E Research in the Humanities and Social Sciences

Arjan Hogenaar, Heiko Tjalsma and Mike Priddy

1 Introduction

The social sciences and the humanities taken together contain a heterogeneous range of research disciplines. Almost all existing methods of research can be found within these two domains. Data handling (collecting, processing, selecting, preserving) and publication methods differ greatly. Attitudes in the field towards Open Access of publications as well to research data vary as well.

It is not possible to cover the total fullness, and complexity, of all the disciplines within these two domains. Our observations will therefore be based upon a number of case studies. Taken together these case studies give a fairly representative picture of the domains, at least of the most common research environments. The main dividing line is between those disciplines creating empirical data, such as survey data in the social sciences and those, especially in the humanities, using existing source material, such as history or text studies. This source material can either be of an analogous or a digital nature. As will be shown in the case studies in many disciplines a mix of created and existing is often combined.

The Data Archiving and Networked Services (DANS[1]) has been chosen as an exemplar within the area of social science and the humanities. DANS promotes sustained access to digital research data. For this purpose, DANS has created the online archiving system EASY[2] which enables researchers to archive and re-use data in a sustained manner, primarily in the social sciences and the humanities. It is expected that this will be extended to other disciplines in the future. In addition, the institute provides training and advice and undertakes research into sustained access to digital information.

[1] http://www.dans.knaw.nl.
[2] http://www.easy.dans.knaw.nl.

Through its activities, DANS is in close contact with a number of researchers in the two domains of this study. The findings in this section are based on interviews with a selection of these. Care has been taken that this selection was as representative as possible for these heterogeneous disciplines. Interviews of approximately 1 hour each were conducted with a range of researchers from both within DANS and with researchers from other institutions that have close ties with DANS: either collaborators on projects or who are using or depositing data in the online archiving system EASY. The interviews were semi-structured with a list of questions and subquestions, but if it was clear that certain groups of questions were not relevant to the interviewee these were not asked. The majority of interviews were conducted with scholars who would identify themselves as working in the Humanities or with humanities data. This emphasis was because DANS had recently conducted similar interviews on usage of digital data and research infrastructures with senior researchers, where the bias was towards the social sciences. A total of 15 interviews were conducted, nine with humanities researchers and six with social scientists. Even with such a small sample of interviews, we attempted to get a broad cross-section of disciplines; however, within archaeology we conducted three interviews to get a deeper insight into one discipline.

The interviews conducted with senior academics and research managers, mostly professors and/or directors of research institutes, occurred in summer/autumn 2010 and in spring 2011. This set of interviews formed part of a strategic plan on widening the scope of social science and humanities disciplines utilizing the services of DANS.

In most interviews, the need for data preservation and (open) data access, as experienced in the specific discipline, were discussed in a very broad sense. The interviews carried out within the humanities and social sciences are of particular importance for the OpenAIRE Project, as they give insight in how researchers within these fields deal with open access to data and publications. They focused on a limited number of fields: economics/econometrics, finances, sociology (survey research) and law.

Furthermore an online survey was used which was carried out into data usage, data archiving and research infrastructures amongst researchers from all disciplines in the Netherlands.

2 Workflows in social sciences and humanities research

2.1 Phases in social science research

Discovery and planning Starting from a theoretical and empirical perspective, the researcher first wants to extend his or her knowledge. The researcher shall need to explore what data will be required to give the best answers to the scientific questions of his or her research project: are there existing (archived) data available or should new data be collected?

Initial data collection This is the phase in which data are actually collected. This could be in the form of a survey held or an experiment carried out, or the acquisition of previously collected data, possibly restructured or linked to other datasets, may form the foundation of the data collection. Essential data management strategies are formulated and executed, including decisions about documentation content and formats.

Final data preparation and analysis In this third phase, the researcher undertakes analysis after having performed final verification and modification of the data. The process of data preparation should be is complete and results are written up.

Publication and sharing In the fourth phase, the researcher will communicate the research findings in publications.

Long-term data management In this final phase, there are two critical goals, seen from the perspective of the wider social science community: providing access to the data and ensuring long-term preservation. Once the data are available for secondary use, they have reached the final stage of the research cycle and could become the start of new projects that begin with their discovery and re-use thus beginning the cycle anew.[3]

2.2 The lifecycle in the humanities

Because of the heterogeneity within the humanities, this example describes historical or textual research in general, but applies less to other domains within humanities, such as archeology.

[3] Green, AG and Gutmann MP. *Building partnerships among social science researchers, institution-based repositories and domain specific data archives.* OCLC Systems and Services: International Digital Library Perspectives 2007, 23, 35–53.

Creation In this first phase, the design of the information structure has to be made through data modeling or text modeling, based on the goals and design of the research project. It also includes the physical production of the digital data either by data entry and text entry tools, or by digitization (optical character recognition) of existing analogous resources.

Enrichment The raw data, in whatever format (images, texts, databases, GIS-files) will have to be enriched with metadata, describing the historical information in more detail and, in particular, the context and provenance of it. Preferably this should be done in a standardized way (Dublin Core for example), but in practice this mostly not the case.

Editing Editing includes the actual encoding of textual information by inserting mark-up tags or entering data in the fields of database records. Enhancement could be considered as a separate phase of the editing process by which data are being transformed. This stage could also include annotating original data with background information, bibliographical references and links to related passages.

Retrieval In this phase, the information should be ready to be selected (by queries), looked up and used (i.e. retrieval). Results of this process should be displayed, possibly in a more advanced visualized representation.

Analysis Analysing information can refer to various activities in historical research, due to the varying methodologies used, ranging from quantitative analysis, using advanced statistical methods, to qualitative descriptions.

Presentation Various forms of presentation are used in the historical sciences, and the humanities generally. Presentation of results could also take place in earlier stages as well.

Presentation of digital historical information may also take quite different forms, varying from electronic text editions, online databases and virtual exhibitions to small-scale visualizations of research results.[4]

Long-term data management Of course, within the humanities, as in the social sciences, the data that are the results of the research should be stored for access and re-use as well, but also long-term preservation should be ensured.

[4] Boonstra, O, Breure, L, and Doorn, P. *Past, present and future of historical information science*. NIWI-KNAW, 2006. pp. 21–23. Available at http://www.dans.knaw.nl/sites/default/files/file/publicaties/Past-present.pdf.

