the quality of their DASs.

In general, we see that within CESSDA the number of high-quality DASs exceeds the number of similar DASs in DARIAH or CLARIN. So, apart from the promotional activities in Southern Europe, special attention should be paid for the creation of high-quality DASs within DARIAH and CLARIN.

SHARE and ESS, as stated earlier, have a special position, as these two ESRFIs have not developed their own deposit infrastructure but rely on services from GESIS and NSD.

5 Suggestions for service improvement

The survey and interviews have made it clear that it is very difficult to develop a DAS with all its deposit- and use-licences that fits for every single discipline within the SSH.
This brings us to the following overview of high-quality deposit services:

Of course there is always room for further improvements. Based on the survey and the interviews, we make the following suggestions for service improvement and for the stimulation of the convergence of data services:

1. Every DAS should use a PID-system. A DAS is free in its choice, but on the level of the five SSH ESFRI the focus should lie in the realisation of interoperability of these PID-systems.
2. Following the conclusion of the DASISH Strategic Workshop on AAI, a Federated Identity Management system has to be developed to promote interdisciplinary use of deposited datasets. EDUGAIN, as part of the Géant services, may play an important role in the realisation of trustworthy exchange of information between federations.
3. In Europe a model based dataPASS-model should be developed, creating a common European Digital Stewardship Alliance.
4. The development of training courses is costly. Sharing already developed courses within and across the five SSH infrastructures will reduce the cost for a specific DAS.
5. Certification of DASs is needed, not only as an indication of the level of their trust, but also as a means to set up new DASs or to upgrade existing DASs according to high-level standards.
6. Especially within a specific ESFRI, deposit agreements have to be harmonised.
7. Clear guidelines, procedures and requirements for management, archiving, and sharing of data should be developed by the five RI in SSH, CLARIN, DARIAH, CESSDA, SHARE and ESS in order to realise harmonisation in this respect.

6 Concluding remarks

In this report we try to map the SSH data archive and deposit services landscape within the SSH in Europe based on the results in a survey and six follow up interviews conducted during the fall of 2013 and spring of 2014 respectively.

As it turned out it was not an easy task to get an overview of the quality and services offered by Data Archive Services in Europe. However, we believe that based on the Data Archive Description Sheets (DADS), the survey instrument made it possible to indicate the state of the art with regard to the existence of trustworthy deposit services within SHH in Europe as to make suggestions for further improvements based on internal discussions and on the interviews conducted this summer.

We would like to thank all DASs for their willingness and cooperation.

References

Roadmap for Preservation in the SSH. Deliverable 4.1 of the DASISH project. Available via http://dasish.eu/publications/projectreports/D4.1_-_Roadmap_for_Preservation_and_Curation_in_the_SSH.pdf/

Report about the Preservation Service Offers. Deliverable 4.2 of the DASISH project. Available via http://dasish.eu/publications/projectreports/D4.2_-_Report_about_Preservation_Service_Offers.pdf/

Appendix 1: Regions

Regions

Southern Europe:	Portugal, Spain, Italy, and Greece
North-Western Europe:	Ireland, UK, France, Belgium, Luxemburg, Netherlands, Germany, Austria, Switzerland, Denmark, Norway, Sweden, Finland, and Iceland
Eastern Europe:	All other European countries

Appendix 2: Codebook

This codebook contains the questionnaire as fielded in the DASISH Data Archive Services Survey.

- The variable names are printed in bold and correspond to the names in the dataset.
- The questionnaire routing is printed in italics for each variable concerned.
- open: answer box; no limit to the length of the answer
- string: answer box that accommodates a maximum amount of characters (255 is standard)
- empty: question may be left unanswered
- With numerical variables, wherever the range within which the respondent could choose an answer was not visible to the respondent, this is printed in italics in the codebook. Wherever no limits applied to the range within which to choose an answer, this is indicated in the codebook as 'integer'.
- The so-called 'fills' (variable text) are indicated between straight brackets [].
- Variables between curly brackets {} are not part of the database, but the associated questions or texts were part of the questionnaire.

nohouse
Administrative number of respondent

intro
Data Archive Services: ***a questionnaire for organizations within the Social Sciences and Humanities***
DASISH (Data Service Infrastructure for the Social Sciences and Humanities) is a European cluster project that brings together all 5 ESFRI research infrastructure initiatives in the social sciences and humanities (SSH). These initiatives are CESSDA, CLARIN, DARIAH, ESS, and SHARE. This survey is designed as part of the DASISH Work package 'Data Archiving'.

Aim
With the survey, we hope to gain insights about the organization and general state of Data Archive Services in the European SSH research infrastructures. The results of this survey will be beneficial to policy advisors and SSH researchers - in their roles of both data depositors and users - as much as to Data Archive Services themselves, especially if they are looking to further develop their offer of trustworthy preservation services. In the end, the survey results will be an instrument in the development of better services for the social sciences and the humanities.

Who should participate in this survey?
The survey has been designed to be answered primarily by Data Archive Services with a scope suited to one of the five SSH ESFRIs. A data archive service is any service which accepts research data and accompanying documentation for the purpose of curating and preserving this data for a given time span, and to make it available for access and re-use by a user community.

Results

A generalized report of the results of the survey will be published online on the public part of the DASISH-website and by doing so will be made available to the SSH community.

If you have indicated at the end of the survey to be interested in further cooperation, we may contact you in the future for additional input into the DASISH-project.

See http://dasish.eu/links/ for an explanation of these acronyms
See http://dasish.eu/activities/, paragraph ‘Data Archiving’

Contact

Organizational Context

To start, please provide the following information.

Organization

Organization: *string*

Country

Country: *string*

Name

Your name: *string*

Email

E-mail address: *string*

q1

To which of the following ESFRI infrastructures is the scope of your organization suited to? Please tick all that apply.

q1_1_: CESSDA

q1_2_: CLARIN

q1_3_: DARIAH

q1_4_: ESS

q1_5_: SHARE

q1_6_: None of the above

0 no

1 yes

q2a

Does your organization offer a data archive service?

1 yes, namely: **q2ayes** *string*

2 no

if q2a=2

q2b
If not, is your organization preparing to launch a data archive service?
1 yes
2 no

if q2b=1
q2byes
If so, what is the planned launch date? *string*

if (not (q2a=1 or q2b=1)) then to 'nodataservice', otherwise continue:

q3
Please give an indication of the source of funding for your organization. Please tick all that apply.
q3_1_: public funding
q3_2_: third-party funding
q3_3_: revenues, e.g. pricing for data service offerings
q3_4_: membership fees
q3_5_: public-private partnership
q3_6_: other, namely: **q3other** *string*
0 no
1 yes

q4
Does your organization provide any other services in addition to data archiving?
1 yes
2 no

if q4=1
q4yes
If so, please describe them briefly. *open*

q5
What is the primary designated community of the data archive service?
1 social scientists
2 linguists
3 historians
4 other communities, namely: **q5other** *string*

q6
Does the data archive service have a mission statement?

1 yes
2 no
if q6=1
q6yes
If yes, please provide the URL where the mission statement can be found.
For instance, for GESIS the mission statement is available at:
http://www.gesis.org/en/institute/the-association/mission/
string, empty

q7
Does the data archive have a Licence Agreement or a Depositor Agreement that data depositors are required to sign?
1 yes
2 no

if q7=1
q7yes
If yes, please provide the URL where the License or Depositor Agreement can be found.
For instance: The UKDA Licence Agreement is available at:
http://ukdataservice.ac.uk/media/28102/licenceform.pdf
string, empty

q8
Does your data archive service have a Usage Agreement or Code of Conduct to be signed by users of the archive?
1 yes
2 no

if q8=1
q8yes
If yes, please provide the URL where the Usage Agreement or Code of Conduct can be found.
For instance for DANS the General Conditions of Use are available at:
http://www.dans.knaw.nl/sites/default/files/file/archief/DANS_General_Conditions.pdf
string, empty

q9
Does the data archive have a preservation policy?
1 yes
2 no

if q9=1

q9yes

If yes, please provide the URL where the preservation policy can be found.
For instance for the UKDA the preservation policy is available at:
http://data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf
string, empty

q10

What rights does the data archive obtain with regard to the data it ingests?

1 The archive becomes the exclusive rights holder of the data
2 The archive is granted non-exclusive rights to make copies of the data and to modify data in any way necessary for the purpose of digital preservation
3 The depositor (or his/her employer) retains all rights to the data
4 Other, please explain: **q10other** *string*
5 No rights obtained

q11

Deposit and Ingest

Which file formats are accepted and curated by the Data Archive Service?

1 The archive accepts data and documentation in any format.
2 The archive accepts only formats that are included in a list of accepted formats. If so, please provide the URL where this list can be found. **q11formats** *string, empty*
3 The archive has a list of preferred formats, but will also accept other formats. If so, please provide the URL where this list can be found. **q11formats2** *string,empty*
4 Other, please define: **q11other** *string*

q12

What are the top three file formats curated by the data archive during the last three years?
open, empty

q14

Which general metadata standards are primarily used by the data archive? Please tick all that apply.

q14_1_:CMDI
q14_2_:DC
q14_3_:DDI
q14_4_:EAD
q14_5_:IMDI
q14_6_:MARC
q14_7_:MODS
q14_8_:TEI
q14_9_: Other, namely: **q14other** *string*
q14_10_: No metadata standards are used

0 no

1 yes

q14b

Are discipline-specific metadata standards used by the data archive? For instance CSDGM (for geographic data), MPEG-7 (for multimedia information), and MEI (for music).

1 yes

2 no

if q14b=1

q14byes

If yes, please name which. *open*

q15

How do depositors submit their data to the data archive service? Please tick all that apply.

q15_1_: Upload via an online deposit form

q15_2_: Send data via email

q15_3_: Send data on a CD/DVD

q15_4_: Other, namely: **q15other** *string*

0 no

1 yes

q16

Are there any costs involved for the depositor in depositing the data?

1 yes

2 no

if q16=1

q16yes

If yes, please indicate the amount of cost involved. Please indicate which currency you use. *open*

q17

Archival Storage and Preservation

What is the size of the current archive in terabytes (1 terabyte=1000 gigabytes)?

String

q18

What is the size of the current archive in terms of the number of datasets and the number of data files? For instance: DANS-EASY consists of 25,000 datasets and 2,000,000 data files.

string

q19

Does the data archive have a maximum deposit size?

1 yes

2 no

if q19=1

q19yes

If yes, please give the size in gigabytes (1 gigabyte=1000 megabytes).

q20

How long does your data archive aim to preserve deposited research data for?

1 Indefinitely

2 Up to a maximum number of years, namely: (please indicate the number of years) **q20years** *string*

3 Other, namely: **q20other** *string*

q21

What preservation strategies are employed to ensure that the preserved data can be used again? Please tick all that apply.

q21_1_: Migration

q21_2_: Emulation

q21_3_: Bitstream preservation

q21_4_: Other

0 no

1 yes

q22

Has the data archive undertaken any activities to determine its trustworthiness?

1 yes

2 no

if q22=1

q22yes

If yes, please indicate the level of trust that has been reached:

1 A self-assessment, e.g. DRAMBORA, DSA, or the Nestor catalog of criteria

2 A peer-reviewed DSA-self-assessment

3 A peer-reviewed ISO 16363 or DIN 31644 self-assessment

4 Full conformance (with external audit) to ISO 16363 or DIN 31644

5 Other, namely: **q22yesother** *string*

q23

What measures are employed to protect the authenticity and integrity of the data stored in the data archive? Please tick all that apply.
q23_1_: Access controls
q23_2_: Check sums
q23_3_: Audit trails
q23_4_: Other, namely: **q23other** *string*
q23_5_: No measures
0 no
1 yes

q24
Dissemination
Through which channels or via which platforms can the data be accessed?
For instance: via a website; an online catalogue; a special interface like DANS-EASY
open

q25
Are there any costs involved for the user in accessing the data?
1 yes
2 no
if q25=1
q25yes
If yes, please indicate the amount of cost involved. Please indicate which currency you use.
string

q26
Are there conditions involved in accessing the data?
For instance: Access only free to Academia, access only after registration, depositor defined access.
1 yes
2 no

q27
Future development
In the previous parts of this questionnaire, questions were asked about different aspects of the data archive service. We would like to know whether you are satisfied with the maturity level of these aspects. If you are not satisfied, we would like to know if you have any concrete plans to improve the situation.

Are any measures needed to improve the Data Archive Service on any of the following aspects? Please tick all that apply. If yes, please elaborate what improvements you plan to make.
q27_1_: Data archive administration **q271yes** *string*

q27_2_: Ingest facilities **q272yes** *string*
q27_3_: Archival storage and preservation **q273yes** *string*
q27_4_: Dissemination facilities **q274yes** *string*
q27_5_: Other measures, namely: **q27other** *string*
q27_6_: No measures needed
0 no
1 yes

q28
Would you be interested in cooperating further with European Research Infrastructures like CESSDA, CLARIN, DARIAH, ESS, or SHARE to discuss options for improving your Data Archive Service in the areas indicated in the previous question?
1 yes
2 no

if (not (q2a=1 or q2b=1))
nodataservice
This questionnaire is only intended for organizations that already provide data archive services or are planning to do so in the near future.
Thank you for your co-operation!

if q2a=1 or q2b=1:
opm
Thank you for your co-operation!

Do you have any comments on the questionnaire?
1 yes
2 no

if opm=1
evaopm
Please insert your comment here below.
open

DatumB	Date start questionnaire
TijdB	Time start questionnaire
DatumE	Date end questionnaire
TijdE	Time end questionnaire

Appendix 3: Questions for interviews

A. Procedure

The people working in DASISH WP4 have already a good notice of policies and procedures within the major European DASs. Therefore it is decided to conduct interviews with a limited number (six) promising DASs. Promising in this respect is related to the fact that the DAS already exists or has in our views a good concept to develop into a DAS of high quality. The selection of the six candidates has been made using the answers given in the survey, introducing a spread over the three ESFRIs CLARIN, CESSDA, and DARIAH. Most important selection criteria are: availability of a mission statement, of deposit and user licences, of trust-related activities, of dissemination activities, and of rights managements.

This approach has led to the selection of the following DASs:

BAS and UFAL/LINDAT within the ESFRI CLARIN
TextGrid and DRI within the ESFRI DARIAH
DDA and ADF within the ESFRI CESSDA.

Within WP4.3 we have decided to focus the interview questions on topics that have not been addressed in the survey. We agreed to ask 10 questions to every interviewee and to have the possibility to ask 4 additional questions, depending on the answers given.

Interviews have been conducted by Bamba Dione (UFAL/LINDAT and DRI), Birger Jerlehag (DDA and ADF), Claudia Engelhardt (TextGrid), and Arjan Hogenaar (BAS).

B. Interview questions

1) Name interviewee
2) Function
3) Organisation
4) Name data archive
5) Country
6) ESFRI
7) Telephone number
8) Email address
9) URL of the data archive
10) Personal URL

Core questions (to be asked in every single interview)

11) a. What is the technical background of your data archive and how was the decision for the specific technical setting used by your archive made?
    b. Technology is changing rapidly. How are the technical developments evaluated?
    How often do these evaluations take place and how does this influence the technical backbone of the data archive?

12) Which authorisation/authentication tools and methods do you currently use, both in ingesting and in accessing datasets in your archive? Are you satisfied with these?

13) In the survey you indicated your primary designated community. To what extent do you know this designated community? Which problems do you have in the defining your designated community? Do you have contact (on a regular basis) with members of this community to discuss their needs regarding the current and future (re-) use of data in the archive?

14) Describe the data archive's training and outreach activities (e.g. organizing workshops, courses for researchers, universities and so on in order to promote proper data management). Do you think you could benefit from support/cooperation in the development and maintenance of these training and outreach activities? If so, of what kind of support/cooperation?

15) By which measures does the archive support its sustainability on an organizational level? How has the preservation strategy been set up, what are the arrangements with third parties for storage and what kind of measures have been taken to ensure confidentiality of data? Which benefits do you see in the starting (or continuation) of risk management procedures like DRAMBORA and certification trajectories like DSA, DIN 31644 or ISO 16363 (and the documentation of all procedures within your archive related to certification)

16) a. Has your archive/data service developed a comprehensive policy framework to govern the total of procedures and guidelines for data management, data archiving, and sharing of research data? And do ethical guidelines fall within this framework and, if so, how have these been implemented?
    b. Which routine procedures do you follow in the archive to protect sensitive data?

17) What is your need for large-scale interconnection of European data archives (for instance in a common access portal)? Could national or international cooperation help your archive to realise economy of scale effects?

18) a. Do you require the presence of a minimum set of metadata in the ingest process? And is the depositor free in choosing her metadata standard?
    b. What is your opinion on the adoption or development of common metadata standards to facilitate the searching of various data archives?

19) a. How do you deal with different legal/copyright aspects of the archived data?
Our survey showed that all kinds of copyright transfer agreements exist in the different archives. An example is the fact that it is difficult for a data archive service to realise sustainability when it does not have (enough) rights to migrate or emulate the data in its archive?
b. How does the archive handle different user licences for the deposited datasets?
c. Which possibilities do you have as an archive to negotiate with the depositor(s) in case of unclarity?

20) Other suggestions/remarks?

Additional questions: to be asked depending on the answers given by the interviewee

21) Sustainability of a data archive is dependent on many factors. We have already discussed the technical aspects. In an economic sense, sustainability is related to the financial viability of the archive. How do you guarantee this financial viability and what is the business model for your archive?

22) Which PID system do you use and why?

23) One of the objectives of DASISH is promoting of convergence of data archive services. Every discipline (and often every archive) has its own tools to facilitate retrievability of datasets. Apart from free-text searching, classification may be used to retrieve datasets within a specific field. An overview of classification codes in use within Europe may be helpful in realising convergence, as this could be the start for concordance activities. Does your data archive use special (classification) codes for retrieval and are you open for the idea of classification concordance?

24) In a data archive service, it is expected that its primary expertise lays in the field of archiving and giving access to datasets. The actual physical storage of data is perhaps a task that may be performed by a third party. How do you think about outsourcing the physical storage of data? In your opinion, what would be the advantages and disadvantages of having a physical back-up service on a European scale?