3 Case studies

3.1 Archaeology

The Archaeology interviews were with: (i) an established university-based researcher; (ii) an early career researcher who re-used data; and (iii) a senior-researcher who conducted excavations on behalf of the municipality (local governmental archaeological agency). The three archaeologists had backgrounds in Neolithic prehistory of the Netherlands, Bronze and Iron Age of East Netherlands, and Roman and Medieval history of The Hague. All three researchers continue to work in these areas, although approaches to data and literature do vary. In this archaeology case study, quotes from these interviewees are printed in italic.

For archaeologists it is normal that an excavation produces digital and analogue data, as well as finds that may require digitization. The data collected may be compared to other excavations recorded in the digital literature archives. Additionally digital data may come from other organizations, for example elevation data, which is often used in archaeology. One of the interviewed archaeologists will include GIS and elevation in a database for the specific project. However, the gathering of other digital sources, as well collecting own resources, is often for personal research needs. "*A lot of archaeologists don't use this [digital] data as heavily as I do. Some will and there are some who still don't even create digital data.*" Archaeology is a diverse field and so is the use of digital data. None of the archaeologists interviewed use standardized digital tools, or workflows "*because it is too diverse, the creation of data*" There are a number of key sources used to start the process of gathering data. For one archaeologist, "*part of job is to look at new methodologies, new visualization, new tools.*"

One archaeologist's research is about the habitation development along a river valley, with a focus on 12 Roman sites, but using existing data from digs that took place between 1960 and 1990 and were not studied and published before.[5] Because of the period the original excavations were conducted, only a few files and images are digital. "*Bringing the old data into the digital domain is crucial and very important to this scientific community.*"

Another archaeological senior researcher is also digitizing unstudied work of past excavations.[6] The first excavations were in the 1930s, the researcher was involved in later excavations in the 1980s and 1990s and "*is now digitizing the memory I have in my head*". DANS has funded the scanning of the field

[5] Subsidized by the Odyssee programme that supports the publication of previously unpublished archaeological excavations.

[6] "Den Haag Ockenburgh: een fortificatie als onderdeel van de Romeinse kustverdediging"; also funded by the Odyssee programme.

drawings of the 1930s, new interpretations are made and misinterpretations from the 1930s are corrected. The digital databases from the later excavations are in a number of different formats "*that were not quite good enough.*" This data is now being enhanced and assessed, but updating the database systems is difficult due to organizational planning.

In archaeology most data is generated through field excavations, whereas data re-use is less common and this is borne out by the three researchers interviewed, where only one interviewee re-uses data. The other researchers tend to use the grey literature that are produced from excavations if these sites are similar to their proposed excavation. "*Only occasionally would one want to re-use data, e. g. the incorporation of house or farmstead plans from different sites and comparison between the different plans. Some required analogue drawings that required digitization, but others were already in digital form. One could combine them together at the same scale, make comparison between house plans.*"

However, it is likely, as the quantity of digital archaeology data grows, that researchers will see data sets from excavations as a source of new research in its own right. "*More people will search for existing data as a resource to answer new research questions.*" In the Netherlands, where it is required that all excavation data be deposited in one archive, "*finding data is not too much of an issue as there is only one EDNA, which makes things easier*".

A common issue for archaeology, but also for other humanities and social sciences, is the ability to search for data across current political boundaries (also identified by the political science researcher). Clearly in ancient history current political boundaries did not exist. "*If there were German and Belgian counterparts to EDNA*[7] *and it was possible to search (multilingual) across these archives it would be useful, for example like the ARENA II project.*" Also: "*There is a lot of interest in being able to cross-search countries, for example amber, which came from Denmark in the Bronze Age, you may want to investigate the distribution of amber artefacts available now and how far they have travelled from the source. [. . .] The Bronze Age just kept on going and burial mounds in Germany are the same as in Belgium, France and the Netherlands.*"

For those archaeologists interviewed, all the data required are collected before metadata enhancements are added. For one researcher, any additional metadata or annotations were for personal research use only as part of a new dataset and so did not see it as an enhancement of the existing data. It was felt that most researchers do not see enhancements as part of an ever-increasing corpus from which everyone will be able to draw benefit. "*What*

[7] EDNA is the e-depot Dutch Archaeology. For more information on this service, see 6.2 e-Depot Dutch Archaeology (EDNA).

do I need, how do I get it, what do I need to do, where can I do research, and at the end I will have a publication and a dataset. The publication is based upon work that is acknowledged, but I might have completely reshuffled everything."

In archaeology in the Netherlands, during an excavation project, everything is stored locally, as each member of the team is working on their own dataset. An excavation project is highly dynamic, where new data is created everyday from up to ten specialists. There would normally be a shared network disk for the gathering of the data and post processing after the excavation, which will be backed up daily. Data is very valuable to archaeologists because an excavation cannot be repeated, and therefore it is important to ensure that data is not lost. There will be a back-up in the field, using external hard disk drives, as well as at the university. Once an excavation project is completed, and the database does not have further edits, the data is, obligatorily, archived in EDNA.[8]

The archaeologist that is reusing existing unpublished data combines his gathered information into a database on the 12 sites. The aim is not to create a detailed study of each of the 12 sites but rather a comparative overview. However, the research will result in new data as several sources of data are combined, metadata is added and new maps, for instance of pottery distribution, are created by integrating existing GIS data. Currently the digital data on the sites is stored at DANS under restricted access and project-specific data is stored on a local computer. This will eventually be deposited at DANS when the project finishes in 2012.

Desk-based archaeological research is now possible using archives such as EDNA as the publications (excavation reports) and data are accessible. Time is saved because there is no longer the need to travel to libraries that hold the analogue publications. "*The effect of digital resources within [archaeological] research will only grow during the next few years.*"

Even the early-career archaeologist prefers to publish in highly rated, peer-reviewed journals before putting his work on a website. The choice of "*a valuable scientific journal*" is more important to him than the aspect of Open Access, even when he would receive money to pay for publication in an Open Access journal.

However, another researcher believed that the "*reputation of the journal is not as important in my area compared to other research areas, and citation index is not as important for me to get further funding*". For this researcher, it is more important to get published than getting papers into the right jour-

[8] EDNA (e-Depot Nederlandse Archeologie) is hosted at DANS using the EASY archive. Some university archaeological departments (e. g. University of Groningen) have their own data archive as well.

nals. Other researchers used conference proceedings and thus were less concerned about journals reputation or they published in "not strictly scientific" journals due to time constraints.

Interestingly, not every researcher interviewed has an online list of publications on a personal homepage or institutional website, nor is particularly worried about having an online presence. However, for some scholars, it does appear to be an essential part of their standing in their community. "*One of my experiences is that I gain a lot of visibility by having things put online. [I] always feed the institutional websites with PDFs, [however] institutional websites are a problem because they are not stable, and repositories are a better bet.*"

Two researchers identified the lack of an institutional literature/pre-print repository within DANS. The current ingest procedure into the repository is directed towards data and many of the metadata questions do not apply to a single paper e. g. the [dataset related] question "how many files does it have?" It was felt that a publication repository would give the researchers more visibility and that "*it would be harvested by other repositories that only harvest metadata*" It was suggested that this repository should be for prepress and not for publications, and should contain, for example, longer discussion papers rather than shortened papers for journal-based publication. "*If you wanted to be more descriptive than is possible in an article for a journal with all the details put somewhere and then extract a certain aspect for the journal paper. Ideally it should be kind of peer reviewed. It doesn't always have to be an international consortium, but rather internal or some sort of editorial control to ensure that quality is maintained. A level of quality that is comfortable for all the group of people who are working with it.*"

Both researchers thought that only using individuals' websites was problematic as "*A personal website doesn't have the persistence of an archive as people move their websites around.*" One researcher felt that: "*The fact you have a third party willing to publish your work properly has a psychological effect. The tangible object of a book is still nice. When it arrives boxed, it is too late to change it, and having something in your hand as a product of your effort is nice.*" Furthermore the researcher commented that the linear process, with iterations of improvement, an editorial process and a formal deadline improved the quality of the final work. "*Having a book, I can show it to my mother, rather than saying I've just had a paper published in an electronic journal.*"

Archaeological publications do not follow the normal pattern of peer-reviewed papers in journals or chapters in books, but results from excavations are typically published as reports as a result of the size of the publication. Publication is normally through institutional series. Journal articles are more

likely to be in the form of a generalized article or something from an important excavation or team.

The excavation reports are usually under institutional copyright, but are made available digitally through EDNA and in local university repositories. This product of the institute is not peer-reviewed but there is an editorial team that will maintain quality. There are a limited number of national and international journals in this field, but there are conference proceedings, which are peer reviewed.

Mainstream archaeological publications are by "internal" official reports. Dutch state services have a number of series that are published and there are a high number of edited volumes. Paper versions are still the main medium for publishing your information from excavations and other archaeological studies. These are usually available in PDF, but the paper (book) is still the main version. "*I am convinced that the more openly you make your publication and data available, the more your research will be cited and re-used. If it were an Open Access PDF document then you would download it immediately and read the chapter you are interested in and will cite it. So I try to be as open with my data and publications as possible.*" Also: "*Standard [archaeology text] books would be a very valuable addition to EDNA*" available as Open Access, was a comment from one researcher who also taught on a masters course.[9] However, another archaeology researcher commented, "*customers want to have analogue printed reports*" (the PDF of the publication is sent with it as well). For example, when an inventory fieldwork is being conducted to explore if there is archaeology in the ground, the publication of it is sent to the "builder", and these publications are freely open.

As identified earlier, there are other forms of publication common in archaeology apart from journal and conference papers, such as local leaflets for the public, local history websites and popular books about local history which are on sale. One researcher commented, "*There are lots of limitations to printed journals and books, particularly when it comes to illustrations and interactive [multi] media. For me the printed journal is a little bit out-dated. I also like hyperlinking to other resources on the Web.*" This researcher self-publishes as "rich internet publications" and considers it as a form of scientific publishing, under a Creative Commons share-alike licence, but does publish in Open Access journals as well where there is a policy such that "*you are free to use [your article] for your own academic purposes*" Also: "*What everyone should have is the right to use their own work. It is stupid to sign away your copyright and then have to ask for permission to use work you have created. [It is] completely crazy how the publishing industry is going, in that you work*

[9] One of the top downloads from EASY is the archaeology book "De steentijd van Nederland".

for free for them then you have to pay if you want to use it (your work). [I] will look for peer-reviewed Open Access journals, but I don't think so much about the (Thompson) ranking of the journals. Colleagues have strong beliefs in other directions, but peer reviewing is very useful, if done in an honest way then it can be very useful."

An archaeologist commented: "*Persistent identifiers (PID) are very important as that will give you the opportunity to cite the data source in a normal way... Data citation [is the] same as literature: author, title, date and PID, plus which file, and this can be verified.*" Gradually, in archaeology people are becoming aware that there is the possibility, a need and a way of citing published data. This is becoming part of the normal workflow of the archaeological community, but mainly since EDNA has been in place. "*Citation of data is growing, but it will take another 10 years before it will become common.*" This awareness is growing due to champions who are promoting Open Access, who believe that it is a good thing that their data is being re-used. Crucial for the re-use of data is the usage of PIDs. DANS is currently evolving the PID system used in EDNA (and EASY) so that a single file can be referenced rather than a collection. There is a move to increasing the granularity of what can be cited.

In archaeology in the Netherlands, the storage of the publication, and possible requirements documentation, with the data (even though the publication is stored digitally elsewhere) means that the publication acts as documentation that aids the understanding of the dataset

For archaeologists in the Netherlands, access to data, grey literature and internal reports are not restricted and this is born out by the comments of the interviewees, who have not found access to data required being restricted.

For the commercial and municipality archaeologists, the list of priorities of excavation and report means that generating data is usually last and largely unfunded. The publications are done but to deposit the larger sets takes time, staff and organization and is currently not part of the workflow.

In the commercial area of archaeology, there are a number of situations in which one does not want to make the information available. There are reasons why embargoes are implemented on archaeological data: for instance, if an inventory of an area is made public, then it could be inferred that the area may be built on or used for other purposes so land/property speculators might use that data, or if there are protests against the location of a road, then it maybe possible for action groups to misuse the archaeological information/data to stall the process. There is an obligation to make everything available because it is cultural heritage; therefore a short-term embargo is used. The Dutch state service has 2-, 4- and 6-year embargoes.

Treasure hunting may also be an issue however; this is not really a problem, as the publications are not released until 2 years (normally) after the excavation has been completed. Treasure hunting is often used as an argument, but it is not a good argument to keep archaeological data private, because if a site is important then it will be classified as a monument and therefore gain legal protection. Before every excavation the archaeologists must inform the state service, who will check that the location is not a listed monument, and provide a report after the excavation has finished with a structured summary. With larger projects, there is a programme (methodologies) requirements documentation that describe how the excavation will be undertaken.