Appendix 4: Descriptions of the interviews

A. BAS interview

Interviewer: Arjan Hogenaar (DANS)
Date of interview: July 8th, 2014
Duration of interview: 1 hour (10.00 – 11.00 h, CET).

1) Name interviewee: Dr. Florian Schiel
2) Function: Co-director BAS
3) Organisation: University of Munich
4) Name data archive: Bavarian Archive for Speech Signals
5) Country: Germany
6) ESFRI: CLARIN and SHARE
7) Telephone number: +49 89 2180 2758
8) Email address: bas@bas.uni-muenchen.de
9) URL of the data archive: http://www.bas.uni-muenchen.de/bas
10) Personal URL: https://www.phonetik.uni-muenchen.de/institut/mitarbeiter/schiel/Schiel.html

Answers on Core questions

11a. After having studied many available systems (like Fedora) the decision was made to develop own software within BAS, based on PERL, PHP, Shibboleth, and OAI-PMH libraries for PERL. The reason for this is that systems like FEDORA did not offer an optimal search system at the time of setting up BAS. Besides, FEDORA did not offer a version control system. Finally, integrating Shibboleth with FEDORA appeared to be difficult.

11b. BAS has to follow technical developments, for instance in Shibboleth, in order to adapt the own software (example: authentication tools for web services).

12. BAS is satisfied with AAI/Shibboleth. Disadvantage is that AAI only works for Academia. BAS, though, has also a lot of costumers from outside Academia (private persons and commercial companies). Authentications of this type of users has been realised using CLARIN IdP (CLARIN Identity Provider).

13. Linguists are the primary designated community. For the academic part of it, AAI is being used for identification. There are more and more users coming from other disciplines (psychologists; anthropologists; neuroscientists and so on). Needs of special user groups are discussed, for instance, the need amongst anthropologists for speech processing.

14. BAS organises an annual workshop for PhDs and post-docs. The newest adaptations are presented and the attendees are asked for giving feedback. Besides, BAS publishes White Books on its website, describing the different procedures in use (for data collection; data annotation and so on).

Cooperation regarding training takes place within CLARIN. Aside from international summer schools BAS has no experience with trans- or interdisciplinary cooperation for training.

15. BAS has chosen to set up its own sustainability model. Finance is coming from Munich University. This university is paying two permanent positions. For a fee, commercial users may access the datasets deposited at BAS; fees are used for maintaining the archive. BAS stores its backup datasets at the Leibniz-Rechenzentrum (part of Max Planck Gesellschaft). As a CLARIN-B centre, BAS has obtained both the DSA-seal and the CLARIN B assessment procedure. BAS has no experience with DIN 31644 or ISO 16363.

16a Most information on this subject is described in internal documents. Some parts of these have been published. A well-defined statement regarding acceptance of datasets offered to BAS is published in https://www.phonetik.uni-muenchen.de/Bas/BasPolicyExternalResources_eng.pdf. Research data in a bad shape will not be accepted. BAS has also strict validation procedures documented in http://www.bas.uni-muenchen.de/forschung/BITS/TP2/Cookbook/. Only data with written consent of the data provider will be accepted.

16b. BAS has datasets with sensitive data. These are only accessible for certain persons/organisation (agreements made during the ingest process).

17. For BAS the CLARIN ERIC is sufficient. Via CLARIN federated content searching will be made possible. BAS is not in favour of setting up a central archive.

18a. BAS requires a minimum set of metadata elements (gender; age; region; language). Without these elements, datasets will not be accepted. BAS prefers CMDI, but metadata in OLAC or DC will also be accepted (and these will be converted into CMDI later on).

18b. CMDI is also used for searching the (metadata) of the datasets. The CMDI metadata are harvestable via OAI-PMH.

19. Copyright is still a minor issue at BAS. Recommendations to CLARIN centres are made by the Institut für Deutsche Sprache IDS, Mannheim. In case of transfer of a dataset to a different university, a legal document will always be composed. There are no standard contracts. Only the University Administration is allowed to sign the legal documents. For usage by individual researchers, there is a standard document.

BAS has two use licences:

a. terms of usage of data and annotations (downloading data)
b. terms of usage of web services.

20. No Other suggestions/remarks?

Additional questions:

21. BAS uses Handle/EPIC by the GWDG (Gesellschaft für wissenschaftliche Datenverarbeitung) as its PID-system.

22. BAS has its own classification system. Users are not required to use it, though.

23. BAS has no need for a European physical back-up service.

24. The relationship with SHARE consists of ELRA (European Language Resources Association). ELRA - part of the SHARE infrastructure-uses the BAS catalogue-data

B. Interview with LINDAT/UFAL

1. Name interviewee: Jozef Mišutka
2. Function: CTO of LINDAT/CLARIN, repository manager
3. Organisation: Institute of Formal and Applied Linguistics
4. Name data archive: LINDAT/CLARIN digital library
5. Country: Czech Republic
6. ESFRI: CLARIN
7. Telephone number: + 420 221 914 278
8. E-mail: misutka@ufal.mff.cuni.cz
9. URL of the DAS: https://lindat.mff.cuni.cz/repository/xmlui
10. Personal URL: https://ufal.mff.cuni.cs/jozef-misutka

1) *a. What is the technical background of your data archive and how was the decision for the specific technical setting used by your archive made?*

The LINDAT technical infrastructure consists of two servers and a virtualized operating system. The platform used is based on Ubuntu Long-Term Support (LTS). Accordingly, new critical updates and security patches will be available for a couple of years. Packages are updated or upgraded regularly (for instance, there may be regular updates every month).

The technical setting is made by the administration department, which decided to go for Ubuntu. There is no particular reason for choosing this Linux distribution; other distributions could have been chosen as well. The crucial point when setting up the infrastructure is the fundamental requirement to have a system that is available 24 hours, 7 days a week. Thus, the decision for the technical setting was made based on this requirement. Accordingly, the virtualized operating system is used to migrate between these servers, making sure that if one server fails, the other server will automatically take over.

Concerning new technical developments, the servers used in LINDAT are quite stable, and include clusters for performing high-computing operations. The clusters are also kept up to date regularly (e.g. every month) by the administration department.

b. Technology is changing rapidly. How are the technical developments evaluated? How often do these evaluations take place and how does this influence the technical backbone of the data archive?

The repository is run on the servers mentioned above. The performance of the repository is closely monitored using standard monitoring services. In addition, there are other services running on the servers, which are automatically checked. For instance, a qualitative assurance framework is used with automatic programs embedded in a monitoring system like Nagios to keep track of both the availability and the performance of the servers.

System evaluation is performed only when real performance problems arise, and to date no such problems have occurred. Thus, so far, there was no reason for evaluating the system. However, in the event that problems were to arise, a monitoring service with special software programs would be used to perform automatic checking of the system properties. Such a check could be done on a regular basis (e.g. every five minutes) depending on the nature of the check. Currently, several checks are being performed; for instance for current running processes, free disk space, the time needed to receive an answer when pinging a server or getting a webpage, etc. These kinds of checks are usually run every five minutes. If there is a specific problem, the administrator is notified via email.

Should problems arise in this area, the evaluation process could definitely influence the technical backbone of the repository. However, at this moment, the servers haven't experienced any real problem.

2) ***Which authorisation/authentication tools and methods do you currently use, both in ingesting and in accessing datasets in your archive?***

As far as data ingest is concerned, LINDAT allows any user to submit data to its repository, provided this user authenticates. After authentication, the user can submit his/her data at one of two different places, the first place is for the LINDAT community, and the second one for outsiders. Authenticated users can choose to which of these places they want to submit their data.

Shibboleth is used as authentication tool. In addition, LINDAT has a local user-based authentication system. Concerning Shibboleth, LINDAT is part of both the Czech national federation, and international federations like the EDUcation Global Authentication INfrastructure (eduGAIN) and CLARIN Service Provider Federation (SPF). Also, LINDAT is harvesting data from Homeless Identity Providers (IdPs). These constitute the main sources for user authentication.

According to Contact, Shibboleth is the best solution regarding authentication of academic users. Non-academic users can also be handled by the system used at LINDAT, although this may be more complicated than providing authentication for academic users, which are the main target for the repository. The status of a user as an academic one is determined by an authority, i.e. the federations LINDAT is part of (the national federation, eduGAIN, CLARIN SPF).

Users without an IdP have two options: 1) they can go to the Homeless IdP of the CLARIN project; or 2) they can apply to open an account to the LINDAT repository. In any case, LINDAT has to find a way to check the user's identity, e.g. if (s)he is an academic user. Such identity does not need to be a real name, an email, a date of birth, etc. This means that identification can be in an anonymized form: A user is identified when (s)he signs an electronic license. This is the only requirement for authentication, allowing the repository for being able to keep track of the users.

Regarding access to the datasets, LINDAT encourages people to use Creative Commons (CC) licenses. Data licensed under CC are accessible without authentication (except for resources under CC-BY-NC-SA license). However, LINDAT has other resources which have specific licenses attached to them. Before accessing/downloading these resources, the user has to authenticate and sign the specific license. In this context, signing the license means that (s)he reads and accepts the terms of the license by clicking a button labeled "I AGREE". During this process, the repository records the user's information needed for the LINDAT licensing system. This information will include among other things the user ID, the license type, the resource (s)he wants to download, the time, etc. The repository also informs the user that these data are being recorded. When (s)he agrees to the terms of the license and provides any extra information required by the LINDAT licensing system, only then (s)he would be able to download the data.

LINDAT is promoting usage of Creative Commons-like licenses. Data depositors are thus encouraged to use terms of the license that are similar to the CC attributions, allowing for public access and use of their data.

In addition, some other items in the LINDAT repository are embargoed, meaning that they will be available after some date. After this date, the items can be accessed in the usual way (authenticated users who have accepted and signed the license attached to these items will be able to download them, as described above).

Furthermore, LINDAT has a **local** authorization system, i.e. authorization information is stored on the server. Thus, in order to set the authorization rights, users have to log in to the local system (i.e. the repository) and perform the setting there. Authorization settings are modified mostly by the administrators of the repository in the repository itself.

Contact thinks that it is a great idea to have a single authorization platform that allows people in one country to give access rights to some other users in another location. Currently, there are authorization platforms called virtual organizations. For instance, in international projects like CLARIN, there are many Centres/service providers. If a user has a resource in all of these repositories, (s)he would have to go to all of these repositories, to log in there and then create/update the authorization on these servers. The theory with these international projects is that there will be a single authorization infrastructure where users can log in and set the authorization. Then, the repository could get this authorization information from this infrastructure.

However, at the present time, LINDAT does not make use of a pan European authorization system. One main reason for not using such a system is that there are technical problems related to user identification. More specifically, authentication on a pan European level is currently facing problems related to attribute releasing: (i) it is difficult to obtain the necessary attributes to identify the user and (ii) it is even more difficult to prevent one real user to be identified differently in different systems.

Hence, an authorization system at a pan European level may work. However, if LINDAT would try to use any of the systems currently used by the international projects mentioned above, its user base would shrink, and this would be problematic, as one of the main objectives of this repository is to have as many users as possible. Also, it is not possible for the LINDAT to use such authorization platforms for all its users, because the repository uses methods/attributes for identification that are different from those the virtual organizations are using.

Are you satisfied with these?

Yes, people at LINDAT are completely satisfied with the authentication tools they are using. According to Contact, federated systems like CLARIN SPF and eduGAIN, Shibboleth and IdPs are the best solutions currently available for academic users.

3) ***In the survey you indicated your primary designated community. To what extent do you know this designated community? Which problems do you have in the defining of your designated community? Do you have contact (on a regular basis) with members of this community to discuss their needs regarding the current and future (re-) use of data in the archive?***

The primary designated community for LINDAT consists of two main groups. The first group is actively working on creating the resources, putting them to the repository, and so on. In contrast, the activity of the second group mainly consists in using the resources stored in the repository.

LINDAT knows the first group very well: the repository organizes regular meetings and workshops with this group to discuss their needs. Also, LINDAT meets with this group in many (linguistics) conferences. Because these conferences are not project specific, they bring together most of the people involved in the CLARIN project and people outside of the CLARIN community. Through these meetings, conferences and workshops LINDAT will seek to get a grasp of what's going on, to understand the reality and the concrete problems of this group and to get an idea on how these problems can be solved / are being solved by other infrastructures.

LINDAT is also in contact with the second group, but knows it to a lesser extent (in comparison to the first group). For instance, through research conferences, LINDAT has an opportunity to meet with this group. Besides the conferences, LINDAT organizes workshops on different topics several times a year. The workshops are designed for specific users (e.g. for the outside community and mostly those people who are really using the infrastructure) in order to discuss their current needs. One of these workshops is planned for the fall 2014 and will be organized at the local institution.

Within the CLARIN community, there might be different concepts on how to do things, but LINDAT has not really encountered any problems in the defining of its designated community.

4) ***Describe the data archive's training and outreach activities (e.g. organizing workshops, courses for researchers, universities and so on in order to promote proper data management).***

The workshops mentioned above may also include training on data management. However, LINDAT is not regularly organizing data management workshops. One of the objectives of LINDAT is to design the repository so that this kind of training and outreach activities will not necessarily be needed. For instance, much work has been done to support the data ingesting process. This consists for example in providing the most relevant features a user expects and avoiding those features that would annoy him/her. Thus, the data ingest procedure is designed in such a way that a user who has never used the infrastructure before will find the process quite intuitive.

Once the data are submitted, only minor changes (e.g. the correction of spelling mistakes) are allowed; the user is not allowed to change the data itself. With respect to data management, it is possible to have new versions, some additional changes, but the user will not be able to concretely "touch" the data. The user has the possibility to delete the data, if (s)he wants. However, the deletion of data cannot be done automatically, but has to go through an administrative process. For instance, the user will have to send a request (e.g. an email) to the repository.

Do you think you could benefit from support/cooperation in the development and maintenance of these training and outreach activities? If so, what kind of support/cooperation?

Contact thinks that the development and maintenance of training and outreach activities with the aim to collaborate and to show different approaches are very welcome. Also, people at LINDAT would certainly be happy to take part in such cooperation. However, Contact does not think they need support at the moment. He is also not sure to be able to give the exact kind of support LINDAT would take advantage of.

5) ***By which measures does the archive support its sustainability on an organizational level? How has the preservation strategy been set up, what are the arrangements with third parties for storage and what kind of measures have been taken to ensure confidentiality of data?***

Sustainability can be considered at least on two different but interrelated levels: technical vs. organizational.

At a technical level, LINDAT supports its sustainability by performing regular data backups at local level and by replicating data to external servers located at different places. For data storage, LINDAT has agreements with the Czech Academic Network (CESNET) and CINES in France. The data preserved by the third parties are just stored there, but are not publicly available. In the agreement, the third parties are engaged to preserve the data safely and in multiple locations. In case they stop being operational, they are required to notify LINDAT about this a certain amount of time before, so that the repository can make arrangement for the transition with some other organizations.

This data replication approach used by LINDAT ensures that the records will still be safely preserved[14] in other places, in the event a natural disaster (e.g. fire) would occur at the Charles University in Prague, where LINDAT is hosted. For instance, the data replicated in France are being replicated in many different cities across Europe. Thus, from this perspective, LINDAT can be considered to be really sustainable at the technical level.

Concerning the organizational level, LINDAT has obtained the Data Seal of Approval (DSA). In addition, the Charles University has been funded for many centuries now and people are

[14] The data are preserved on a bit level.

quite confident that the University will continue to exist for a very long time in the future. However, in case there would be no funding at all for the University, one of the priorities in LINDAT has been to create a repository that is sustainable at almost zero cost (there would only be costs for the hardware). Accordingly, the infrastructure has foreseen cases in which all the developers and curators would leave the project. Thus, the system has been implemented in such a way that a non-expert user (someone who has never used the infrastructure) will be able to take over the repository management in these cases. Also, LINDAT has performed tests successfully aimed to ensure the success of such a project.

In case there will be no funding for the system hardware, sustainability will be supported by the fact that the data replication approach used by LINDAT is performed in a standard way that allows other CLARIN Centres to easily take over the records stored in the repository. The only problem that may arise will concern handling the legal issues. Before being able to take over these records, the Center would be required to have a licensing framework similar to the one used by LINDAT (the metadata are freely available).[15] For all the data included in the LINDAT repository, the depositors sign a Deposition License Agreement, which states that the repository can replicate the data outside, and LINDAT has to adhere to the licensing agreement.

Furthermore, the implementation of the LINDAT repository is publicly available. Thus, anyone should be able to acquire a server, install this repository and harvest all the public data (harvesting non-public data requires first to resolve the legal issues, e.g. sign agreements, ensure copyright compliance, etc.).

Regarding confidentiality of the data, LINDAT has the only right to manage and work with the data. The third parties are used only for storing the data, but not for making them available.

Which benefits do you see in the starting (or continuation) of risk management procedures like DRAMBORA and certification trajectories like DSA, DIN 31644 or ISO 16363 (and the documentation of all procedures within your archive related to certification)?

LINDAT has obtained the DSA, which is a requirement for all CLARIN Centres. However, before being DSA certified, the repository already had everything in place. Thus, with DSA there were no changes. The only impact / main advantage of using DSA is that it allows for improving the documentation. Contact is not sure if the use of other risk management procedures would present a potential benefit, as people at LINDAT feel really secure at the moment.

6) ***a. Has your archive/data service developed a comprehensive policy framework to govern the total of procedures and guidelines for data management, data archiving, and sharing of research data?***

[15] There are no official agreements between CLARIN Centres on this.

Yes, LINDAT has developed a policy framework[16] to govern the procedures and guidelines for data management, data archiving, and sharing of research data. This framework can be found at the LINDAT webpage and includes the following policies:

- Terms of Serviceshttps://lindat.mff.cuni.cz/repository/xmlui/page/about - terms-of-service
- License Agreement and Contractshttps://lindat.mff.cuni.cz/repository/xmlui/page/about - about-contracts
- Intellectual Property Rightshttps://lindat.mff.cuni.cz/repository/xmlui/page/about - about-ipr
- Privacy Policyhttps://lindat.mff.cuni.cz/privacypolicy.html
- Metadata Policyhttps://lindat.mff.cuni.cz/repository/xmlui/page/about - metadata-policy
- Preservation Policyhttps://lindat.mff.cuni.cz/repository/xmlui/page/about - preservation-policy
- Citing Data Policyhttps://lindat.mff.cuni.cz/repository/xmlui/page/about - citing-data-policy
  https://lindat.mff.cuni.cz/repository/xmlui/page/about - citing-data-policy
  https://lindat.mff.cuni.cz/repository/xmlui/page/about - citing-data-policy

And do ethical guidelines fall within this framework and, if so, how have these been implemented?