One archaeologist deposited his doctoral data in EASY at the e-depot (EDNA), at first with an embargo to restrict access and then later making it available openly. The reason for doing this was so that the data would be assigned a PID, which could then be cited in the PhD. Only after the PhD was defended the data was made public. Although there is not strictly a legal or privacy requirement for an embargo, it is an example where embargos on data are useful. Shortly after changing it to Open Access it appeared in the top ten downloads in EASY.

Archaeology produces a lot of data but communication of results occurs at the report publication level. Within archaeology, it is very difficult to draw definitive conclusions about data. "*Sometimes it is more interpretation of the type of artefact, the date or the cultural relevance/significance: it is all knowledge-based inference and interpretation. If you gave the same artefact to different archaeologists they may come up with different interpretations of what it is, or what it is used for. Even measuring the length of an artefact is open to interpretation as most are broken. [. . .] Data is open to interpretation, and conclusions are soft, so this makes people cautious of reusing data.*"

Another comment was: "*It is not a formal part of research to share the data as part of an university department, but it is for commercial archaeologists in the Netherlands.*" Normally there is no checking whether data is deposited. "*There was the idea that they [funding organizations] would not pay the last 10% of grant if you did not deposit your data and DANS signed off that you have done so. However, I don't know if they have followed through on the threat.*"

Archaeology in the Netherlands is a small community of researchers, so it is normal to share data on an informal basis between researchers upon request. "*If someone wants to extend your work they may contact you, for example, for your GIS file, and perhaps ask to re-use it.*" One archaeologist is willing to grant individual requests for access to data when it is being studied and generated, but will not yet put the data on open. They are always welcome to ask for the data before publication.

Main issues in the field Archaeology

- There is a need for retro-digitization: digitization of data gathered before the start of the digital era.
- Excavation results are normally speaking published in reports. Excavation results should be distinguished from research data in archaeology. These are selected and collected from the excavation results and subsequently analysed by university researchers as part of their normal academic research. These research results will be published in peer-reviewed academic journals.
- The interest for, and possibilities of, reusing archaeological data is clearly growing. There is, however, still a way to go regarding data sharing and standardization of metadata. National repositories like EDNA can help here.
- Sharing of data may be made easier by the implementation of standard data formats and by clearly defined embargo regulations.

3.2 Political science

One researcher from a political science and computer science background now works with political data from many sources. He enriches these data with all kind of connections between the different data objects. In this way he uses the Dutch parliamentary proceedings as a secondary data source. This procedure leads him to comment upon the quality issues of open and publicly available data. In this case, it could be safely assumed that the data source was "authoritative"; however, when errors were discovered, the researcher was initially "blamed".

This researcher works with a broad collection of data of many formats, including controlled vocabularies (in all 24 EU languages); however, the main data source is the proceedings from the Dutch parliament (and others). Other sources include: political blogs, RSS feeds, Twitter streams and newspaper articles with user-generated comments. The very large Dutch proceedings are all scanned and are complete from year 1814 to current. All the data used exist as digital sources, but some scanning is undertaken. These sources are often structured text, but the structure is not explicit, so there is considerable effort in making the implicit structure explicit and machine-readable

When it comes to political data, the most effective way of managing and enriching the large quantity of data that is used is to download and store all data sources locally to the research team. This is because the most costly process is transformation of the data. "*The raw data is stored on disk, and backed up to a dual-redundant RAID. The transformed data is stored on disk in an eXist XML database which is also backed up. The database has 20 GB of text, which is a significant amount of data.*"

The data coming from the Dutch parliamentary proceedings and other sources have been greatly enhanced. The text from the proceedings is enriched by cross referencing, based on named entities, for example to the names of speakers (their bio page), political parties, dates and controlled vocabularies of political issues (which is the hardest). Hyperlinks to laws, which are identified in the speech (with a specific number that points to the legislation), are also included. Votes are extracted from the data. New queries can be asked, for example "everything said by a person" which is a completely new view compared to the documentary of day by day, and analyse the language used by a speaker over time.

The data enrichment is automated using machine-learning algorithms; however, the rules are hand coded. This is because the structure is quite consistent throughout the corpus. It is also consistent between countries as six countries use the same the proceedings methodology, so it is possible to apply the same automated enrichment. "*Then it is possible, with machine translation, to work with a large database across political boundaries.*" The addition of Dublin Core, TEI (for text mark-up), and ISO country and language code metadata is also automated. "*Persistent identifiers are added to every paragraph to make very specific linking possible.*" This means millions of identifiers are generated for a data collection, which is unique amongst the researchers interviewed.

This research is the only interviewee who published data directly after enrichment and before articles are published. This research team, which is enriching the parliamentary proceedings, use the community of users to check the data, "*which is the main advantage of openness*". They have discovered that it is much better to be completely open and honest, that there are a lot of mistakes and be willing to hear from users and correct the errors. "*If it is open immediately then quality becomes important to you because people will be checking it, using it and building on it. Openness will just improve quality [of data] through these simple social mechanisms. Open data can be a really big thing.*"

The researcher did identify a number of disadvantages to this approach. For example, if someone takes the data, builds upon it, and researchers using that dataset finds that there is a gap or it is not as reliable as they think, then the use of the data is already quite far away from the origin so it is difficult to identify the cause of the problem. Enriching raw data created by others makes this team brokers between the original data producers and the users. Regarding the original sources of data used by him: "*you would think [these] are more authoritative than we are, but actually they are not. This is dangerous and could kill your reputation*" as the researcher will be blamed for the errors in the enriched material which come from the original source

material. This method of data publishing is not typical for the field, and the researcher believed it to be "state of the art", since "*most others in field do not publish data, or use it is a bench mark. The data itself is not valued*". Also: "*Everything we do is completely public so it is quite dangerous. Once it is in the eXist database it is public. There is a URL to every item in the database and you can query it. This is highly public and therefore this makes it extremely easy to share. It is important to be fast [with corrections]. Only when people use it you find mistakes.*"

Main issues in the field Political science

- Datasets in many formats and are used with controlled vocabulary.
- In this particular area new sources like blogs, RSS-feeds and Twitter are becoming important, but some of these are also extremely difficult to get hold on for researchers.
- Enrichment of existing data (texts) is an important activity in this discipline.
- Openness of data leads to enrichment by the research community.

3.3 History

3.3.1 Oral history

Quantitative oral history recordings are the sources material for one specific scholar. An important aspect of these recordings is that the data collected from individuals may be sensitive.

The researcher gathering oral history data does so "*with the aim of making this information accessible to others*", and this is "*implicit in the discipline*". The data used can be very sensitive, however, the data is often being "*handed down on the basis of trust*". The researcher considers that for each form of qualitative data it is necessary to have a specific protocol, for the people that are questioned in an interview and for the people that want to use that information, and this cannot be generic, "*because then it would be useless*". In the USA there are strict review boards to assess the research protocols used. The interviewee thought that this might be useful in the Netherlands, but argued that such review boards might make research very inefficient because of the difficulty of getting permission to use the data. The difficulty with the oral history data is the selection being made. Statistically the selection is often not representative of the situation or the whole group of interviews: "*it is the group of people that want to be heard that become part of the selection and not the ones that do not want to be heard*" This is important to realize when using this kind of data for research. It was suggested by the researcher that there should be a pilot project to test the sharing of qualitative data

in Europe, but suspects there will be problems and solutions will be found. Eventually the sharing will work but the researcher stressed the importance of respecting peoples privacy and the contract someone signs need to be very clear on this aspect.

For the oral history interviews, the researcher established an embargo commission to restrict access to the recordings and data in sensitive cases. Four people with very different backgrounds who will provide each a different insight on the embargo cases form this commission. People who are being interviewed are mostly not asked to give permission for the use of everything they say by researchers, and this situation should be taken more serious. The commission should prevent sensitive information becomes public when the person in question does not want this to happen.

There is also the issue of how long information should be kept private. Sometimes it is desirable to release the embargo before the set date, therefore it is important to put something to cover that in the contract. For example when the person in question brings his information to a public source, the original holder of that information is then permitted to release the embargo and use the information publicly as well.

For researchers in this field, a major concern is that the research data created as part of their work is not re-used for commercial purposes, particularly when this is interview data gathered.

3.3.2 Cultural, social and economic history

A scholar in the field of historical demography, working with the HSN (historische steekproef Nederland/historical sample of the Netherlands[10]) noted that "*most historians are not able to work with digitized HSN material*", so his assistance is required in preparing the data. In such cases the scholar will be a coauthor of the publication. This researcher also commented that he had "*easy access to everything I need*" with respect to data and literature.

A researcher investigating intergeneration mobility (in the field of professions in education,) also creates dataset based on HSN. However, this researcher commented that "*HSN is not user-friendly and it is difficult to create subsets*". The subsets will not be published or stored, as "*there is no facility for this. HSN should create space for storage of subsets and results of research*".

This researcher re-uses sociological surveys (statistical data) archived on EASY as a historical source and mainly uses digital sources of data. However, there are obstructions to using these sources such as the need for registration

[10] http://www.iisg.nl/hsn/abouthsn/index.html.

and specialist skills to handle the data. Furthermore the metadata may not be of a high-enough quality to make full use of the data source.

Other researchers tend to store digital data locally mainly because it is work in progress or because the researcher is working on their own, but other comments include: "*I lost data while working in a team due to overwriting of files stored in cloud*", "*I have relatively small amounts of data*", "*part of the material is digital, and a part is still analogue*". Back-up on institutional infrastructure seems to be the normal way of ensuring that the data is safe, but some additionally store data at home on a private PC as well.

Working with the HSN (historische steekproef Nederland/historical sample of the Netherlands), the users are obliged to add the release of the project-specific subset of HSN with the publication in order to make it possible to re-use the data. This is a form of data citation, but again the researcher commented that the rules for referencing are not formalized. Another researcher believes that "*data creation should be given credit*" and applies the citation rules set down by the data producer; however, even in his own project this facility has not been implemented.

For one researcher, the right of display of copyright material on academic research websites is problematic. This is because cultural heritage institutions are "*exploiting their rights by asking large sums of money for a single image. Mostly they are generous to academia but it should be free, as we have paid for the paintings and objects in the museums and it should be free on a website for everybody to re-use, but obviously not for commercial gain*". It can take considerable time and money to ask permission and obtain the right to re-use each item, especially in a visually rich digital publication.

There is no obligation for researchers to publish data, and at DANS this is currently free for self-depositors. In the same way as paying for publishing an article in a journal, it is likely that depositors will have to pay for data archiving. This researcher considers this is better than letting the user pay because that is against the idea of Open Access; however, this paid deposit is something that is a point of discussion for the future. "*There is an opportunity for humanities to take the forefront, because we do not have many commercial barriers, and researchers in humanities want a big public. In the humanities a considerable result can be achieved with relatively little energy. Next to intellectual talents, you need extra resources, an infrastructure to store and re-use data and research outcomes.*"

Main issues in the field History

- Openness of the oral history data often conflicts with its confidential character.

- Skills to handle complex datasets are sometimes lacking within the discipline.
- Low quality of the metadata hinders optimal re-use of the data.
- Often cultural heritage institutions ask fees for accessing their data, in combination with copyright barriers, so that researchers have to negotiate with them to (re-)use the data.

3.4 Law

For law research, there are two main sources that can be considered as data. These data, more commonly referred to as resources, have not been created *within* the law research environment, but in the administrative world, including the administration of justice.

The first type of source is the jurisdiction consisting of all the judgments issued by the law courts and other bodies administering justice. In the Netherlands, only a limited number of the latter category is publicly available on the internet, mostly only those cases that are important from a jurisdictional point or having large public attention. In the law world there is an ongoing debate on the desirability of making far more (or in fact systematically all) judgments available online. There is, however, strong resistance against this, mainly based on privacy and financial considerations.

Theoretically a solution to this problem could be to deposit (all) these judgments in a research repository to which only law researchers (i.e. not the general public) have access. The Personal Data Protection Act of The Netherlands (Wet Bescherming Persoonsgegevens, WBP) enables this possibility. Access to data files containing personal data is allowed for "academic, statistical or historical" research. This principle is based on the European Directive on personal data protection. There is, however, a pessimistic view on the materialization of this in the near future. Too often, privacy reasons are used as a way of avoiding publication, but there are certainly ways to get around the issue of personal data protection. There is already software solving this problem available. From the political world, attention is increasing towards the quality of administering justice. This could be seen as a hopeful development, as it could lead to more attention to the problems mentioned. Other sectors of the society can be considered as stakeholders here, such as journalists, but also lawyers. Besides from the privacy issue, there is a financial hurdle: who is going to pay for this? Anyway there is a clear lack of willingness in the juridical world itself to do this and to give these "data" away to another (research) organization.