No. Ethical guidelines are responsibility of the depositor, and therefore not included in this framework. Thus, LINDAT is not performing manual inspection of the deposited data with respect to ethical issues. One of the reasons for not performing a manual check of the data is that the repository would have the liability for this, according to the Data Protection Direction of the European Union.

For deposited data, the metadata as well as the format and the availability of data (not the data themselves) are checked automatically. In case a depositor submits data that do not respect the ethical norms, the repository has his/her ID and can take countermeasure afterwards. All the users of the repository come from the academic world and they know they can be tracked down.

Also, the deposit license agreement, which the user signs when submitting his/her data to LINDAT, specifies that the depositor is responsible for the content (s)he submits. Finally, the repository has a complex data curation process regarding the metadata, but not for the data itself.

b. *Which routine procedures do you follow in the archive to protect sensitive data?*

LINDAT can give advices, but has not established guidelines on protecting sensitive data. There are no routine procedures (e.g. manual checking or inspection of data) to protect sensitive data. Procedures such as anonymization are the responsibility of the depositor, not of the repository. [17] LINDAT assumes that data should be anonymized by the depositor before submission.

[16] Contact is not sure if the policy framework can be defined as being comprehensive.

[17] LINDAT will anonymize data related to the user access like logs, ip addresses etc. that is completely different from the data submitted to the depository.

7) ***What is your need for large-scale interconnection of European data archives (for instance in a common access portal)?***

LINDAT is already part of several large-scale international data archives. The Virtual Language Observatory (VLO), which is part of CLARIN, is one of these. In addition, LINDAT is in the process of being registered to the linguistic meta-catalogue Open Language Archive Community (OLAC), which contains many other repositories. Also, LINDAT is included in the data index of Thomson Reuters, is part of Google Scholar, and the repository is being harvested by other projects. From this point of view, Contact is not sure if LINDAT would really benefit from a large-scale interconnection of European data archives. Of course, the more collaboration with data archives (on harvesting metadata records), the better. However, LINDAT is against the proliferation of persistent identifiers. It is not a problem that the data are stored in many places, but the most fundamental thing is that the data must always have the same persistent identifier.

According to LINDAT policies, the metadata are openly available, and there is no problem that a European data archive harvests the metadata using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which is also used by LINDAT.

Contact has nothing against an interconnection of European data archives; however, he is a bit skeptical of the benefits of this project for LINDAT. He can imagine that this could be beneficial for the European Union.

Could national or international cooperation help your archive to realise economy of scale effects?

Contact cannot answer this question. The only thing to say here is that LINDAT will definitely try to collaborate with libraries.

8) ***a. Do you require the presence of a minimum set of metadata in the ingest process?***

Yes.

And is the depositor free in choosing her metadata standard?

No. There are several metadata standards. Currently, LINDAT offers 9 metadata standards (an additional standard will be offered quite soon). In Contact's opinion, it doesn't make sense to let the user choose a specific standard, because the repository would then have to take care of the metadata conversion afterwards. Thus, LINDAT uses an abstract data ingestion process, where the user does not see which metadata standard is used to store his/her submission. (S)He only fills out some values required for the submission (e.g. title, author, keywords). These values are defined in a semantically precise way (i.e. they are not ambiguous). Then, the information entered by the user is first stored in an abstract format, and then, converted into a standard metadata scheme (e.g. Dublin Core, IMDI, etc.).

LINDAT is supporting the Component MetaData Infrastructure (CMDI) initiated by CLARIN. CMDI provides a framework to create and use self-defined metadata formats. It also allows the user to integrate existing schemas. LINDAT is also using multiple metadata profiles, Dublin Core and MetaShare, which can be exported to CMDI format using OAI-PMH.

b. What is your opinion on the adoption or development of common metadata standards to facilitate the searching of various data archives?

There are a lot of projects or meta-search engines, which require their own metadata formats. LINDAT is strongly against the proliferation of the metadata schemes. There are a lot of formats, and none is the best one.

However, Contact is very skeptical to the fact that there will be a common metadata standard in the near future. If someone chooses one format and says this is the best one, then he will fail or will not be able to convert it to other projects, because there are lots of different formats. Previous experiences with projects like MetaShare and Dublin Core have proved that this kind of approach is not going to work or is not really possible.

According to Contact, users should take one of the standards that are already available, and then define the minimum set of metadata in a way they will be able to export it to different formats. This is one of the reasons of using CMDI, which may have its own disadvantages. However, CMDI allows users to create different formats, and the necessity for having this infrastructure comes from the fact that there were data submitters who, for some reason, tried to create their own format, because they haven't found a suitable one. The experience in LINDAT is that there will always be authors who will deposit their work and who do not understand that it is a really good idea to have a one uniform metadata standard. Such depositors will try to put their own format, leading to a continuous proliferation of the metadata schemes.

A realistic approach with respect to this problem would be, for instance, to require those who want to be part of a specific project, to support the metadata standard established by this project. Such an approach is adopted projects like OLAC, which requires those who want to be part of its meta-search engine to support its metadata format. Those who do not want to be included, do no need to support this format. This is an approach that the European Union can also adopt. In Contact's opinion, forcing one metadata format will be doomed to failure.

9) ***a. How do you deal with different legal/copyright aspects of the archived data?***

LINDAT has its own policy framework, where licenses and the related attributes are defined. For each of the resources submitted, the depositor can attach a specific license to it, and the resource will be protected by this license. It can be one of the licenses supported by LINDAT. If the license is not supported, the repository can create a license for the user. However, LINDAT recommends using the Creative Commons licenses.

If somebody breaches the copyright license, he can be identified via the information he provides during the authentication process. So far, LINDAT hasn't experienced problems related to a breach of copyright.

Our survey showed that all kinds of copyright transfer agreements exist in the different archives. An example is the fact that it is difficult for a data archive service to realise

sustainability when it does not have (enough) rights to migrate or emulate the data in its archive?

The deposition license agreement specifies the conditions under which users submit their items to the LINDAT repository. These data can be replicated under the conditions specified in the deposition agreement. The agreement can be found at the LINDAT webpage[18]. During the submission process, the depositor should read the agreement, and accept the conditions specified in this agreement before being able to submit his/her data.

The migration of data is performed differently according to the type of license attached to the resources. Migration for all public data can be done without any constraints. For protected resources (non-public data like academic or restricted data), however, there are specific circumstances under which these resource should be migrated. By signing the deposit license agreement, the user grants the rights to the repository to perform data replication. One of the conditions specified in this agreement is that the depositor agrees that the repository keeps more than one copy of his submission for purposes of security, back-up and preservation. Thus, the repository has the right to migrate, but does not use it for protected resources unless specific conditions are met. For instance, LINDAT would not migrate a non-public resource to another repository, unless this repository would conform to the LINDAT standards of the licensing.

b. ***How does the archive handle different user licenses for the deposited datasets?***

See the licensing framework mentioned above.

c. ***Which possibilities do you have as an archive to negotiate with the depositor(s) in case of unclarity?***

LINDAT has not encountered problems related to unclarity so far. In the event such problems arise, LINDAT has established a framework or curation process and clear means on how to find the liable person. However, the procedure does not involve downloading or manually checking the data submitted. The repository disclaims any liability or responsibility for depositing contents which breaches copyright. If a user submits an item, for which (s)he does not have the permission, (s)he will be held responsible. (S)He will be tracked down on the basis of the information entered during authentication. This will be a fairly straightforward process, as only users from academic institutions can authenticate and the repository has IDs for each user. The user who deposits an item to the repository has to explicitly check the deposit license agreement, which makes him/her liable of everything he deposits.

Furthermore, if a user cannot properly choose a license, (s)he can notify the repository (e.g. by sending an email). Then, LINDAT will provide help by selecting a license for him/her. If the user does not agree to the deposition license, (s)he can send an email to the repository which would then have to decide what to do in this case.

10) ***Other suggestions/remarks?***

[18] https://lindat.mff.cuni.cz/repository/xmlui/page/contract

No suggestion.

Additional questions: to be asked depending on the answers given by the interviewee.

11) ***Sustainability of a data archive is dependent on many factors. We have already discussed the technical aspects. In an economic sense, sustainability is related to the financial viability of the archive. How do you guarantee this financial viability and what is the business model for your archive?***

LINDAT does not have a business model that would consist in selling the resources deposited to the archives. There are different projects where the repository is being used. However, one of the most important things, as mentioned above, is to make the repository at almost zero cost from the administration part (see above for more details).

12) ***Which PID system do you use and why?***

LINDAT uses handle as persistent identifier service. The handle system has several advantages. One of the greatest benefits of using this system is the flexibility it provides. People can manage the system in the way they like. For instance, they can run their own handle server. Also, people have all the power with respect to using the prefix assigned to the infrastructure (LINDAT has acquired the prefix for ca. 50 dollars a year). Furthermore, handles can be used for partial identification, which is not supported by some other persistent identifier services like DOIs. Names can be assigned to the handle, etc.

13) ***One of the objectives of DASISH is promoting of convergence of data archive services. Every discipline (and often every archive) has its own tools to facilitate retrievability of datasets. Apart from free-text searching, classification may be used to retrieve datasets within a specific field. An overview of classification codes in use within Europe may be helpful in realising convergence, as this could be the start for concordance activities. Does your data archive use special (classification) codes for retrieval and are you open for the idea of classification concordance?***

No, LINDAT is not using special (classification) codes, but rather keywords as suggestion for retrieval. People at LINDAT found that the use of keywords has greatly improved the quality of the metadata.

Regarding the idea of classification concordance, yes, people at LINDAT will definitely be interested in and will be very happy to adopt it. However, they will be first interested in hearing how this is being done and why this should be used. Next, they will be considering to looking at this and if they are convinced, they can use it.

14) ***In a data archive service, it is expected that its primary expertise lays in the field of archiving and giving access to datasets. The actual physical storage of data is perhaps a task that may be performed by a third party. How do you think about outsourcing the physical storage of data? In your opinion, what would be the advantages and disadvantages of having a physical back-up service on a European scale?***

Outsourcing the physical data storage is a good thing. A physical back-up service on a European scale would definitely help to ease the burden for LINDAT. However, some conditions may need to be fulfilled, before this can happen. First, Contact is not sure how such a service will work from the legal point of view. Second, the service has to be secure and to ensure at least the same sustainability, as is the case with the current LINDAT infrastructure.

A main disadvantage with such a service would be to only offer storage, because LINDAT is automatically performing some integrity checks and some other automatic processing of the data.

C. TextGrid

Held 21 July 2014 at Göttingen State and University Library
Interviewer: Claudia Engelhardt

1) a. Name interviewee: Sibylle Söring
2) a. Function: Consortium Manager
3) a. Organisation: Göttingen State and University Library
4) a. Name data archive: TextGrid Repository
5) a. Country: Germany
6) a. ESFRI: DARIAH
7) a. Telephone number: +49-551-39 13777
8) a. Email address: soering@sub.uni-goettingen.de
9) a. URL of the data archive: http://www.textgridrep.de/
10) a. Personal URL: ---

1) b. Name interviewee: Stefan E. Funk
2) b. Function: Developer
3) b. Organisation: Göttingen State and University Library
4) b. Name data archive: TextGrid Repository
5) b. Country: Germany
6) b. ESFRI: DARIAH
7) b. Telephone number: +49-551-39 7700
8) b. Email address: funk@sub.uni-goettingen.de
9) b. URL of the data archive: http://www.textgridrep.de/
10) b. Personal URL: ---

Core questions (to be asked in every single interview)

11) a. What is the technical background of your data archive and how was the decision for the specific technical setting used by your archive made?

Stefan: TextGrid was started in 2005. At the beginning, we did evaluate many techniques and came up at last with the SOAP protocol that we were using then. And we used some programs and protocols connected to that SOAP technique and on that basis we build the TextGrid Repository in the first place.

b. Technology is changing rapidly. How are the technical developments evaluated? How often do these evaluations take place and how does this influence the technical backbone of the data archive?

Stefan: Yes, indeed. TextGrid has been running for nearly 10 years now and the repository underwent some technical changes, for example we added REST to the protocols we are using, so every service is capable of doing REST now. In the beginning the technical background was chosen for XML file management. Many partners would use TEI and the repository is now capable of coping with XML files. And how the technical developments were evaluated, I'm not really sure … we just decided: We would need REST – and we implemented REST.

Sibylle: Not in a systematic way, but just following up the contemporary technologies. And also the search engine was changed some months ago. Now we have a faster and more powerful performance (elasticSearch).

Stefan: And the API, the interfaces stayed the same for a long time, and we try to keep that. So the underlying technology changes, but the API stays in place.

12) Which authorisation/authentication tools and methods do you currently use, both in ingesting and in accessing datasets in your archive? Are you satisfied with these?

Stefan: We are using two kinds of authorisation and authentication. One thing is OpenRBAC role based access control system. That is a system that is working with roles, which means you have for example a role "editor", and every person who has the role "editor" may create documents and read documents in one project. And the second authorization and authentication method is Shibboleth and that is going together with the DARIAH AAI in the future, so things will be simpler and quicker.

Claudia: So that means that every person with a DARIAH account would be able to log into TextGrid with their DARIAH credentials.

Stefan: Yes. Everyone who is allowed to do so, he must be put in some LDAP group.

13) In the survey you indicated your primary designated community. To what extent do you know this designated community? Which problems do you have in defining your designated community? Do you have contact (on a regular basis) with members of this community to discuss their needs regarding the current and future (re-) use of data in the archive?

Sibylle: The [definition of the] primary designated community was accompanied by a long discussion in the project about who actually is our target group, who are we working for? And there's no such thing as THE humanities, but it is a very heterogeneous group and everyone has their own needs and demands. So it is a problem to create something that fits most of the needs or to identify the mutual needs of the humanities. So yes, we do know some of them, particularly the ones we are cooperating with. And we have user meetings twice a year, which is very helpful to get to know your customer or your clients and your users. And we started with user meetings like two years ago, and this is very fruitful to find out about their needs. But there is no final point to this – as TextGrid gets new projects, there are new demands or requirements, how to explore data in the repository, how to make the data interoperable and so on.

14) Describe the data archive's training and outreach activities (e.g. organizing workshops, courses for researchers, universities and so on in order to promote proper data management). Do you think you could benefit from support/cooperation in the development and maintenance of these training and outreach activities? If so, what kind of support/cooperation?

Claudia: So these TextGrid user meetings are part of your outreach. What other outreach activities do you have? And do you also have training?

Sibylle: Yes, we are putting a lot of effort into training and workshops and so on. We give training to people who want to use TextGrid or who are evaluating their needs and the possibilities and the different options to use TextGrid, and we are training them. We are also giving workshops, for example to scientific libraries. And in addition to all this, we have mailing lists, we have a blog, we have a user message board which is not as successful and active as we would have wanted it to be, but it was a demand by the community itself. So, let's see, maybe it will start to become more active in the future.

Claudia: So you already do a lot and it sounds quite interesting. Do you think you could benefit from support or cooperation in the development and maintenance of training courses with other archives or on a higher level, for example, within DARIAH?

Sybille: Definitely. First, because the TextGrid Repository will be unified with a prospective DARIAH repository. And secondly, we at TextGrid, we can't do everything. But we often notice that researchers or scholars initially don't see why they should put so much attention to the management of their data, and to using standards and so on. And this is something another partner could well get an active role in: training and offering an established framework, so our user sees why you should do this and how you can do it.

15) By which measures does the archive support its sustainability on an organizational level? How has the preservation strategy been set up, what are the arrangements with third parties for storage and what kind of measures have been taken to ensure confidentiality of data?

Sibylle: Right now, the organizational framework is a research union which will end in 2015. Then TextGrid will be operated by an association which has been funded in 2012. It consists of the main key players who play an active role in TextGrid right now, such as several universities and academic institutions. And then there's the Gesellschaft für wissenschaftliche Datenverarbeitung (GWDG) in Göttingen, who is our partner in maintaining the service and the repository.

Claudia: If TextGrid is then maintained or sustained by this "Verein" (registered union/association), will it be independent then from project funding?

Sibylle: Yes, that's the idea. But it will be a kind of mixed funding – we are also hoping to get funding by several German ministries. So it will be a mixed funding structure.

Stefan: As for the preservation strategy: We have an arrangement at the moment with the GWDG who is hosting our data. And that will be done by the GWDG in the future too, but the scope will be the DARIAH repository. And so we will have service level agreements, most likely, that are coping with all the legal questions. And our preservation strategy at the moment: We have backups, of course, and at the moment we have bit preservation. Further preservation strategies will be coped with within the DARIAH repository, I hope.
And for the confidentiality, we have MD5 checksums to test if the data stayed the same and other measures will be then taken by the DARIAH repository in the future.

Which benefits do you see in the starting (or continuation) of risk management procedures like DRAMBORA and certification trajectories like DSA, DIN 31644 or ISO 16363 (and the documentation of all procedures within your archive related to certification)?

Stefan: We just started the process to get the Data Seal of Approval for the TextGrid Repository and I think we will get it for the DARIAH repository, too. We just started to answer the questions. I don't know exactly when we will be finished with it, but I think in the near future, we will get the DSA. And, others, DIN or ISO, are not really in the scope of TextGrid until 2015, but the GWDG is also looking into DIN and ISO.

16) a. Has your archive/data service developed a comprehensive policy framework to govern the total of procedures and guidelines for data management, data archiving, and sharing of research data?
b. Which routine procedures do you follow in the archive to protect sensitive data?

Stefan: At the moment we have documentation of the repository and the interfaces for importing and exporting data. I think in [the process of getting] the data seal of approval we will cope with these questions. There are a lot of questions in this direction that TextGrid is going to answer.

And do ethical guidelines fall within this framework and, if so, how have these been implemented?

Stefan: We have terms of use that must be accepted by every user who uses TextGrid. And they say, you are not allowed to put in personal data, or you are not allowed to put in …

Sibylle: … anything that is non-scientific data.

Stefan: Yes, and data you don't own yourself and such things. And we have got the URL to the terms of use.

Claudia: Is there sensitive data in the TextGrid archive?

Sibylle: It depends how you define sensitive data. There is no personal data. Basically it's primary research data.

Claudia: Ok, so there is no need for procedures for the protection of sensitive data?

Sibylle: No. So far, we haven't had such a case.

17) What is your need for large-scale interconnection of European data archives (for instance in a common access portal)? Could national or international cooperation help your archive to realise economy of scale effects?

Sibylle: I think it is a very interesting question, but it's very difficult to answer, because we can't really foresee the needs of our users. We are about to inquire about their demands, but we are not really sure whether they want something like an international TEI portal. I think, it would be useful to inquire a bit more about the demands and needs. It's difficult as someone who is offering these services to answer this question. So far, we experience our users to be very heterogeneous with very different needs. There are people who would embrace this idea and say: This is fantastic, this is everything we always dreamt of! But there would be others who would say: This is of no use for us.

Stefan: And of course we have the DARIAH repository. In DARIAH, there is a collection registry where collections of data, research data can be put in. And the DARIAH generic search is using the data of the collection registry and indexes it, so you can search over all collections or research data in the DARIAH collection registry. And TextGrid will be put in there too, so you can use the DARIAH generic search to search TextGrid data, too. I think that this is a first step in the direction of a European data archive.

Claudia: And is this for DARIAH-DE only or for DARIAH-EU?

Stefan: At the moment it's for DARIAH-DE, but I think it's one of the contributions to DARIAH-EU.

18) a. Do you require the presence of a minimum set of metadata in the ingest process? And is the depositor free in choosing her metadata standard?

Stefan: At first, we have two sorts of repositories. We have the repository the TextGrid Lab is working with. There applies the AAI. That is not public data. And we have static data, where there is only published data and data that is open for everyone to use. And in the TextGrid Lab we have data that researchers are working with, and if you have completed the project you can publish the data. And the metadata requirements differ in those two sorts of repositories. For the TextGrid Lab, for every file you are putting in, you only need the format and the title. And if you publish the data, then you need a bit more of metadata. Then you need, for every file the rights owner and a licence. And you only can publish editions or collections in TextGrid and for these collections and editions there is some mandatory metadata you must put in, that is, for example, the license in every case, and the date and some more metadata that is mandatory.

Claudia: So, do you think you could describe this as a kind of TextGrid specific metadata set?

Stefan: Yes, we have a XML schema for that and that is very specific to TextGrid. But we used standardised vocabularies. For example, our title is the dc.title. We used standards where we could. But the metadata schema is the TextGrid metadata schema, using standardised vocabularies.

b. What is your opinion on the adoption or development of common metadata standards to facilitate the searching of various data archives?

Stefan: In the beginning of TextGrid, we had the idea of a baseline encoding, but that was not the metadata, that was the data. We had the idea of having a simpler TEI to generalise the data that is put into the TextGrid Repository. For example, every project is using TEI, but every project is using different TEI. There are many, many versions of TEI. And to be able to search over heterogeneous data, we developed in TextGrid a baseline encoding that is a general encoding that every project could convert its TEI to, so that you have every TEI file also in this baseline encoding, to search over all TextGrid TEI data. At the moment, that is discussed in some workgroups and we just have…

Sibylle: But we won't have the time to do it properly. It's a lot of work and we only have some months left. But this problem is reflecting the problem about TEI and XML itself. TEI is fantastic, because it gives you instruments to process any humanistic text genre you can think of; there are modules for encoding letters or for encoding drama plays or lyrics, whatever. But it would never suit all demands you have in a certain project, and it's left to the editors or to the researchers how deep they encode or how they encode in general. And so, it gives you a lot of freedom which means you have the pitfall that you will never have one TEI model. .. There is something like "TEI light" and "TEI simple" and so on, there are loads of activities trying to cover this problem, but probably this will never be really solved, because the demands are too varied.

19) a. How do you deal with different legal/copyright aspects of the archived data?

Claudia: You said that in the TextGrid terms of use, it is stated that the users are only allowed to ingest data of which they own the copyright, is that right?

Sibylle: Data of which the copyright situation is cleared.

Our survey showed that all kinds of copyright transfer agreements exist in the different archives. An example is the fact that it is difficult for a data archive service to realise sustainability when it does not have (enough) rights to migrate or emulate the data in its archive?

Claudia: How do you deal with this problem?

Stefan: We don't at the moment. I think we have to develop a policy for that. And I think there are, let's say, two stages of data curation. The first is, if you can migrate the data without loss of information, then maybe the archive just can do it. But if there is a possibility of content loss or content change or so, then the depositor must be involved, the researcher must be asked which of the content and which of the significant properties of the data are to be kept. And there must be a clear policy in the future. And for that information the user or the depositor must be involved. And I think in DARIAH there is, at least at the level of research data lifecycle, that is being discussed. I don't think that DARIAH will have a solution for this… We had some projects here at the SUB [Goettingen State and University Library], for example kopal and DP4Lib tried to make policies and tried to develop some technical solutions for that. But I think it's a wide field so that maybe not even DARIAH will have technical solutions at the end of the project.

b. How does the archive handle different user licences for the deposited datasets?

Stefan: What is meant by user license?

Claudia: I think what is meant is a license that tells users or re-users of the data what they are allowed to do with the data.

Stefan: If you publish data, the user must give a license.

Claudia: Is he free to choose the license?

Stefan: Yes, he is. Of course he has to be sure that the license does not conflict with the terms of use.

c. Which possibilities do you have as an archive to negotiate with the depositor(s) in case of unclarity?

Sibylle: I guess we would not take such an active role. Users have to adapt to our point of view and so on. We can't take a role like stepping in, obtaining rights or whatever, we can't do this. We can only propose solutions, but, it's not our job to take an active role.

Claudia: Yes, I guess you would probably need extra staff for that…

20) Other suggestions/remarks?

see below…

Additional questions: to be asked depending on the answers given by the interviewee

21) Sustainability of a data archive is dependent on many factors. We have already discussed the technical aspects. In an economic sense, sustainability is related to the financial viability of the archive. How do you guarantee this financial viability and what is the business model for your archive?

[We already talked about this at an earlier point of the interview…]

22) Which PID system do you use and why?

Stefan: We are using the handle system that is provided by the EPIC consortium where the GWDG is involved. And we are using – in the new TextGrid Repository that we are deploying tomorrow – version 2 of the EPIC handle system. We use it because it's provided by the GWDG.

Sibylle: Maybe we should mention the future SUB PID policy. There will be a future policy and as we are part of the SUB…

Stefan: Yes, there will a policy for the SUB and of course it will apply to TextGrid. And the handle system, the EPIC version 2 handle system will also provide automatically a DOI for everything.

Claudia: So every data set or data file in the TextGrid Repository gets a handle or only the published ones?

Stefan: Only the published ones.

23) One of the objectives of DASISH is promoting of convergence of data archive services. Every discipline (and often every archive) has its own tools to facilitate retrievability of datasets. Apart from free-text searching, classification may be used to retrieve datasets within a specific field. An overview of classification codes in use within Europe may be helpful in realising convergence, as this could be the start for concordance activities. Does your data archive use special (classification) codes for retrieval and are you open for the idea of classification concordance?

Sibylle: I'm not really sure what this would look like. To my understanding, the TEI is already a classification system, so what else could there be for the humanities, it's the established classification system. Or what would be examples for other classifications?

Claudia: I'm thinking of discipline specific thesauri or vocabularies. So, the more content-wise aspects, are they already covered in TEI?

Sibylle: Well, in TEI, there's a tag for everything, so… But maybe I'm not fully understanding the question.

Stefan: There is a possibility in TextGrid that projects can use their own metadata schemas, but I don't know if this applies to the question.

Claudia: Well, I think the main use of such a classification concordance would be to improve the searchability across archives.
Sibylle: So is it about controlled vocabularies?

Claudia: Yes. … But you don't see an urgent need for that at the moment?

Sibylle: Yes, we are advising any partner or user to use them, but users we are not working closely with - we don't have the possibility to influence their research design. So we are trying, within our user meetings, we will have one only about normed data and controlled vocabularies to strengthen the importance of using these standards and so on. But it's up to the users what they choose.

24) In a data archive service, it is expected that its primary expertise lays in the field of archiving and giving access to datasets. The actual physical storage of data is perhaps a task that may be performed by a third party. How do you think about outsourcing the physical storage of data? In your opinion, what would be the advantages and disadvantages of having a physical back-up service on a European scale?

Claudia: You already told us that the physical storage of TextGrid is done the GWDG. So what do you think about the outsourcing of physical storage? Or do you have any more comments on that.

Stefan: Yes, in a way it is already outsourced to the GWDG. And when the data is hosted in the DARIAH repository in the future, there are physical backups, not yet on a European scale but on the scale of at least Germany. There are various data centres in Germany that are partners in DARIAH-DE and the data that is in the repository will be replicated, according to a certain policy, among these data centres in Germany. So we will have a copy in Munich/Garching and a copy at the KIT in Karlsruhe and so on. And I think that will be spread around Europe when the repositories are working on a European level. But when that will be, I don't know. It is one further contribution to DARIAH-EU.

25) Do you have any more questions or comments you want to share with DASISH?

Sibylle: Maybe for TextGrid we could mention a fact which is quite interesting: TextGrid consists of the Lab and the Repository. And initially, in the first years of TextGrid we thought that the Lab would be the most interesting environment for the users, but in the past couple of years, users are mainly interested in the Repository. So there is a much bigger focus now on the repository than it used to be or than we thought. So, that's quite interesting.

And another thing which we came across is: The interface of such a thing needs to be both easy to use and attractive. It's something we only got aware of a bit too late or we didn't have the capacities or resources to put much effort into it, but we do notice that this is quite important to the users or to the use in general of such a service.

Stefan: And the DARIAH collection registry provides access to the collections via an OAI-PMH protocol and in TextGrid we implement an OAI-PMH interface. And that is the interface to the generic search. So TextGrid can be accessed via OAI-PMH.

Claudia: Do you have the impression that the redesign of the TextGrid logo/design one year or so ago is attracting more users now?

Sibylle: We can't say. I think, to get more users or bind the users to us, the most important thing is training and workshops. Anything else is minor, it's nice to have, but it's not as important.

C DRI

Name interviewee:	Aileen O'Carroll
Function:	Policy Manager
Organisation:	Digital Repository of Ireland
Name data archive:	Digital Repository of Ireland
Country:	Ireland
ESFRI:	DARIAH
Telephone number:	+353867722380
Email address:	aileen.ocarroll@nuim.ie
URL of the data archive:	http://www.dri.ie

Personal URL: http://www.dri.ie/dri-team/aileen-ocarroll

11 a. ***What is the technical background of your data archive and how was the decision for the specific technical setting used by your archive made?***

The first thing to know is that the Digital Repository of Ireland (DRI) is a four-year infrastructure project (2,5 years have now elapsed and there is 1,5 year to go). DRI is a federation of different universities.[19] The DRI infrastructure is not completely established yet. It is in the process of being built, and will be based on a Fedora backend using Hydra. Drupal will be used for the user interface.

Contact is not the right person to talk about particular technical details.

With a view of using open source solutions, there was an internal review process of different technical options and a number of different platforms. The decision for the specific technical setting was then based on the results of this review by choosing the platform that best suits the needs of DRI. The information on this review process may be available later, but is currently not published yet.

11b. ***Technology is changing rapidly. How are the technical developments evaluated?***

DRI will have technological watches (software watches and hardware watches) implemented. These will be based on Fedora and Hydra in the future. But at the moment, these systems are not in place yet.

How often do these evaluations take place and how does this influence the technical backbone of the data archive?

Evaluations of technical developments did not take place yet. People at DRI don't know (haven't established that) yet. They will be looking into the DASISH report and see what other people do.

12. ***Which authorisation/authentication tools and methods do you currently use, both in ingesting and in accessing datasets in your archive?***

As this question is a technical one, Contact is not the best person to know how this is implemented, but can provide some information about this.

DRI aims at to have some ways of allowing user authenticating. At the moment, the repository does not have any users yet; but there will be a requirement to have an authentication and authorization infrastructure (AAI) for both data ingest and access. The AAI system should allow different types of users to be authenticated and granted with different permissions: anonymous, authenticated, and authorized users. A person as an anonymous user will be able to see some of the content. The more access the users will have, the more they will be able to do with the content.

[19] Contact is attached to one of these organizations called the Irish Qualitative Data Archive.

In fact, the initial authentication system to be used is Edugate. There is however one main concern with this system: the Edugate system is designed for users from universities and research organizations in Ireland, and will thus not be sufficient, as DRI wants users who are not necessarily university members to have access to the data.

Thus, in addition, there is a plan to use eduroam, and eventually the Shibboleth federated identity solution. However, a procedure dealing with restricted access to data has not been developed yet. This constitutes one of the next steps of the project and should be done by the end of December 2014.

Are you satisfied with these?

Since DRI is not currently using an AAI system, this question cannot be answered yet.

13 *In the survey you indicated your primary designated community. To what extent do you know this designated community?*

The project includes humanities and social science (HSS) data. The primary designated community is people from the whole HSS. Data to be included within DRI include data from both universities and cultural institutions (e.g. museums, archives, art galleries, and art libraries). DRI engages / interacts with the different stakeholders in a number of ways:

- Requirement interviews have been conducted for a wide range of stakeholders.[20]
- A stakeholders committee made up of national stakeholders has been established;
- There are demonstration projects to test the system: using data from people from the humanities and social sciences to test the system;
- DRI has developed parallel short-term projects with different stakeholders as partners.

Which problems do you have in defining your designated community?

Contact is not sure if it has been difficult to define the designated community.

The community is defined when the project was set up. Thus, DRI was built surrounded by the community, and did not exist first and then have to find its designated community.

Obviously, there are issues in terms of building an infrastructure that meets the requirements concerning data archiving best practices and norms. As one of the aims of the project is to link different data collections coming from different domains, DRI wants to make sure to understand the need of each stakeholder. Thus, in fact, defining the primary designated community is not really the problem, but rather to design the project so that it responds to the needs of the different communities in the humanities, scholars, social science scholars, people from data archives and museums, etc.

[20] More details of interviewed stakeholders can be found in the report here: http://dri.ie/digital-archiving-in-ireland-2012.pdf

Do you have contact (on a regular basis) with members of this community to discuss their needs regarding the current and future (re-) use of data in the archive?

DRI has a stakeholder advisory group. Thus, every year, there are regular meetings with the group, including forum presentations. Work produced by DRI is sent to the advisory group for review. In addition, there is collaborative work between DRI and the advisory group within parallel projects. For example, one of these projects is called *Inspiring Ireland*[21]. Developed with many cultural institutions (museums, art galleries) in Ireland, this project is on developing and sharing an online curative display of the designated community's contents (e.g. "high quality images of Ireland's treasured cultural assets"). As a ways of working with the designated community, DRI provides the technological backend, when it comes to run projects like the aforementioned one.

Furthermore, DRI and its designated community are engaged in a number of collaborative projects, as outlined here: http://www.dri.ie/projects. Within these projects, DRI is working with the community on a particular task to better understand the type of content the latter owns/collects.

14 Describe the data archive's training and outreach activities (e.g. organizing workshops, courses for researchers, universities and so on in order to promote proper data management).

Part of the Contact's role is to discuss with most of the researches from humanities and social sciences about good practices for managing data. The issues discussed include e.g. preparing data for archiving, and considering ethical issues when archiving HSS data. In addition, there are in HSS training courses with presentations from diverse stakeholders, including members of the Research Data Alliance and members of some other seminar organizations. In a recent past, DRI has organized several workshops and seminars on different topics related to data archiving, including e.g.:

- Workshop on "Open Access to Humanities Data"[22] (May 2013)
- Workshop on "Getting Started in Digital Preservation"[23] (November 2013)
- Workshop on "Linked Data and OpenRefine"[24] (December 2013)
- Seminar on "The new Creative Commons 4.0 Licence – what's new and why it's important"[25] (March 2014)
- etc.

Later on, in phase II of the project, DRI intends to organize more intensive courses, e.g. on

- Preparing data specifically for deposit within DRI;
- Generating metadata;
- Using the technical infrastructure;
- Ingesting data into the archive, etc.

[21] http://www.inspiring-ireland.ie/

[22] http://www.dri.ie/open-access-humanities-data

[23] http://www.dri.ie/getting-started-digital-preservation

[24] http://www.dri.ie/workshop-linked-data-and-openrefine

[25] http://www.dri.ie/new-creative-commons-40-licence-%E2%80%93-what%E2%80%99s-new-and-why-it%E2%80%99s-important

In addition, DRI is working on producing a range of factsheets and user guides (help pages, support on how to use the DRI infrastructure, guide on how to use Dublin Core, on the way DRI will interpret users' Dublin Core metadata in the archive, etc.). DRI encourages users to follow the best practices in the domain, and gives them guidance on what the infrastructure will do with the data they have provided.

Do you think you could benefit from support/cooperation in the development and maintenance of these training and outreach activities?

Absolutely. It is important to have and work with shared standard and best practices.
If so, of what kind of support/cooperation?

The support/cooperation DRI thinks could benefit from includes support in developing best practices for long-term digital preservation, including generating metadata, data protection, handling ethical issues, data (re-)use, legal framework, etc. This constitutes one of the reasons for DRI to be interested in the reports (like the one with the requirements interviews[26]), in order to get feedbacks.

15 By which measures does the archive support its sustainability on an organizational level?

To ensure sustainability is a major problem (perhaps one of the biggest problems the repository has). People at DRI are focusing on this issue right now, trying to find ways to make the repository sustainable after the project ends (e.g. by getting more funding). The core project is funded until 2015; there are smaller projects funded until 2019. However, at the moment, Contact is not able to provide details on the measures the archive will take to support its sustainability on an organizational level. There are many negotiations going on right now. But, as these are part of a long process that had started, it is difficult at the present stage to predict the measures that will be proposed. Contact cannot address this question in more details.