There are the laws and official regulations at all kinds of level themselves. These are available, nowadays mostly in digital form on the internet, albeit not in the most ideal format. Regarding the publication of laws, another

important source for law research, the situation is considerably better, but far from perfect, as it can be very difficult to reconstruct the development of certain laws or regulations over time.

According to some, mostly younger, researchers, research in the law faculty is "*old-fashioned*" legal, mostly textual, work on jurisdiction and law making. Law researchers on the whole are not very much interested in more quantitative analyses, in the direction of social science. This could explain why there are not so many resources available for that kind of research. These younger researchers very much would like to have available a substantially larger corpus of law resources, in particular court decisions on all levels (district, appeal, etc.) than is presented today.

As for academic publications on law research, most journals do not have Open Access policies at all. There are, however, some hopeful developments here as well. There is for example the *Leiden Law Review* of Leiden University. In this repository there are often Open Access articles. Most articles are published in Dutch, unfortunately.

Another field, strongly related to law, is that of criminology. In the way research is carried out here, there is a strong difference with that in law properly, as described above. Criminology is a form of social science and research methods are those as used in the social sciences. There are research projects with strong qualitative elements, but the whole research process still can be considered as social science. Access to data in this field is also mostly very restricted. This is again mainly because of privacy reasons. Most of the research data are of a highly sensitive nature, including personal data on youth offenders or criminals generally. The Ministry of Justice only allows access to the data by researchers under very strict conditions of use.

Contrary to Open Access to law data, both for law and criminological researchers, publications are not in particular problematic regarding Open Access, as these would normally only contain anonymized data.

Main issues in the field Law

- For research reasons, all judgments should be available online, at least for academic researchers.
- Most law research journals do not have Open Access policies yet.

3.5 Economics, social science

Looking from the perspective in what way Open Access might be beneficial to research infrastructures in these fields it should be observed that, as in law, economic researchers are using, for a large part, data which has not been created *within* their research environment. Financial research uses data from

commercial firms as well. That means that these data are not always easily available for researchers and, if available, can have (sometimes very strong) restrictions on dissemination after processing. This could, to a lesser degree, also hold true for publications based on these data.

In economics research, data created in the administrative world are far more important than survey data, which are often created by researchers. One institute with a central position here is the CBS – Statistics Netherlands. They have a large quantity of data that could be linked with each other as well as other data easily, technically speaking. In practice, however, things are not that easy. In the world of academic research the CBS is seen as a difficult organization from which to obtain data, in particular microdata to be linked to other microdata. The CBS has a strong policy of protecting personal data. In particular data on the financial world are difficult to obtain. The willingness of the central banks of Europe to make these available varies from country to country. Even more difficult are data from private banks as well as commercial companies generally, and in particular multinationals.

It seems that most researchers in economics are very much in favour of a very limited period of embargo. One researcher sees as a maximum, a period of 10 months, which is considered as short, especially by PhD students. According to this researcher, on average, a period of 2–4 months is the needed for translating all the data labels and relevant information into English before the data are published. In his field (economics, household surveys) a longer period is unacceptable; information would really be outdated. One thing of importance is international usefulness of the data: that is why all documentation is obligatory in English. He would like to have as much Open Access to the data as possible, only restrained by privacy laws or contracts with commercial parties. However, users of the data should always be traceable. Registration is absolutely necessary.

Also for publications, he is very much in favour of Open Access for all publicly funded research. According to him, there is not yet enough awareness amongst researchers for this, when they are publishing. They are giving their copyright away too easily.

Interestingly enough however, there are conflicting views on this in the social sciences. Another researcher, a professor in the social sciences was rather hostile towards the idea of Open Access and she wonders where this comes from. Her main reservation is that the data collected and processed by a team of dedicated researchers over long periods cannot easily be understood or appreciated by other researchers because of lack of methodology and consequently potential wrong use of data. Furthermore coauthorship of the original investigator should always be appreciated. According to her, research teams which do not provide Open Access to data are publishing far more than those

who give full access by way of the internet (for example, in the Netherlands the large-scale projects TRAILS versus NKPS). In her major research project, a large data corpus exists which is extended over a number of years. All these additions continue to enrich the dataset. Other researchers are invited to use this material but under strict conditions: they have to consent in members of the research team being coauthors of the publications.

Another point made by her is that, if universities, funders (like NWO in the Netherlands) and others really should enforce Open Access to data on researchers, social sciences undoubtedly would reduce its lead over the medical sciences regarding the "easy" availability of research data. In the medical field data acquisition, in particular from patients, is already a much more bureaucratic process.

Regarding publications she admits that "publication inflation" has been going on in academia for years, being very focused on articles. She would very much prefer a system in which there is more attention for quality than quantity. However, she does not see developments in this direction. In social sciences, hardly any high-quality Open Access journals exist yet.

Main issues in the field Economics, social science

- Financial research data are not always easily available.
- Re-use of financial research data is often restricted.
- Economics resembles fields like biology and chemistry in that only no embargo or a short embargo period is acceptable for the progress of research.
- The number of high-quality Open Access journals is (still) low.
- In social sciences, "Open Access" is still valued very differently.

3.6 Linguistics

Literature is a data source for some humanities research and as identified by one researcher it is sometimes a barrier to conducting research if there is no access to an academic library that pays for journal subscriptions. This researcher would pragmatically tend to use literature that is Open Access "*because it is easier to get hold of*".

Only one researcher was interviewed who came partly from the linguistics world and also partly now works for the linguistics digital infrastructure CLARIN. In linguistics copyright is a very large problem, in particular for using large text corpora and text mining. The large and growing text databases of newspapers for example, is of great interest for scholars of contemporary language use, are not available for research, except in a very limited way. It is his experience, however, that it is possible to circumvent the copyright

restrictions of commercial publishers, as long as there are no commercial interests involved. Good communication with commercial partners is essential for this to succeed.