How has the preservation strategy been set up, what are the arrangements with third parties for storage and what kind of measures have been taken to ensure confidentiality of data?

DRI is in the process of developing a preservation strategy. This is being developed via in-house cross team discussion with reference to international practice.

DRI does not have arrangements with third parties for storage. The storage is going to be a federated system, a federated project involving 7 universities.

Concerning confidentiality of the data, the first issue is that not all data within DRI are confidential (i.e. some of these are available under open access). In case there is no requirement to have confidentiality, then DRI will not take any measures with regard to this issue. In other words, DRI does not have to take measures all the time.

Otherwise, if DRI does have to take measures to ensure confidentiality of data, the digital repository can operate under several different forms:

[26] http://dri.ie/digital-archiving-in-ireland-2012.pdf

- DRI has a legal framework, which requires that the data depositors undertake to ensure that any confidential agreements are met appropriately, probably through anonymizing data.
- The repository has options for access control with a restricted access to the data on the system; in which case the data depositors will be people who have the authority to grant the restricted access. Restricting access to the data can be done in a number of ways:
    - Restriction depending on the user type: The depositor can restrict access to the data by granting permission to a certain user type (e.g. university researchers);
    - Time restricted: Data can be restricted for a certain amount of time, beyond which the data become unrestricted.
- DRI also requires depositors to ensure that appropriate confidentiality agreements are in place. They advise on the use of appropriate consent forms.

Which benefits do you see in the starting (or continuation) of risk management procedures like DRAMBORA and certification trajectories like DSA, DIN 31644 or ISO 16363 (and the documentation of all procedures within your archive related to certification)

DRI is being developed using both the Data Seal of Approval (DSA) and ISO 16363. The main benefit of using these certification trajectories is that they give a structured framework, in which DRI people can address their own development.

The second advantage is that DSA and ISO 16363 are useful in terms of validating DRI's expertise and trust in the eyes of its stakeholders and other third parties. However, this use is quite limited, as many of DRI's stakeholders would not be aware of these certification frameworks.

16 a. ***Has your archive/data service developed a comprehensive policy framework to govern the total of procedures and guidelines for data management, data archiving, and sharing of research data?***

The policy framework developed at DRI follows the DSA and ISO 16363. The development is designed such a way that DRI can automatically apply for certification later on.

And do ethical guidelines fall within this framework and, if so, how have these been implemented?

Yes, absolutely. Ethical guidelines fall within the framework being developed at DRI.
Ethical aspects constitute a main concern for DRI. At the moment, the repository has not taken any restricted data yet, but a legal framework is being developed to take ethical issues into account. Considering negative examples in the past (e.g. the Boston College Case) where data archives break ethical agreements with the researchers, people at DRI want to make sure that, if they give ethical commitments, these will be commitments they can stand over.

Furthermore, DRI has partially implemented ethical guidelines in their framework. The repository will have deposit agreements that data depositors will have to sign. Only members of the organizations / federated institutions are eligible for depositing data. In this agreement, DRI will specify that the repository will protect restrictions to deposit data to the extent the law allows (i.e. the repository can't operate outside the law).

DRI gives training on how to anonymize data, how far the anonymization procedure should be, what kind of data can be archived by researchers.

16b Which routine procedures do you follow in the archive to protect sensitive data?

In practical terms, DRI does not have routine procedures to protect sensitive data. The federated system tries to avoid having an approach that would be based on such routine procedures. Thus, most of the responsibilities for protecting sensitive data fall on the depositor's side. Accordingly, DRI proposes a framework that enforces users/researchers to go through four different processes and a variety of these depending on the data.

- There is a legal framework which ensures that confidential agreements are met appropriately.
- DRI has established best ethical practices and everything works according to these practices, which include Consent forms, and approaches ensuring that people have ethical permissions. Accordingly, users are required to sign documents attesting that they are following these practices.
- DRI provides guidance on data anonymization, but does not take the responsibility to perform/check anonymization of the data. It is the depositor's responsibility to take all necessary steps to protect sensitive data, including anonymization.
- The infrastructure implements different types of access restrictions.

DRI is considering asking users to do evaluation of their data at the very beginning.
Thus, when people will have to deposit their data, DRI will ask them to evaluate the sensitivity of these data. Then, DRI will

- put a sensitivity level on the data;
- and develop appropriate strategies to each sensitivity level.

Not all data are equally sensitive. For some data, DRI might consider not taking the anonymized version; there will be embargo for a given amount of time. In some other data or the same dataset, there might be a set of interviews where people can be identified, and DRI will therefore require from the data depositor to be put restrictions (e.g. anonymization procedure) on these.

17 *What is your need for large-scale interconnection of European data archives (for instance in a common access portal)?*

Obviously, a large-scale interconnection of European data archives is a good thing. But Contact thinks that this question is too broad to be answered.

Could national or international cooperation help your archive to realise economy of scale effects?

Contact cannot answer this question (does not know much about the economy of scale effects).

18 *a. Do you require the presence of a minimum set of metadata in the ingest process?*

Yes, DRI does require the presence of a minimum set of metadata in the ingest process.

And is the depositor free in choosing her metadata standard?

No. The data depositor is not free in choosing her metadata standard. DRI supports a fixed set of standards and the depositor has to choose one scheme from that set. Currently, i.e. in phase I of the project, the very small set of standards supported by DRI includes: Dublin Core, qualified Dublin Core, Metadata Object Description Schema (MODS), Encoded Archival Description (EAD), and MARC XML. Accordingly, for each of these schemes, DRI will require users to fill a low value of Dublin Core fields. For instance, a data depositor may be required to fill five compulsory fields and perhaps twelve suggested fields (the suggested fields are optional).

In phase II of the project, the set of metadata schemes will probably be expanded by considering other schemes such as EBUCore, The Data Documentation Initiative (DDI), but people at DRI haven't really made a concrete agreement on this yet.

18b. What is your opinion on the adoption or development of common metadata standards to facilitate the searching of various data archives?

The adoption or development of common metadata standards to facilitate the search of various data archives is a great idea. The main problem is the development of tools that enables non-experts to use metadata. This is an area where further development is really needed.

Contact thinks that people should be encouraged to use metadata standards that already exist and are most widely used in whatever domain it is. They should not be encouraged to develop their own standards or metadata schemes.

19 *a. How do you deal with different legal/copyright aspects of the archived data?*

When people deposit data within DRI, they will have to attach a copyright statement to these data. They also have to specify a (re-)use statement. DRI encourages people to use Creative Commons where possible.

Our survey showed that all kinds of copyright transfer agreements exist in the different archives. An example is the fact that it is difficult for a data archive service to realise sustainability when it does not have (enough) rights to migrate or emulate the data in its archive?

DRI is aware of the issue that it is difficult to a data archive service to realise sustainability when it does not have (enough) rights to migrate or emulate the data in the archive. To handle this issue, DRI adopts the following approach: people who want to deposit data within the repository have to sign an agreement which will allow DRI to migrate data to third parties if it is required in terms of being sustainable. Also, when an organization becomes a member of DRI, it has to sign this agreement. In case the data depositor does not have ownership of the data, it is her responsibility to get the rights for these data from the owner and to transfer these rights to DRI.

19 b. How does the archive handle different user licences for the deposited datasets?

DRI encourages people to use Creative Commons licenses. The repository has not looked at restricted data yet, nor developed restrictive licenses besides Creative Commons.

A data depositor can attach to a data object any license (s)he wants. However, if (s)he would like to grant bespoken licenses, it is his/her responsibility to manage them. DRI will not take this responsibility. Data depositors have to clearly specify for each object, what kind of license will attach to it. Also, if, for instance, a license specifies that end users have to get the permission from the depositor, DRI will not act as an intermediary between the depositors and the end users. Thus, end users' requests will be sent directly to the data depositors.

In addition, DRI has end user agreements, which specify that the user commits to respect the terms of the data license.

For the data objects, DRI allows the depositor to set a license to the metadata.

DRI encourages the use of Creative Commons licenses on the digital objects. All metadata must have a CCO or a CC BY license attached to it. For more information, see http://www.dri.ie/sites/default/files/files/Fact%20Sheet%20No%202%20Copyright%20and%20Licensing%20ver%203.pdf.

19c. Which possibilities do you have as an archive to negotiate with the depositor(s) in case of unclarity?

DRI has developed a policy that often works. In addition, the repository has defined a set of action policies (i.e. a process) to follow if there is any concern about the licensing of the objects. For instance, in case somebody contacts DRI with a query, he will first be asked to provide some documentation. Next, DRI would then relay this documentation to the depositor, and would ask the depositor about this documentation. Then, DRI will review the documentation, make a decision and take an action. One possibility could be to unpublish the data object.

20 *Other suggestions/remarks?*

DRI is interested to know what kinds of approaches are being taken by other archives to delete data. Do they have a policy stating if data should be deleted or not.

Additional questions: to be asked depending on the answers given by the interviewee

21 *Sustainability of a data archive is dependent on many factors. We have already discussed the technical aspects. In an economic sense, sustainability is related to the financial viability of the archive. How do you guarantee this financial viability and what is the business model for your archive?*

Contact cannot answer this question.

22 ***Which PID system do you use and why?***

DRI uses DataCite DOI as persistent identifier service. This choice was made on the basis of a reviewing process. The repository has performed a review of the persistent identifier services available, and picked the one that fits best their needs.

23 ***One of the objectives of DASISH is promoting of convergence of data archive services. Every discipline (and often every archive) has its own tools to facilitate retrievability of datasets. Apart from free-text searching, classification may be used to retrieve datasets within a specific field. An overview of classification codes in use within Europe may be helpful in realising convergence, as this could be the start for concordance activities. Does your data archive use special (classification) codes for retrieval and are you open for the idea of classification concordance?***

Contact cannot answer the question.

DRI recommends that depositors use the vocabularies appropriate to their field and direct them towards the ones most commonly used. The repository would consider it very difficult to have a single classification that spanned all domains.

24 ***In a data archive service, it is expected that its primary expertise lays in the field of archiving and giving access to datasets. The actual physical storage of data is perhaps a task that may be performed by a third party. How do you think about outsourcing the physical storage of data?***

DRI does not use outsourcing the physical storage of data, and does not intend / is not thinking about using it.

In your opinion, what would be the advantages and disadvantages of having a physical back-up service on a European scale?

Contact thinks that this question is too broad or abstract to be answered. Advantages and disadvantages of having a physical back-up service on a European scale may depend on the cost, robustness, and validity.

E. DDA

Caveat, to be emphasised in advance.
The Danish State Archives, of which DDA is a part, are currently in the middle of a major organizational restructuring, which may cause changes in procedures for archiving etc. Therefore, the answers given below do not necessarily reflect the reality of DDA in a few years' time.

Ad 11)
A: -

B: -

Ad 12) At the moment the DDA-users must submit a signed form to access our service, be it ingesting or accessing datasets. This is of course not satisfactory. We are therefore in the process of implementing the single sign-on system WAYF (http://wayf.dk/), whereby academic staff and students at Danish universities and research institutions can access our service through their local authorization. WAYF will according to plan be implemented at DDA in 2015.

Ad 13) DDA's designated community is the research and higher education area within the fields of social science and public health science, primarily at the Danish universities but also public and private research institutions. We are in regular contact with management, researchers and students at the institutions to strengthen our mutual cooperation and to improve our service according to their needs and wishes.

Ad 14) Several times a year DDA organizes workshops/course at universities. The targeted group is students and PhD students, and the focus of the courses is on reuse of research data and data management. Furthermore we have a guide to data management on our website, providing advice, checklists and links to data management tools (http://samfund.dda.dk/dda/datamanagement.asp).

Ad 15) -

Ad 16)
A: Yes, we have comprehensive written guidelines on the procedure of ingesting, managing, archiving and sharing research data.
B: When ingested the data is scanned for sensitive information, such as social security number, name and Email address, which is then extracted from the data and stored separately on a secure server. To get access to this information, scientist must first seek and get permission from the Danish Data Protection Agency.

Ad 17) -

Ad 18)
A: Yes, the minimum metadata requirements is presented on the ingest form. We use the metadata standard ddi-l, but at the moment the depositor is free to choose metadata standard when ingesting.
B: DDA actively supports the efforts of developing common metadata standards like ddi-l.

Ad 19)
A: In the deposit form it is stated from what date the data producer allows his data to be published and the access level is likewise stated. The access levels run on a scale from 1: free access for scientific/statistical use; to 5: access only by permission from donor. We do not accept any data without the researcher having signed a deposit form.

B: When seeking access to the data at DDA, the users are informed of access level on the particular data set.
If access level 5, the user does not get access to data before donor's explicit approval.
C: If the access level is very high = difficult to get access, DDA at a regular basis asks the data producer if he is willing to accept a lower access level.

Ad 20) No

Ad 21) DDA is part of the Danish State Archives, which ensure a certain degree of financial and organizational security and stability.

Ad 22) DDA uses the PID system DOI. DOI is well known in scientific circles and gives optimal exposure. Furthermore, we have a good cooperative relationship with the DOI's Danish provider.

Ad 23) DDA is using the CESSDA Topic Classification (http://www.gesis.org/en/services/research/thesauri-und-klassifikationen/cessda-topic-classification/)

Ad 24) DDA currently has no plans to outsource the physical storage of data. However, data storage capacity is a recurring issue and challenge, which will only increase in foreseeable future. Due to the personally identifiable and sensible nature of (some of) the data in our archive, such outsourcing decisions should not be taken lightly. The same concern applies to the idea of a European back-up service.

F. ADP

Name interviewee:	Irena Vipavc Brvar
Function:	Expert assistant
Organisation:UL,	ADP
Name data archive:	Arhiv družboslovnih podatkov = Social Science Data Archives
Country:	Slovenia
ESFRI:	CESSDA
Telephone number:	+386/1/5805-293 (skype id: irena.vipavc)
Email address: Irena.vipavc@fdv.uni-lj.si	
URL of the DAS:	http://www.adp.fdv.uni-lj.si
Personal URL:	-

Core questions (to be asked in every single interview)

11a. What is the technical background of your data archive and how was the decision for the specific technical setting used by your archive made?

11 b. Technology is changing rapidly. How are the technical developments evaluated? How often do these evaluations take place and how does this influence the technical backbone of the data archive?

Programs and applications used in ADP:

- some general like MS Office Suite 2010, Adobe professional
- DB currently in Access / files on folders but we are in the process of moving everything to FEDORA with JAVA application on top
- JIRA – for task management
- Django – web

- Oxygen – XML editor, Nesstar Publisher
- Nesstar
- Tortoise SVN, Owncloud

All of the above are upgraded so the newest versions are used on regular basis. Like in the case of any other application, it might happen that the newest version of software will not perform tasks the way they were executed with previous versions of the same software. Therefore, regular checks and adjustments are needed. When a need is recognized, we evaluate new technologies and establish if they can be used, how they can be used and what is the expense of implementing new solutions.

12. Which authorisation/authentication tools and methods do you currently use, both in ingesting and in accessing datasets in your archive? Are you satisfied with these?

At the ingest part: – Depositors are usually members of Slovenian institutes / universities so they hold official e-mail address. At the same time we know most of them personally and have phone contacts with them in the depositing process.

As for accessing – Users need to register and confirm End user Licence agreement. Their e-mail address is later on used for authentication. Access to scientific use / sensitive data is more strict and subject to special agreement and authentication of users.

At the moment the tool that we are using for distribution (Nesstar) does not support AAI authentication. But we are considering other options to be able to use it in the future.

13. In the survey you indicated your primary designated community. To what extent do you know this designated community? Which problems do you have in the defining your designated community? Do you have contact (on a regular basis) with members of this community to discuss their needs regarding the current and future (re-) use of data in the archive?

Designated community for ADP is defined in the previous document. It includes students, researchers and professors from social science and humanities fields from all Slovenian universities. We communicate with our users via e-mail (monthly newsletter and other promotional notifications), via blog and at training courses / workshops and other promotional events of ADP or conferences where ADP or its employees are present. At the end of training courses we ask participants to fill in a questionnaire which provides some feedback about training and our services on general.

Bachelor and Master theses and PhD dissertations, which are (partially) based on microdata distributed by ADP, compete for the annual Klinar Fund award.

ADP offers services which are general to data archives all around the world and especially to CESSDA members. We follow changes and needs of users abroad and try to implement them locally if we recognize the need.

14. Describe the data archive's training and outreach activities (e.g. organizing workshops, courses for researchers, universities and so on in order to promote proper data

management). Do you think you could benefit from support/cooperation in the development and maintenance of these training and outreach activities? If so, of what kind of support/cooperation?

We offer training courses in the form of workshop for users / students at different universities / faculties across Slovenia, invited or open for everyone and organized several times per year. We also offer training courses to researchers / possible depositors. Lately we have been covering RDM as well. Not long ago we organized quite a successful workshop on the Role of librarians in opening up research data and are planning to organize another bigger event for researchers in December 2014. Both of the events are supported by the FOSTER project /Eifl.

Presence at several local and international conferences in the form of presentation or participating in poster sections, is also a form of promotional activity and creates a possibility to communicate with our community and colleagues from our field.

CESSDA provides and will provide training even more frequently in the future. The training is and will be organized within CESSDA and beyond to provide quality information on the best practices surrounding operational processes and data management. CESSDA ERIC - Archive and Data Management Training and Information Centre is under development.

We also cooperate with some other related centres / infrastructures, especially in the field of open science / data and long-term digital preservation.

15. By which measures does the archive support its sustainability on an organizational level? How has the preservation strategy been set up, what are the arrangements with third parties for storage and what kind of measures have been taken to ensure confidentiality of data? Which benefits do you see in the starting (or continuation) of risk management procedures like DRAMBORA and certification trajectories like DSA, DIN 31644 or ISO 16363 (and the documentation of all procedures within your archive related to certification)

ADP has internal documentation on procedures of preserving and working with files following the OAIS model. Files are saved on multiple servers and are backed-up at appropriate regular intervals. Backed-up copies on the mirrored disk are also saved in separate locations outside ADP / Ljubljana.