This researcher confirms that he is very much in favour of Open Access, fully in line with CLARIN policies. The only two unavoidable barriers are privacy protection and copyrights. The copyright barrier, in particular, could partly be surmounted with extra effort and ingenuity by the researcher. Privacy protection could also be lifted for strictly academic research. Under certain conditions researchers are allowed to access confidential personal data, as long as they only publish about their findings in *anonymized ways*. Concerning embargos, he would advocate temporary restrictions only on the ground of protecting a PhD student and then until the actual graduation moment.

Main issues in the field Linguistics

- Research is built on existing text corpora.
- Copyright hinders re-use (for instance via text mining) of the resources. There are ways to overcome this problem.

3.7 E-Science

Many of the researchers interviewed are difficult to categorize, as they often conduct research that could be said to be at the interface between disciplines. This can be particularly seen with one interviewee, a senior researcher who has a background in physics, mathematical modelling, philosophy and history of science, and who is now conducting research in modelling the economics of innovation (social sciences).

This researcher uses large bibliographic databases to investigate the growth of a research field and innovation by extracting the network information of collaboration and citations. However, the researcher identified the difficulty with access to "*cleaned data, which is essential*", but these sources are private. Other large data sources, such as the Web of Science database (a favourite) has a very specific view, and in this instance is only a selection and is biased towards English language and specifically American journals. These difficulties in finding suitable openly available sources lead the research team to use a Wikipedia "dump" from 2008. Although, with Wikipedia, they have "*as much reliability with Wikipedia as you have with Wikipedia*" They observed a "*lot of self-correction and self-cleaning going on*"; however, this does not guarantee 100% correctness. "*You do a selection of data sources that are selections in themselves. [...] The bias doesn't come from the data sources available, but I think the bias probably comes from the research attitude and the epistemic knowledge inside of the field.*"

For this researcher whose team is using the Wikipedia dump, a selection of the raw data was made and statistical methods along with clustering algorithms were tested. The volume of data challenged the algorithms, with some 2.8 TB of raw data to process the BigGrid grid infrastructure was utilized to both store the data and run parallelized versions of the analysis algorithms. The data was parsed and extracted, looking at changes over time. Rapid changes that occur in a short period of time which represent vandalism, or "noise", were discarded. The decision of how and when the data was agglomerated to feed into the statistical analysis tools. The algorithms were written in Java, and for clustering existing algorithm were adapted. Analysis of the results files was also run in parallel on the grid. "*As a humanities group they [BigGrid] took us on board as a pet project. We were one of their successful Humanities projects.*" Data generated was also stored on the grid. When the formal collaboration with BigGrid finished the secondary data was downloaded to a laptop hard drive, and a backup was made to an external hard disk drive. The Wikipedia dump was also stored on a desktop computer as a 100 GB zipped file. "*Currently the data is not accessible, but planning to do so.*" The team of the researcher did not make qualitative annotations, "*but more quantitative notes, done very primitively using excel sheets, or looked it up in the raw data and just counted the occurrences e. g. we compared number of times a term came up Wikipedia and UDC library data.*"

The range of publications and forms of publication is again broad in the humanities and social sciences. Clearly, peer reviewing is crucial, and the reputation of the journal is the first priority before accessibility, but some researchers do wish to reach the widest possible audience including other forms of media. "*Definitely peer reviewed but making the pre-press prints available as soon as possible as well.*" Also: "*Would I choose a journal according [to] its public availability? No. I would choose a journal according to its standing in the community rather than gain visibility.*"

The time take to get a paper published in a journal is an issue for some researchers and this does influence the choice of journal. "*We have submitted a paper to a journal more than a year ago, but they are not so fast at putting together the special issue even though they are only short papers.*" And: "*I'm most interested in short turnaround on publishing and don't like to wait for half a year or more for publishing. [In my field of study] things are out of date in 2 months time. I prefer to publish [literature] online as Open Access/open source, e. g. in the online journal of document information, [where] authors keep copyright.*" Also: "*Submitted a journal article some time ago, and recently, before publication we have found that some hyperlinks didn't work.*" In the first case the researcher was concerned whether publishing a pre-press version locally or depositing it with a subject-specific pre-press repository

would risk the inclusion in the journal. The researcher also mentioned that the ability to deposit an article in a pre-press repository "*depends upon who you are talking to in the [publisher's] administration and how you bring it [to them]*".

Main issues in the field E-Science

- Open available resources are sometimes difficult to find.
- Sharing information is in certain situation more important than publishing articles in high-quality journals.
- Time between submitting a paper to a publisher and the actual publishing is too long.

3.8 Important general issues

Apart from the discipline-related issues, there are issues relevant to the whole field of humanities and social sciences.

- Time between sending in a manuscript and the publishing of the final article takes too long.
- Researchers should be given credits for creating datasets.
- There is a strong need for searching datasets across political boundaries.
- Persistent identifiers given to both traditional publications and datasets are important for the retrievability and citing of resources.

4 Current status of Open Access

4.1 Data

For the Netherlands, EASY is the main access point to datasets in the humanities and social sciences. DANS gives access to the description of datasets, but the depositors determine the access rights to the deposited datasets. There is no totally free Open Access within EASY, but access free after registration. This registration is used to identify meant to retrieve who the users of the datasets and generate usage statistics.

In EASY there are 19,900 descriptions of datasets (August 2011), and 8583 of these datasets are Open Access. From the other datasets, 1743 have some form of restricted access, 9135 are related to the restricted-access "Archaeology" group and 439 have another form of restriction. Major concern of DANS is that all datasets are described in EASY. The policy remains, however, "open if possible, protected if necessary".

Summarizing the findings from the case studies, it could be said that generally speaking researchers seem positive to Open Access to data. Some reservations, however, could be observed, as discussed below.

4.1.1 Personal data

Researchers are clearly extremely cautious in handling records containing personal data in whatever form. This is both for ethical and legal reasons. The European laws are very strict, in some research fields still stricter than in others. In particular, medical data can only handled with extreme care. In other research fields there is a grey legal area. The privacy laws are not always perfectly clear: what is still allowed to do and what is not allowed anymore? Specific examples mentioned are oral history interviews on sensitive topics like war memories and deaf-mute children (not being patients) filmed for research purposes. Making these kinds of personal data available on the internet, even restricted for research reasons, creates all kinds of challenges of safety and security. There is a common understanding amongst researchers that these data have to be handled, disseminated and archived with great care. Handling personal data is, however, unavoidable in research and should certainly not be made impossible or deterred by stricter regulation. Clearer regulations on some points should certainly help.