Separate procedure is in place for backup registered user database, data on the Nesstar server and the website. This is done automatically once per month.

More thorough preservation policy and strategy for backup copies is in the preparation and should be fully functional in a year. Data and documentation will be saved on servers of the National and University Library (NUK) which use MD5 checksums. NUK is not TRD compliant, but uses formal approach procedures for digital preservation. NUK published digital preservation strategy and works on its implementation (URN:NBN:SI:DOC-HPEWEXEN). Additional encrypted backup will be made on servers of ARNES (The Academic and Research Network in Slovenia).

Obligations of CESSDA service providers are an extensive version of the DSA. We are in the process of obtaining the DSA. The first evaluation inside CESSDA already circled around and currently we are preparing some documents and improving some processes, especially for digital preservation, as well as writing down all necessary procedures and translating the text to English. The Slovene version of the description of our procedures was published in the Knjižnica journal (URN:NBN:SI:DOC-I29E9V0L).

16. a. Has your archive/data service developed a comprehensive policy framework to govern the total of procedures and guidelines for data management, data archiving, and sharing of research data? And do ethical guidelines fall within this framework and, if so, how have these been implemented?
    b. Which routine procedures do you follow in the archive to protect sensitive data?

We have several documents that support data management, archiving and sharing. Some of them are used for internal processes, others are meant for depositors / researchers. Publicly available documentation (in Slovene only) could be found on our website, additional help could be provided if asked for. We follow standards proposed by CESSDA and we promote materials on this topic provided by other CESSDA members (UKDA, GESIS).

There is no Ethical Review Committee in our field in Slovenia. Users are advised to follow the codes of ethics of the Slovenian Sociological Society and the Statistical Society of Slovenia. Relevant national Data Protection and Intellectual Property legislation has to be observed by data producers in order for their data to be accepted into preservation and retention. This ensures that users work in line with the legal requirements.

Data users agree on two sets of terms. On the first set when creating the account / registering for the access and on the second one when using data. Each study may have its own special access conditions defined. In any of these, the requirement not to try to identify an individual respondent is stated.

Regardless of the above, we still check the data when we receive them and discuss sensitivity of variables (normally, demographic variables are checked) with the depositor if we consider it necessary. In some cases we create public use file (PUF) and more anonymised versions that is scientific use file (SUF) which are then available on special request only. We feel that the majority of our users do not need detailed information (e.g. detailed geographical location) but we still prefer to keep them in data files for more detailed analysis.

17. What is your need for large-scale interconnection of European data archives (for instance in a common access portal)? Could national or international cooperation help your archive to realise economy of scale effects?

As a CESSDA member we are obliged to follow its rules defined in the statue, chapters 3 and 7. Among other issues, it is defined that CESSDA shall facilitate access to social science data resources for researchers regardless of the location of either researcher or data within the European Research Area (ERA), and beyond Enable, extend and promote access agreements, licensing models, and any other legal and organisational measures that enable and extend such access to distributed data resources, while taking specific national requirements into account. We shall coordinate and support development, installation and maintenance of the technical infrastructure which would allow access to distributed data resources (see http://www.cessda.net/export/sites/default/about/docs/Statutes-for-CESSDA-18-June-2013-final-version-brand.pdf). Economy of scale is relevant to us under the condition that any tool which supports these needs is flexible in allowing local implementation (language, style, functionality, etc.)

18. a. Do you require the presence of a minimum set of metadata in the ingest process? And is the depositor free in choosing her metadata standard?
    b. What is your opinion on the adoption or development of common metadata standards to facilitate the searching of various data archives?

Depositor can deposit their metadata either by providing study information in the deposit form (available online or can be download) or directly in Nesstar Publisher and then by sending us appropriate files. Both options are based on the DDI 2.x specification. If metadata and documentation provided are insufficient for long-term preservation of data, the depositor is contacted with a request to provide missing metadata. The minimum amount of provided metadata should at least include the CESSDA recommended fields. However, we encourage depositors to provide as much information as possible to fill in DDI fields to the fullest possible (http://www.ddialliance.org/sites/default/files/cessda-rec.pdf).

We support adoption or development of common metadata standards for searching. If catalogue records for books could be common for different scientific fields, it should be possible to agree on the common standards for data files as well. Moreover, we would like to be involved in such development if possible.

Perhaps also see development on this topic in DwB project, especially in WP8 (e.g. http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d8-3_workflows-and-dataflows-for-resource-discovery-models_report.pdf). Some development in that direction has also taken place in DataCite (DataCite Metadata Schema 3.0).

19. a. How do you deal with different legal/copyright aspects of the archived data?
    Our survey showed that all kinds of copyright transfer agreements exist in the different archives. An example is the fact that it is difficult for a data archive service to realise sustainability when it does not have (enough) rights to migrate or emulate the data in its archive?

    b. How does the archive handle different user licences for the deposited datasets?
    c. Which possibilities do you have as an archive to negotiate with the depositor(s) in case of unclarity?

We use a deposit form which has only slightly changed through time since the establishment of ADP. Until now, anything specific that we have had to agree upon with depositors or their successors / institutes was done without coming across any issues. We build on trust.

Different licences are handled inside Nesstar server. It is defined which types of users have sufficient rights to access specific data.

20. Other suggestions/remarks?

Additional questions: to be asked depending on the answers given by the interviewee

21. Sustainability of a data archive is dependent on many factors. We have already discussed the technical aspects. In an economic sense, sustainability is related to the financial viability of the archive. How do you guarantee this financial viability and what is the business model for your archive?

We described funding in the table provided few months ago. In additional to that, we should note that the organization is financed on a 5 year project basis.

CESSDA is in the list of Slovenian Research Infrastructure Roadmap 2011-2020. Slovenia signed MoU for joining the new CESSDA, and ADP is chosen as the service provider for Slovenia.

22. Which PID system do you use and why?

Due to the collaboration with the Slovenian National and University Library, we will use URN in the near future. Identification will follow the internal ADP organizational rules related to identification of studies, related documentation and datasets.

23. One of the objectives of DASISH is promoting of convergence of data archive services. Every discipline (and often every archive) has its own tools to facilitate retrievability of datasets. Apart from free-text searching, classification may be used to retrieve datasets within a specific field. An overview of classification codes in use within Europe may be helpful in realising convergence, as this could be the start for concordance activities. Does your data archive use special (classification) codes for retrieval and are you open for the idea of classification concordance?

I am not sure if I understand the question right.

We use DDI controlled vocabularies which are suggested on DDI Alliance website.
At the moment we use CESSDA topic classification for topics and in the future we plan to use CESSDA ELSST.

For most of the other fields we developed defined elements which are used when preparing metadata study descriptions and are offered in Nesstar Publisher templates to depositors as well.

24. In a data archive service, it is expected that its primary expertise lays in the field of archiving and giving access to datasets. The actual physical storage of data is perhaps a task that may be performed by a third party. How do you think about outsourcing the physical storage of data? In your opinion, what would be the advantages and disadvantages of having a physical back-up service on a European scale?

It is a fact that archives do not have primary knowledge about storage. Therefore, in case some general recommendations and procedures, as well as tools or even infrastructure were developed, a lot of archives, especially smaller ones like ours, would use it. There are companies / organizations which are specialised in long-term preservation and have necessary knowledge and means; consequently it would be wise to use them

We currently have our own storage system in the organization, but we also “outsource” it by storing it at the second (NUK) and the third (ARNES) storage location.

There is a question though, if we would need to change the Depositor agreement form due to this change. Even though that most files in outsourced location are encrypted. However, we see the advantage of being involved in common preservation collaboration such as the model of DataPass in US for disciplinary data services.

Appendix 5: Interviewees' organisations

BAS

The Bavarian Archive for Speech Signals (BAS) is a public institution hosted by the University of Munich founded with the aim of making speech resources of contemporary spoken German as well as tools for the processing of digitized speech available to research and speech technology communities. Speech material will be structured in a manner allowing flexible and precise access, with rich annotations, metadata and linguistic-phonetic evaluation forming an integral part of it. Since 20th of June 2013 the BAS is a licensed CLARIN B center.

LINDAT / UFAL

The LINDAT/CLARIN Centre for Language Research Infrastructure provides technical background and assistance to institutions or researchers who want to share, create and modernise their tools and data used for research in linguistics or related research fields. The project also provides an open digital repository and archive open to all academics who want their work to be preserved, promoted and made widely available. LINDAT/CLARIN is funded by the Ministry of Education, Youth and Sports of the Czech Republic.

TextGrid

TextGrid is a joint effort of ten partners. It started in 2006 and is funded by the German Federal Ministry of Education and Research (BMBF). The aim of TextGrid is to establish a research infrastructure for digital and collaborative research in the humanities. Heart of TextGrid is the Virtual Research Environment (VRE) that allows users to edit, store and publish data with a number of tools. Published data is stored in the TextGrid repository, which is operated and maintained by Göttingen State and University Library.

DRI

The Digital Repository of Ireland is a national trusted digital repository for Ireland's social and cultural data. The repository will link together and preserve both historical and contemporary data held by Irish institutions, providing a central internet access point and interactive multimedia tools. As a national e-infrastructure for the future of education and research in the humanities and social sciences, DRI will be available for use by the public, students and scholars. A research consortium of six academic partners working together to deliver the repository, policies, guidelines and training builds the Digital Repository of Ireland.

DDA

The Danish Data Archive (DDA) is a national data bank for researchers and students in Denmark and abroad. DDA is dedicated to the acquisition, preservation and dissemination of machine-readable data created by researchers from the Social Sciences and the Health Sciences communities. DDA, furthermore, has quantitative historical data materials, especially transcribed historical censuses. DDA is an independent unit within the group of Danish National Archives.

ADP

Social Science Data Archives offer access to data that are interesting for social science analysis, with emphasis on problems related to Slovenian society. Priority is given to theoretically significant and methodologically well-designed studies, especially data gathered over a period of time and international comparative data that include Slovenia.

Target users are national and international researchers, teachers and students who are data and statistical literate.

Appendix 6: List of Abbreviations

AAI	Authentication and Authorisation Infrastructure
CCSDS	Consultative Committee for Space Data Systems
CESSDA	Council of European Social Science Data Archives
CLARIN	Common Language Resources and Technology Infrastructure
DADS	Data Archive Description Sheet
DAS	Data Archive Service
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DASISH	Data Services Infrastructure for the Social Sciences and Humanities
Data-PASS	Data Preservation Alliance for the Social Sciences
DDI	Data Documentation Initiative
DRAMBORA	Digital Repository Audit Method Based On Risk Assessment
DSA	Data Seal of Approval
ESFRI	European Strategy Forum on Research Infrastructures
ESS	European Social Survey
OAIS	Open Archival Information System
PID	Persistent Identifier
SHARE	Survey of Health, Ageing and Retirement in Europe
SOAP	Simple Object Access Protocol
SSH	Social Sciences and Humanities

Appendix 7: Data Archive Description Sheets

Språkbanken (Swedish Language Bank)

Functionalities	Short Description	References
	Administrative Context	
Funding	Public	
Depositor Agreements	No	
Usage Agreements, Code of Conduct to be signed	No	
Policies in Place	No	
Rights on Data Claimed by the Archive	Depositor retains all rights; in-house resources normally released under CC-BY/LGPL	
Data Curation Strategy	Migration; linguistic annotations added (POS; lemma; syntax trees)	
	Pre-ingest	
Primary Community in Focus for Deposits	Researchers in language technology and linguistics	
Secondary Communities accepted for Deposits	SSH researchers	
	Ingest	
Formats accepted and curated	Negotiated on a case-by-cases basis, since Språkbanken is not primarily a data archive	
Formats accepted and not curated	-	
Metadata formats accepted	CMDI; DC; TEI	
User-based ingest	Decided on a case-by-case basis	
	Archival Storage and Preservation	
Size of current archive in TB	< 0.5	
Size of current archive in other means (collections, files, etc.)	over 200 corpora/5 billion words + 23 lexicons / 700,000 entries	
Maximal deposit size in TB	No	
Long-term guarantees/standards of trust	No	
Checks on quality/quality control	-	
	Dissemination	
Costs/conditions for access	Free online access; free download of lexicons and scrambled corpora	
Tools/Interfaces used for access	Korp corpus interface; Karp lexicon interface; REST web services; complete resource download	<http://spraakbanken.gu.se/korp/>; <http://spraakbanken.gu.se/karp/>; <http://spraakbanken.gu.se/eng/research/infrastructure/korp/sentencesets>; <http://spraakbanken.gu.se/eng/resources/lexicon>
ESFRI	CLARIN and DARIAH	
Name of the archive	Språkbanken (Swedish Language Bank)	

Slovene Social Science Data Archive (ADP)

Functionalities	Short Description	References
	Administrative Context	
Funding	Since 2004, the funding of operations has been provided by MVZT (Ministry of Higher Education, Science and Technology) within the infrastructure programme 'Network of Research Infrastructure Centres at the University of Ljubljana"	
Depositor Agreements	Licence agreement (Deposit form) signed by depositor and head of ADP; it includes conditions for distribution (Creative Commons License) and list of deposited materials.	
Usage Agreements, Code of Conduct to be signed	Users fulfil Online registration form; every request is reviewed and confirmed manually, usually within one working day. Special procedures apply to protected data files.	
Policies in Place	Copyright, Access Conditions and Acceptance of Terms and Conditions are included in the Licence Agreement. As well as in User Registration form.	
Rights on Data Claimed by the Archive	No. Data depositors retain all rights to their own data.	
Data Curation Strategy	ADP uses migration strategy for providing long-term preservation of ADP's data holdings. The data formats that ADP uses for storing data are chosen with long-term preservation in mind, avoiding proprietary, closed or rarely used file formats. Extensive metadata are collected and stored to ensure usability of data. ADP transforms documentation and data when received. Routines for handling data are described in several internal documents. Basic description published in Knjiznjica journal.	
	Pre-ingest	
Primary Community in Focus for Deposits	Researchers and students from various, mainly social sciences, disciplines; sociology, political sciences, psychology, social work, economy, humanity, pedagogy, … Priority is given to theoretically significant and methodologically well designed studies, especially data gathered over a period of time and international comparative data that include Slovenia.	
Secondary Communities accepted for Deposits	Marketing agencies – data should be free of charge (at least for academic users; CC license)	
	Ingest	
Formats accepted and curated	List of preferred formats	
Formats accepted and not curated	-	
Metadata formats accepted	DDI	
User-based ingest	Via mails	
	Archival Storage and Preservation	
Size of current archive in TB	0.03	
Size of current archive in other means (collections, files, etc.)	700 datasets	
Maximal deposit size in TB	-	
Long-term guarantees/standards of trust	Indefinitely; self-assessment	
Checks on quality/quality control	access controls; check sums; audit trials	
	Dissemination	
Costs/conditions for access	No	
Tools/Interfaces used for access	Website and Nesstar	
ESFRI	CESSDA	
Name of the archive	Slovene Social Science Data Archive (ADP)	

ADPSS (*Archivio dati per le scienze sociali*)

Functionalities	Short Description	References
	Administrative Context	
Funding	Public	
Depositor Agreements	Data deposit form	http://www.sociologiadip.unimib.it/sociodata/wp-content/uploads/Modulo-acquisizione-dati.pdf
Usage Agreements, Code of Conduct to be signed	Yes, a licence agreement	http://www.sociologiadip.unimib.it/sociodata/wp-content/uploads/Request-Form.pdf
Policies in Place	Access conditions and acceptance of terms and conditions are included in the licence agreement	
Rights on Data Claimed by the Archive	Non-exclusive rights granted	
Data Curation Strategy	Not specified	
	Pre-ingest	
Primary Community in Focus for Deposits	Sociologists	
Secondary Communities accepted for Deposits	-	
	Ingest	
Formats accepted and curated	Any	
Formats accepted and not curated	-	
Metadata formats accepted	DDI	
User-based ingest	No	
	Archival Storage and Preservation	
Size of current archive in TB	0,15 TB	
Size of current archive in other means (collections, files, etc.)	1,370 datasets	
Maximal deposit size in TB	No	
Long-term guarantees/standards of trust	No	
Checks on quality/quality control	Manual checks and quality controls	
	Dissemination	
Costs/conditions for access	Free or refund required (depends on datasets)/ Licence agreement	
Tools/Interfaces used for access	NESSTAR; SPSS	
ESRFI	CESSDA; ESS	
Name of the archive	ADPSS	

ADS (Archaeology Data Service)

Functionalities	Short Description	References
	Administrative Context	
Funding	Other than income from research and development projects the main income comes from one-off deposits charges levied at the point of deposit	
Depositor Agreements	See reference	http://archaeologydataservice.ac.uk/attach/guidelinesForDepositors/ads_licence_form.pdf
Usage Agreements, Code of Conduct to be signed	See reference	http://archaeologydataservice.ac.uk/advice/termsOfUseAndAccess
Policies in Place	See reference	http://archaeologydataservice.ac.uk/attach/preservation/PreservationPolicyV1.3.1.pdf
Rights on Data Claimed by the Archive	Non-exclusive Rights to the Archive	
Data Curation Strategy	Migration	
	Pre-ingest	
Primary Community in Focus for Deposits	Archaeologists	
Secondary Communities accepted for Deposits	-	
	Ingest	
Formats accepted and curated	List of preferred formats	
Formats accepted and not curated	-	
Metadata formats accepted	Dublin Core; UK MIDAS heritage data standards, plus other specific metadata	
User-based ingest	online upload; DVD/CD; mail	
	Archival Storage and Preservation	
Size of current archive in TB	2.65 TB	
Size of current archive in other means (collections, files, etc.)	1119 collections and 1M+ files	
Maximal deposit size in TB	-	
Long-term guarantees/standards of trust	DSA peer-reviewed self-assessment	
Checks on quality/quality control	Access control	
	Dissemination	
Costs/conditions for access	Access is free, subject to the ADS terms and conditions of use, which are broadly equivalent to CC-BY-NC, although many forms of commercial not-for-profit use are allowable	
Tools/Interfaces used for access	Most data sets are available for download, but some resources have tailored databases or map-based interfaces which allow them to be queried online	
ESFRI	DARIAH	
Name of the archive	ADS	

APIS (Portuguese Social Information Archive)

Functionalities	Short Description	References
	Administrative Context	
Funding	Public	
Depositor Agreements	No	
Usage Agreements, Code of Conduct to be signed	No	
Policies in Place	See reference	http://www.apis.ics.ul.pt/depositar.html
Rights on Data Claimed by the Archive	No rights obtained	
Data Curation Strategy	Migration	
	Pre-ingest	
Primary Community in Focus for Deposits	Sociologists	
Secondary Communities accepted for Deposits	-	
	Ingest	
Formats accepted and curated	Any	
Formats accepted and not curated	-	
Metadata formats accepted	DDI	
User-based ingest	mail	
	Archival Storage and Preservation	
Size of current archive in TB	-	
Size of current archive in other means (collections, files, etc.)	100,000 data files	
Maximal deposit size in TB	-	
Long-term guarantees/standards of trust	-	
Checks on quality/quality control	Access controls	
	Dissemination	
Costs/conditions for access	No	
Tools/Interfaces used for access	PHP, MySQL	
ESFRI	CESSDA, ESS, SHARE	
Name of the archive	APIS	

BAS (Bavarian Archive for Speech Signals)

Functionalities	Short Description	References
	Administrative Context	
Funding	Public; Third party revenues	
Depositor Agreements	Bilateral agreement with each depositor. See reference	http://www.phonetik.uni-muenchen.de/Bas/BasTemplateContract.pdf
Usage Agreements, Code of Conduct to be signed	See reference	https://www.phonetik.uni-muenchen.de/Bas/BasTermsOfUsage_eng.pdf
Policies in Place	See reference	https://www.phonetik.uni-muenchen.de/Bas/BasPolicyExternalResources_eng.pdf
Rights on Data Claimed by the Archive	Granting of non-exclusive rights to third parties	
Data Curation Strategy	Validation according to BAS quality guidelines; metadata production; ingest in repository	
	Pre-ingest	
Primary Community in Focus for Deposits	Academics in humanities	
Secondary Communities accepted for Deposits	Science, engineering	
	Ingest	
Formats accepted and curated	List of preferred formats	http://www.bas.uni-muenchen.de/forschung/Bas/BasFormatseng.html
Formats accepted and not curated	-	
Metadata formats accepted	CMDI; DC; IMDI; support for CMDI production via web services (COALA)	
User-based ingest	No	
	Archival Storage and Preservation	
Size of current archive in TB	3.5 TB	
Size of current archive in other means (collections, files, etc.)	-	
Maximal deposit size in TB	No limit	
Long-term guarantees/standards of trust	Data Seal of Approval (granted in 2013); CLARIN Centre Type B (granted in 2013); Member of CLARIN Service Provider Federation; long-term commitment of University of Munich	
Checks on quality/quality control	Validation according to BAS quality guidelines	http://www.bas.uni-muenchen.de/forschung/BITS/TP2/Cookbook/
	Dissemination	
Costs/conditions for access	Free for European academics (AAI); user fees for others depending on resource	
Tools/Interfaces used for access	CLARIN BAS repository; CLARIN Virtual Language Observatory; CLARIN Federated Content Search; Web Catalogue; CMDI browser ARBIL	http://www.phonetik.uni-muenchen.de/forschung/bay_arch_sprsig/, http://catalog.clarin.eu/vlo/, http://weblicht.sfs.uni-tuebingen.de/Aggregator/, http://www.bas.uni-muenchen.de/forschung/Bas/BasKorporaeng.html
ESFRI	CLARIN; SHARE	
Name of the archive	Bavarian Archive for Speech Signals (BAS)	http://www.phonetik.uni-muenchen.de/forschung/bay_arch_sprsig/

CSDA (Czech Social Science Data Archive)

Functionalities	Short Description	References
Administrative Context		
Funding	Public Funding; Third Party Funding	
Depositor Agreements	See reference	http://archiv.soc.cas.cz/sites/default/files/dohoda_o_depozici_dat.doc
Usage Agreements, Code of Conduct to be signed	See reference	http://archiv.soc.cas.cz/login/zp_aj.pdf ; http://archiv.soc.cas.cz/login/s_aj.pdf
Policies in Place	See reference	http://archiv.soc.cas.cz/download/1907/Archivacni_rad_CSDA.docx
Rights on Data Claimed by the Archive	Depositor retains all rights	
Data Curation Strategy	Migration; Bitstream preservation	
Pre-ingest		
Primary Community in Focus for Deposits	Sociologists	
Secondary Communities accepted for Deposits	All other social scientists	
Ingest		
Formats accepted and curated	List of preferred formats	Section 2.8 – page 9 of PP (in Czech only)
Formats accepted and not curated	-	
Metadata formats accepted	DC; DDI	
User-based ingest	CD/DVD; email	
Archival Storage and Preservation		
Size of current archive in TB	0.9 TB	
Size of current archive in other means (collections, files, etc.)	647,8492	
Maximal deposit size in TB	No maximal size	
Long-term guarantees/standards of trust	Data Seal of Approval (ongoing process, we will apply for DSA in 2014)	
Checks on quality/quality control	access control; check sums; audit trials	
Dissemination		
Costs/conditions for access	Free	
Tools/Interfaces used for access	Nestar	http://nesstar.soc.cas.cz
ESFRI	CESSDA; ESS	
Name of the archive	CSDA (Czech Social Science Data Archive)	

FORS - DARIS (Data and Research Information Services)

Functionalities	Short Description	References
	Administrative Context	
Funding	Public; third party	
Depositor Agreements	See reference	http://www2.unil.ch/daris/IMG/pdf/end-user_contract_quanti_normal.pdf
Usage Agreements, Code of Conduct to be signed	See reference	http://www2.unil.ch/daris/IMG/pdf/end-user_contract_quanti_normal.pdf
Policies in Place	No	
Rights on Data Claimed by the Archive	Non-exclusive rights granted	
Data Curation Strategy	Migration; bi stream preservation	
	Pre-ingest	
Primary Community in Focus for Deposits	Sociologists	
Secondary Communities accepted for Deposits	Political science	
	Ingest	
Formats accepted and curated	List of preferred formats	http://www2.unil.ch/daris/IMG/pdf/Guide_to_depositing_quantitative_data_at_FORS.pdf
Formats accepted and not curated	-	
Metadata formats accepted	DDI	
User-based ingest	Mail; DVD/CD	
	Archival Storage and Preservation	
Size of current archive in TB	0.06 TB	
Size of current archive in other means (collections, files, etc.)	500 datasets and 2000 data files	
Maximal deposit size in TB	No	
Long-term guarantees/standards of trust	Self-assessment	
Checks on quality/quality control	Access controls	
	Dissemination	
Costs/conditions for access	No costs; for research purposes only	
Tools/Interfaces used for access	Nesstar (for a small subset of studies); online data catalogue (for all studies with data)	
ESFRI	CESSDA; ESS; SHARE	
Name of the archive	FORS - DARIS (Swiss Centre of Expertise in the Social Sciences - Data and Research Information Services)	

DRI (Digital Repository of Ireland)

Functionalities	Short Description	References
	Administrative Context	
Funding	Public funding	
Depositor Agreements	Under development	
Usage Agreements, Code of Conduct to be signed	Under development	
Policies in Place	Under development	
Rights on Data Claimed by the Archive	Non-exclusive rights to distribute to the archive	
Data Curation Strategy	Other (not elaborated)	
	Pre-ingest	
Primary Community in Focus for Deposits	Humanities and social scientists	
Secondary Communities accepted for Deposits	Cultural institutions and libraries	
	Ingest	
Formats accepted and curated	List of preferred formats (under development)	
Formats accepted and not curated	-	
Metadata formats accepted	DC; EAD; MARC; MODS	
User-based ingest	Under development	
	Archival Storage and Preservation	
Size of current archive in TB	-	
Size of current archive in other means (collections, files, etc.)	-	
Maximal deposit size in TB	-	
Long-term guarantees/standards of trust	-	
Checks on quality/quality control	Access control; check sums; audit trials	
	Dissemination	
Costs/conditions for access	Currently no costs for access	
Tools/Interfaces used for access	Under developments	
ESFRI	CESSDA; DARIAH	
Name of the archive	Digital Repository of Ireland (DRI)	

IQDA (Irish Qualitative Data Archive)

Functionalities	Short Description	References
	Administrative Context	
Funding	Public; third party	
Depositor Agreements	Depositor Agreement is downloadable from website. Signed copy must be returned to IQDA. Signed Depositor Agreement must be submitted for every deposit.	http://www.iqda.ie/content/deposit-data
Usage Agreements, Code of Conduct to be signed	Data Access Request Form is downloadable from website. Signed copy must be returned to IQDA. All proposed users must sign Data Access Request Form. If the applicant is a student, his/her supervisor must apply as the Lead User.	http://www.iqda.ie/content/access-data
Policies in Place	Copyright, Access Conditions and Acceptance of Terms and Conditions are included in the License Agreement	http://www.iqda.ie/content/deposit-data
Rights on Data Claimed by the Archive	The depositor or the depositing body retain the ownership and copyright to the dataset and related material, unless otherwise stated in the Depositor Agreement.	
Data Curation Strategy	None at present	
	Pre-ingest	
Primary Community in Focus for Deposits	IQDA archives qualitative social science data generated in or about Ireland. IQDA archives non-numerical data from projects conducted by researchers and students at publicly funded higher education institutions, publicly funded research institutes and state and semi-state bodies.	
Secondary Communities accepted for Deposits	-	
	Ingest	
Formats accepted and curated	Preferred audio formats: Free Lossless Audio Codec (FLAC) (.flac), WAV file (.wav) Accepted audio formats: MPEG-1 Audio Layer 3 (MP3), Audio Interchange File Format (.aiff) Preferred text formats: Rich Text Format (.rtf), Plain text data, ASCII (.txt), eXtensible Markup Language (XML) marked-up text according to an appropriate Document Type Definition (DTD) or schema Accepted text formats: Hypertext Markup Language (HTML), widely-used proprietary formats e.g. Microsoft Word (.doc/.docx), Proprietary/software-specific formats such as MAXQDA, NUD*IST, NVivo and ATLAS. Preferred image formats: TIFF (version 6) uncompressed Accepted image formats: JPEG (.jpeg, .jpg), TIFF (other versions), Adobe Portable Document Format (PDF/A or PDF), raw image format (.RAW)	http://www.iqda.ie/content/deposit-data
Formats accepted and not curated	None	
Metadata formats accepted	DDI; DC	
User-based ingest	Deposit form	http://www.iqda.ie/content/deposit-data
	Archival Storage and Preservation	
Size of current archive in TB	very small	
Size of current archive in other means (collections, files, etc.)	10 dataset; 500 data files (not including photographs)	
Maximal deposit size in TB	No	
Long-term guarantees/standards of trust	No	
Checks on quality/quality control	Access controls; audit trails	
	Dissemination	
Costs/conditions for access	No costs; Access restricted to bonafide researchers	
Tools/Interfaces used for access	Searchable catalogue, map based search	
ESFRI	None at present	
Name of the archive	IQDA (Irish Qualitative Data Archive)	

ODSAS (Online Digital Sources and Annotation System)

Functionalities	Short Description	References
	Administrative Context	
Funding	The main allocations come from CNRS (Centre National de la Recherche Scientifique), French research public organization through digital archives or research projects	
Depositor Agreements	Yes	
Usage Agreements, Code of Conduct to be signed	The depositor gives specific rights to users	
Policies in Place	Yes	
Rights on Data Claimed by the Archive	The ODSAS software is on creative common licence and the archives are each depositor can chose about the rights on the archives	
Data Curation Strategy	Migration	
	Pre-ingest	
Primary Community in Focus for Deposits	Sociologists, anthropologists, linguistics	
Secondary Communities accepted for Deposits	Social sciences researchers	
	Ingest	
Formats accepted and curated	List of accepted formats	http://odsas.fr/index.php?action=documentation
Formats accepted and not curated		
Metadata formats accepted	DC; RDF	
User-based ingest	CD/DVD; mail; external hard disk; USB-stick	
	Archival Storage and Preservation	
Size of current archive in TB	100 TB	
Size of current archive in other means (collections, files, etc.)	125000 files (texts and pictures), 1000 movies and 1500 sound recordings	
Maximal deposit size in TB	No	
Long-term guarantees/standards of trust	No	
Checks on quality/quality control	Access controls	
	Dissemination	
Costs/conditions for access	Some resources are open access and some have a restricted access. There's no cost for access to the data but not everybody can have an access. You have to be authorized by the author/researcher.	
Tools/Interfaces used for access	To ask for an access to restricted collections, you have to complete an online form	http://www.odsas.net/index.php?action=create_account
ESFRI	CLARIN; DARIAH	
Name of the archive	ODSAS (Online Digital Sources and Annotation System)	

Oxford Text Archive

Functionalities	Short Description	References
	Administrative Context	
Funding	The Oxford Text Archive is funded as a core activity of the University of Oxford IT Services, with some contribution from externally funded projects.	
Depositor Agreements	Yes, see reference	http://www.ota.ox.ac.uk/documents/user_agreement.xml
Usage Agreements, Code of Conduct to be signed	Yes, see reference	http://www.ota.ox.ac.uk/documents/user_agreement.xml
Policies in Place	Yes	
Rights on Data Claimed by the Archive	Non-exclusive rights granted to the OTA by the depositor.	
Data Curation Strategy	Bit stream preservation and migration where this is possible. Depositors are also encouraged to offer updated versions of the resource to the archive.	
	Pre-ingest	
Primary Community in Focus for Deposits	General humanities research, corpus linguistics researchers, language leaners and the general public	
Secondary Communities accepted for Deposits	-	
	Ingest	
Formats accepted and curated	Any format	
Formats accepted and not curated	-	
Metadata formats accepted	TEI	
User-based ingest	Mail	
	Archival Storage and Preservation	
Size of current archive in TB	1 TB	
Size of current archive in other means (collections, files, etc.)	5,000 datasets and 140,000 data files	
Maximal deposit size in TB	No	
Long-term guarantees/standards of trust	Service Level Definition (SLD)	http://www.oucs.ox.ac.uk/internal/sld/ota.xml
Checks on quality/quality control	Access controls	
	Dissemination	
Costs/conditions for access	None	
Tools/Interfaces used for access	Http downloads. Textual analysis tools under development with the CLARIN research infrastructure. The resources are made available for download only. Our aim is to provide direct access to the content at persistent locations so that others can build interfaces and access services at other locations.	
ESFRI	CLARIN	
Name of the archive	Oxford Text Archive	

Réseau Quetelet

Functionalities	Short Description	References
	Administrative Context	
Funding	Public funding from CNRS (National centre for Scientific Research), Ministry Research infrastructure budget and Universities involved in the Consortium	
Depositor Agreements	Contracts with data providers	http://www.reseau-quetelet.cnrs.fr/spip/rubrique.php3?id_rubrique=68&lang=en
Usage Agreements, Code of Conduct to be signed	Licences signed by individual user and the research institution (at the level of the research department)	http://www.reseau-quetelet.cnrs.fr/spip/rubrique.php3?id_rubrique=67&lang=en
Policies in Place	Acceptance of terms and conditions (research only, best practices)	http://www.reseau-quetelet.cnrs.fr/spip/rubrique.php3?id_rubrique=67&lang=en
Rights on Data Claimed by the Archive	Granting non-exclusive rights	
Data Curation Strategy	Bit stream preservations	
	Pre-ingest	
Primary Community in Focus for Deposits	Official microdata and research surveys under public funding	
Secondary Communities accepted for Deposits	Poll institutes	
	Ingest	
Formats accepted and curated	All formats accepted	http://www.reseau-quetelet.cnrs.fr/spip/article.php3?id_article=3&lang=en
Formats accepted and not curated	-	
Metadata formats accepted	DDI	
User-based ingest	Email - CD/DVD	
	Archival Storage and Preservation	
Size of current archive in TB	Currently not available	
Size of current archive in other means (collections, files, etc.)	1100 references + Secure remote access centre holding confidential files	
Maximal deposit size in TB	-	
Long-term guarantees/standards of trust	No	
Checks on quality/quality control	Manual checks and quality controls ; Check sums	
	Dissemination	
Costs/conditions for access	All Scientific files distributed within the frame of the Réseau Quetelet are distributed free of charge for use in research. Users accredited by a specific authority for access to Scientific Confidential Files form for official microdata (highly detailed) accessible via the secure remote access centre CASD are charged for the time used.	http://www.reseau-quetelet.cnrs.fr/spip/rubrique.php3?id_rubrique=67&lang=en
Tools/Interfaces used for access	Access by a web application on Reseau Quetelet's website. Access by a secure remote box for CASD.	http://www.reseau-quetelet.cnrs.fr/spip/rubrique.php3?id_rubrique=75&lang=en
ESFRI	CESSDA; ESS; SHARE	
Name of the archive	Réseau Quetelet	

RODA (Romanian Social Data Archive)

Functionalities	Short Description	References
	Administrative Context	
Funding	Public Funding	
Depositor Agreements	Web-based Licence Agreement	http://www.roda.ro/documente/LicenceAgreement.pdf
Usage Agreements, Code of Conduct to be signed	The user has to sign a one-time only user agreement when registering at RODA. For each particular application, if the data is available for scientific research it's going to be directly available for download, with the exception of special requirements specifically asked by the data owners.	http://www.roda.ro/documente/Individual.pdf
Policies in Place	Acceptance of Terms and Conditions, various access conditions depending on the combination of a) access level of the dataset, b) type of user (e.g. whether academic or not) and c) nature of the research project (e.g. an academic user working for a commercial project is treated like a commercial user)	http://www.roda.ro/documente/DataPreservation.pdf
Rights on Data Claimed by the Archive	Data owners retain all rights, RODA being only an intermediate body between them and the users.	
Data Curation Strategy	RODA is currently in the process of completely redrafting the whole procedure, by the end of 2014	
	Pre-ingest	
Primary Community in Focus for Deposits	Social scientists, data generated from a scientific research project.	
Secondary Communities accepted for Deposits	Official microdata produced by the National Institute of Statistics and other official bodies (e.g. Local administration)	
	Ingest	
Formats accepted and curated	Any rectangular format for quantitative data.	
Formats accepted and not curated	Any kind of format, provided sufficient comprehensible metadata that allows usage.	
Metadata formats accepted	DDI, web-based self-completion plus fine-tune by the archive employees.	
User-based ingest	Data deposit forms, self-depositing tools.	
	Archival Storage and Preservation	
Size of current archive in TB	0.1 TB	
Size of current archive in other means (collections, files, etc.)	100 datasets	
Maximal deposit size in TB	no maximal deposit size	
Long-term guarantees/standards of trust	Data Seal of Approval (process started)	
Checks on quality/quality control	access control; check sums; audit trials	
	Dissemination	
Costs/conditions for access	Free for academic research, other costs specified by the data owners.	
Tools/Interfaces used for access	Under construction, web-based interface that will be released under the open-source.	
ESFRI	CESSDA	
Name of the archive	RODA	

SLDR (Speech and Language Data Repository)

Functionalities	Short Description	References
	Administrative Context	
Funding	Public (Research/Education Govt agencies). SLDR is in the process of merging with other services in the ORTOLANG network infrastructure. SLDR/ORTOLANG is currently interacting with Huma-Num VLRI, CINES and CC-IN2P3.	http://www.ortolang.fr http://huma-num.fr http://www.cines.fr http://cc.in2p3.fr
Depositor Agreements	Depositors agree with Producer's licence certifying ownership of all rights on deposited data.	http://sldr.org/wiki/Licences_en
Usage Agreements, Code of Conduct to be signed	Users agree with User's licence and, whenever necessary, a complimentary licence designed by the producer to regulate specific usage of the data.	http://sldr.org/wiki/Licences_en http://sldr.org/wiki/accessRightsManagement_en
Policies in Place	Related CINES archival service is a Trusted Digital repository compliant with Data Seal of Approval.	http://sldr.org/wiki/DSA
Rights on Data Claimed by the Archive	Depositor retains all rights	
Data Curation Strategy	1) Indexing item and assigning persistent identifiers to all documents (by default); 2) Validating file formats and contents for long-term preservation; 3) Migrating item to the dissemination site; 4) Submitting item to the CINES archival service if long-term preservation is planned; 5) Versioning item to take into account modifications of data and/or documentary files.	http://sldr.org/wiki/Handle_en
	Pre-ingest	
Primary Community in Focus for Deposits	Linguists	
Secondary Communities accepted for Deposits	Scholars working on data associated with acts of language, i.e. speaking, singing, gesture, writing, reading Medical research on speech pathologies, communication deficit and prevention of neurological diseases Non-profit organisations supporting cultural heritage, e.g. endangered regional languages	
	Ingest	
Formats accepted and curated	List of formats supported by CINES archival service	http://sldr.org/wiki/Formats
Formats accepted and not curated	Any format applicable to video/audio streaming Proprietary formats produced by hardware performing measurements of speech production and perception	
Metadata formats accepted	Dublin Core (OAI-DC, OLAC), RDFa, CMDI	
User-based ingest	Mail, CD/DVD, ftp transfer	
	Archival Storage and Preservation	
Size of current archive in TB	2 TB	
Size of current archive in other means (collections, files, etc.)	284 data sets and 345,000 data files	
Maximal deposit size in TB	No maximal deposit size, No restriction on file hierarchy File names encoded in Unicode UTF8	
Long-term guarantees/standards of trust	Self-assessment	
Checks on quality/quality control	Access control and check sums	
	Dissemination	
Costs/conditions for access	Access and downloading are free of charge. Links may be provided with LDC and ELRA for paid-basis access. Access may be restricted in compliance with the French Heritage Code. Access may be granted to individuals or institutions by way of shared licences.	http://hdl.handle.net/11041/sldr000034 (example) http://sldr.org/wiki/accessRightsManagement_en http://sldr.org/wiki/table_derogations_en http://sldr.org/wiki/SharedLicence
Tools/Interfaces used for access	Standard web browsers. Multilingual navigation and interactions: English, Spanish, French, Chinese.	
ESFRI	CLARIN; DARIAH	http://clarin.eu http://dariah.eu
Name of the archive	Speech and Language Data Repository (SLDR	http://sldr.org

TARKI (Hungary Social Research Institute)

Functionalities	Short Description	References
	Administrative Context	
Funding	The TARKI Data Archive receives its funding from the TARKI Foundation as well as from grants and research contracts.	http://www.tarki.hu/en/services/da/da_description.html
Depositor Agreements	The database deposit form is available on the TARKI Data Archive page	http://www.tarki.hu/hu/services/da/da_use.html
Usage Agreements, Code of Conduct to be signed	Access to the Data Archive is free for university students, teachers and researchers. The user can download the user declaration form from the Data Archive page, this verifies the eligibility for the free data access. Within one or two days after the application is arrived, the user will receive the datasets by e mail.	http://www.tarki.hu/en/services/da/docs/user_declaration.pdf
Policies in Place	Copyright and access conditions are included in the database depositor form	
Rights on Data Claimed by the Archive	The depositors determine the data access category - that is, the dissemination rules we apply to their data collections.	http://www.tarki.hu/hu/services/da/docs/adatbank_leteti_nyilatkozat.pdf
Data Curation Strategy	During archiving we create a metadata sheet based on DDI standards for each data collection. Datasheets are stored in an SQL database, which serves our catalogue and the National Digital Archive's catalogue. We distribute our collections, according to the dissemination policy determined by depositors, to anyone who is interested in social research.	http://www.tarki.hu/en/services/da/da_description.html
	Pre-ingest	
Primary Community in Focus for Deposits	The Data Archive has collected and archived more than 750 empirical social research data collections that are suitable for secondary analysis. These tend to be Hungarian. Most of our collection comes from nationally representative sample-survey studies (i.e. micro datafiles). One section of the databases archived is made up of TARKI's own surveys, and the other section comprises surveys from other Hungarian research institutes.	
Secondary Communities accepted for Deposits	The mission of our archive is to provide infrastructure service, and support for all stakeholders in social research.	
	Ingest	
Formats accepted and curated	The data collections contain numerical data.	
Formats accepted and not curated	The data collections contain numerical data.	
Metadata formats accepted	DDI, Dublin Core	
User-based ingest	Database deposit form	
	Archival Storage and Preservation	
Size of current archive in TB	0.01 TB	
Size of current archive in other means (collections, files, etc.)	Including administrative data, documentations: approx. 2.6 GB	
Maximal deposit size in TB	Not specified, evaluated for each new depositor	
Long-term guarantees/standards of trust	Mission statement of TÁRKI Data Archive contains this guarantee	http://www.tarki.hu/en/services/da/da_description.html
Checks on quality/quality control	Manual checks and quality controls are part of the archiving process	
	Dissemination	
Costs/conditions for access	Access to the Data Archive is free for academic or research purposes	
Tools/Interfaces used for access	Data Archive catalogue search /SPSS, Stata	http://www.tarki.hu/en/help/search.html
ESFRI	CESSDA; ESS	
Name of the archive	Tarki	

TextGrid Repository

Functionalities	Short Description	References
	Administrative Context	
Funding	Third party (BMBF – German Ministry of Education and Research)	
Depositor Agreements	See reference	http://www.textgrid.de/en/registrierungdownload/tou/
Usage Agreements, Code of Conduct to be signed	See reference	http://www.textgrid.de/en/registrierungdownload/tou/
Policies in Place	See reference	http://textgrid.de/en/about-textgrid/project/
Rights on Data Claimed by the Archive	No rights obtained by the archive	
Data Curation Strategy	Bitstream preservation; migration	
	Pre-ingest	
Primary Community in Focus for Deposits	Scholars (Humanists)	
Secondary Communities accepted for Deposits	-	
	Ingest	
Formats accepted and curated	XML (TEI), JPEG, TIF (preferred formats)	http://textgridlab.org/schema/textgrid-metadata_2010.xsd
Formats accepted and not curated	All other formats	
Metadata formats accepted	DC, TEI, TextGrid Metadata Schema	
User-based ingest	VRE import tool (TextGridLab), external import Tool (koLibRI)	
	Archival Storage and Preservation	
Size of current archive in TB	2.5 TB	
Size of current archive in other means (collections, files, etc.)	900 datasets and 700000 data files	
Maximal deposit size in TB	None	
Long-term guarantees/standards of trust	DSA in preparation	
Checks on quality/quality control	Access controls	
	Dissemination	
Costs/conditions for access	None	
Tools/Interfaces used for access	TextGridLab, web browser, OAI-PMH, REST/SOAP APIs, web based tools (such as Digivoy, CollateX)	
ESFRI	DARIAH	
Name of the archive	TextGrid Repository	http://textgridrep.de, http://textgrid.de/en

Stichting Nederlands Instituut voor Beeld & Geluid

Functionalities	Short Description	References
	Administrative Context	
Funding	Government funding added to by project funding	
Depositor Agreements	Yes	
Usage Agreements, Code of Conduct to be signed	Yes	
Policies in Place	Yes	
Rights on Data Claimed by the Archive	Right to store, catalogue, preserve and make available according to agreed conditions	
Data Curation Strategy	-	
	Pre-ingest	
Primary Community in Focus for Deposits	Public broadcast media organisations	
Secondary Communities accepted for Deposits	Heritage institutions, commercial media institutions	
	Ingest	
Formats accepted and curated	List of preferred formats	
Formats accepted and not curated	Analogue materials	
Metadata formats accepted	As been agreed in Depositor Agreements ; Dublin Core ; Metadata according to proprietary metadata model based on FRBR	
User-based ingest	Broadcast production: via a direct interface with the archive's repository and catalogue. Other depositors: via a digital file importing system	
	Archival Storage and Preservation	
Size of current archive in TB	7,000 TB	
Size of current archive in other means (collections, files, etc.)	850.000 hours of analogue, digitized and digital born AV-materials	
Maximal deposit size in TB	No	
Long-term guarantees/standards of trust	Requirements for OAIS compliancy formulated (policies, workflow, object management etc.) to be able to apply for a basic level DSA in 2014	
Checks on quality/quality control	Access controls; check sums; audit trials; integrity monitoring	
	Dissemination	
Costs/conditions for access	No charge for accessing the metadata/catalogue, variable costs for downloading/streaming high or low res materials depending on the user category (media professionals, commercial parties, educational users, general public).	
Tools/Interfaces used for access	IMMIX catalogue accessible via extranet for media professionals and via internet for the general public	
ESFRI	CLARIN	
Name of the archive	Stichting Nederlands Instituut voor Beeld & Geluid	

UFAL (Institute of Formal and Applied Linguistics)

Functionalities	Short Description	References
	Administrative Context	
Funding	Public funding by Ministry of Education, Sports, and Youth under the 'Large Infrastructures' Programme (LINDAT/CLARIN project)	
Depositor Agreements	Yes	https://lindat.mff.cuni.cz/repository/xmlui/page/about#about-contracts
Usage Agreements, Code of Conduct to be signed	Yes	https://lindat.mff.cuni.cz/repository/xmlui/page/about#about-contracts
Policies in Place	Yes	https://lindat.mff.cuni.cz/repository/xmlui/page/about
Rights on Data Claimed by the Archive	The depositor retains all rights to the data but the archive is granted non-exclusive rights to make copies and to translate the submission (metadata) without changing content.	
Data Curation Strategy	Migration; bit stream preservation; automatic and manual curation framework for reviewers in the repository	
	Pre-ingest	
Primary Community in Focus for Deposits	Linguists	
Secondary Communities accepted for Deposits	Everybody with academic credentials (through several national and international federations (SPF, eduGAIN)) who is somehow related to linguistics	
	Ingest	
Formats accepted and curated	Preferred formats on a list, but other formats accepted as well	only available during the deposit-process
Formats accepted and not curated	-	
Metadata formats accepted	CMDI;DC; METS	
User-based ingest	online upload	
	Archival Storage and Preservation	
Size of current archive in TB	0.1 TB	
Size of current archive in other means (collections, files, etc.)	72 datasets	
Maximal deposit size in TB	No	
Long-term guarantees/standards of trust	DSA	
Checks on quality/quality control	Access control; check sums; Preservation on different levels (VM, OS, HDD, submissions), syncing submissions to two external data houses.	
	Dissemination	
Costs/conditions for access	Access to metadata is free, see metadata policy. Access to data depends on the licence the data is associated with but mostly we have CC* licences.	
Tools/Interfaces used for access	-	
ESFRI	CLARIN	
Name of the archive	UFAL	

SASD (Slovak Archive of Social Data)

Functionalities	Short Description	References
	Administrative Context	
Funding	Public	
Depositor Agreements	Yes	http://sasd.sav.sk/en/rtf/sasd_data_provision_agreement.rtf
Usage Agreements, Code of Conduct to be signed	Yes	http://sasd.sav.sk/en/pristup_formular.php
Policies in Place	No	
Rights on Data Claimed by the Archive	Non-exclusive rights granted	
Data Curation Strategy	Other (not specified)	
	Pre-ingest	
Primary Community in Focus for Deposits	Sociologists	
Secondary Communities accepted for Deposits	-	
	Ingest	
Formats accepted and curated	Any format	
Formats accepted and not curated	-	
Metadata formats accepted	DDI	
User-based ingest	Mail; DVD/CD; personal delivery	
	Archival Storage and Preservation	
Size of current archive in TB	0.001 TB	
Size of current archive in other means (collections, files, etc.)	30 datasets, cca 180 files (i.d. xml documentation, PDF questionnaire, data files)	
Maximal deposit size in TB	No	
Long-term guarantees/standards of trust	No	
Checks on quality/quality control	Data are stored as ZIP files, no other measures protecting the authenticity and integrity are taken	
	Dissemination	
Costs/conditions for access	No costs; access after registration (http://sasd.sav.sk/en/pristup_kategorie.php)	
Tools/Interfaces used for access	SASD adapted DDI viewer; possibility to review the DDI itself; data distributed in SPSS or SPSS and xls files	
ESFRI	CESSDA	
Name of the archive	SASD	

DDA (Danish Data Archive)

Functionalities	Short Description	References
	Administrative Context	
Funding	Public funding from Ministry of Culture and Ministry of Higher Education and Science	
Depositor Agreements	Web-based deposit form	http://samfund.dda.dk/dda/data-aflevere-en.asp
Usage Agreements, Code of Conduct to be signed	The depositor states in the deposit form if there must be any restriction to re-use of data. All metadata is open access.	http://samfund.dda.dk/dda/data-aflevere-en.asp
Policies in Place	Access conditions and acceptance of terms and conditions are included in the deposit form.	http://samfund.dda.dk/dda/data-aflevere-en.asp
Rights on Data Claimed by the Archive	Non-exclusive rights granted to the archive, creative commons agreement for metadata	http://samfund.dda.dk/dda/om-dda-en.asp
Data Curation Strategy	Data and metadata is processed to DDI-Lifecycle format.	http://samfund.dda.dk/dda/om-dda-en.asp
	Pre-ingest	
Primary Community in Focus for Deposits	Social and health science research and education	http://samfund.dda.dk/dda/om-dda-en.asp
Secondary Communities accepted for Deposits	Health science and research	http://samfund.dda.dk/dda/om-dda-en.asp
	Ingest	
Formats accepted and curated	Preferred formats are SPSS, SAS and STATA but other formats accepted as well	http://samfund.dda.dk/dda/data-aflevere-en.asp
Formats accepted and not curated	Text formats	
Metadata formats accepted	DDI	
User-based ingest	Deposit form	http://samfund.dda.dk/dda/data-aflevere-en.asp
	Archival Storage and Preservation	
Size of current archive in TB	3 TB	
Size of current archive in other means (collections, files, etc.)	2000 datasets	
Maximal deposit size in TB	Not specified	
Long-term guarantees/standards of trust	Peer-reviewed Data Seal of Approval	
Checks on quality/quality control	Manual checks and quality controls, check sums	
	Dissemination	
Costs/conditions for access	No costs, open access and restricted access according to statement in deposit form	http://samfund.dda.dk/dda/data-bestille-en.asp
Tools/Interfaces used for access	DDI-Lifecycle based search service/indexing platform, Nesstar	http://samfund.dda.dk/dda/data-bestille-en.asp
ESFRI	CESSDA	
Name of the archive	DDA	

DTARe (Deutsches Textarchiv Repository)

Functionalities	Short Description	References
	Administrative Context	
Funding	Public Funding plus funding from BMBF and BBAW	
Depositor Agreements	dependent on data type and rights	
Usage Agreements, Code of Conduct to be signed	dependent on data type and rights	
Policies in Place	DSA	https://assessment.datasealofapproval.org/assessment_87/seal/pdf/
Rights on Data Claimed by the Archive	non-exclusive rights granted to the archive	
Data Curation Strategy	Migration and bitstream preservation	
	Pre-ingest	
Primary Community in Focus for Deposits	Linguists	
Secondary Communities accepted for Deposits	literary scholars, historians, social sciences, legal scholars, economists, …	
	Ingest	
Formats accepted and curated	Only formats on the list of accepted formats, cf. DSA and CLARIN's "Standards for LRT" document	https://assessment.datasealofapproval.org/assessment_87/seal/pdf/, http://www.clarin.eu/node/2320
Formats accepted and not curated	-	
Metadata formats accepted	CMDI, DC, TEI	
User-based ingest	-	
	Archival Storage and Preservation	
Size of current archive in TB	0.01 TB	
Size of current archive in other means (collections, files, etc.)	1300 data sets containing 5200 files	
Maximal deposit size in TB	No maximal deposit size	
Long-term guarantees/standards of trust	Peer-reviewed DSA self-assessment	
Checks on quality/quality control	Access controls, check sums and back-ups	
	Dissemination	
Costs/conditions for access	Cf. DSA	https://assessment.datasealofapproval.org/assessment_87/seal/pdf/
Tools/Interfaces used for access	Fedora, Web Interface	http://clarin.bbaw.de+
ESFRI	CLARIN-D	
Name of the archive	CLARIN-D Service Center at the BBAW/Deutsches Textarchiv Repository (DTARe)	