4.1.2 Copyright issues

This barrier to a free and open use of material that in some disciplines (linguistics, social science, economics) is indispensable for research is often mentioned. There are clearly huge hurdles here. It could, however, also be concluded that there seem to be flexible and creative ways to go around this problem. A more generic legal solution would always be preferable, like a general exemption for research purposes. This exemption exists more or less for privacy data. A European directive would help here.

4.1.3 The reluctance of scholars to share data

This is an attitude that could be found in particular amongst social scientists, in particular in social psychology it seems. The main reason here is protecting research data on which research teams worked for years in which a lot of money is invested. The great fear is that other researchers will gain an advantage, in particular too soon, and steal the show. Another, maybe even stronger fear is that other, non-competent researchers or amateurs will misuse the data and willingly or unwillingly misinterpret them.

Regarding the last point the only concession that would be acceptable from the perspective of Open Access to data is temporary protection in the form of a short embargo. This is acceptable for most researchers, especially when it concerns PhD research. Opinions differ on the duration: 1 or 2 years are mostly mentioned. In humanities (history, archaeology), researchers do favour on the whole longer embargo periods than in the social sciences and in particular economics. Especially in the social sciences, data may become outdated within a relatively short period. In this field, a long embargo period may be a serious hinder for research progress.

For the rest it seems a question of mentality, maybe of generations of researchers. There are indications that younger researchers are more inclined toward Open Access generally. On the other hand, as some interviewees hinted at, younger researchers could be more protective as well. They still have to make a scientific career with their data. It should be mentioned that funding organizations are becoming stricter on this point. The largest research funder in the Netherlands, NWO, has recently adopted a stricter policy on Open Access to data, which will make it impossible for researchers to keep their data locked away from anybody else for years.

Implications for OpenAIRE

- Especially in disciplines like health sciences, social sciences and history researchers are dealing with personal data. There is a need for a European regulation that will allow the handling of these data for research purposes.
- Apart of the settlements for Open Access to publications and created datasets (datasets that are produced during the research), a copyright regulation for existing resources like text corpora or financial data that would allow free access for academic research would benefit the progress of research in disciplines like linguistics and economics.
- Advocacy in favour of Open Access to data is needed to overcome the reluctance of scholars to share their data. The readiness to share date could be augmented by allowing an embargo period of 2 years as a maximum, possibly dependent upon discipline.

4.2 Publications

As already explained, the differences between the disciplines within the social sciences and the humanities are broad. Nevertheless, one can conclude that there is a tendency in favour of Open Access, although this impression coming from the interviews will have to be confirmed by additional questionnaires amongst the researchers in this field.

Traditionally, the book plays a more important role – especially in the humanities – than compared to the MST fields. Open Access to books has just begun, as publishers were reluctant to give up their traditional business model. An interesting project is OAPEN[11] in which European publishers cooperate in producing Open Access books in the fields of Humanities and Social Sciences. Meanwhile, the OAPEN site offers access to 832 books (August 2011).

Journals are less important than in the MST field, partly because of the big differences in editorial policy. Peer review is not always possible as in for instance the medical sciences. Nevertheless, the Directory of Open Access Journals (DOAJ)[12] counts the number of journals per (sub-)discipline, as shown in Table E.1. Double counts in DOAJ are possible, but one can see the vast amount of Open Access journals.

The number of available journals in DOAJ is given in combination with the figures from Thomson's Citation indexes.

Table E.1 Number of journals per DOAJ category

DOAJ category	Soc.Sci.Cit. Index	Art and Hum Index	DOAJ
History		273	181
Archaeology		81	32
Religion		128	78
Philosophy		165	163
Linguistics, language and literature		310	428
Social sciences	2475		1445

Other resources for Open Access publications are the repositories Social Science Research Network (SSRN[13]) and Research Papers in Economics, RePEc.[14] These very popular repositories (with a high rank on the Ranking Web of World Repositories[15]) are comparable with PubMedCentral in the biomedical field.

The interviews revealed an important new aspect: the major role of data in the production of publications. This role is that strong, that one could see a shift from traditional publications (articles, reports, books) to enhanced publications (publications that are a combination of the traditional publication and the underlying datasets). This new way of publishing will stimulate the re-use of datasets. Unfortunately, Open Access to these enhanced publi-

[11] http://www.oapen.org.
[12] http://www.doaj.org.
[13] http://ssrn.com.
[14] http://repec.org.
[15] http://repositories.webometrics.info/toprep.asp.

cations is somewhat cumbersome. This is caused by the fact that the Open Access status of the different components of an enhanced publication will vary.

In comparison with other disciplines, Open Access has not reached the same level of importance, but the trend is positive.

An alternative for the traditional publisher is the publishing of research results via institutional repositories. The instability of the URLs is, in this respect, a source of concern. A system assigning PIDs to object could resolve this problem.

Finally, researchers do accept that, in some situations, an embargo period will be necessary.

Implications for OpenAIRE

- Although there is a tendency in favour of Open Access to publications, the fields of humanities and social sciences are too broad to come to definite conclusions. A more detailed questionnaire is needed to validate this preliminary conclusion.
- In the humanities and social sciences, there is a very close relationship between the (traditional) publications and the resources that have been used in writing these publications. Therefore, there is a growing need for so-called "enhanced publications". In these, the relationships between the different resources can be made clear to its readers. This important new development will only be successful if all components of an enhanced publication will have the same level of accessibility (as open as possible).
- Related to both enhanced publications and the re-use of objects deposited in (institutional) repositories is the problem of sustainable access to these objects. Sustainable access could be reached by introducing PIDs to any type of object (text, dataset, image and so on), in combination with a European resolver services that will lead a user from a PID to an actual URL. The next stage would be the connecting of future continental resolvers to realize a world-wide system of resolver nodes. Concurrently, there should be developed a policy of discouragement for using non-persistent URLs to identify objects.

5 Current research infrastructure projects

DANS is involved in a number national and international research infrastructure projects both in preparation and development. These projects may include the archive facility or infrastructure in a particular research area. They may also relate to availability, to the software used or to technical aspects of

the desired data infrastructure. The goals of the projects are the stimulation of collaboration and the sustainable access to data to serve data-intensive research.

5.1 AlfaLab[16]

AlfaLab is a 2-year project (2009–2011) that aims to redress the lack of co-operation and the absence of a technical information infrastructure which forms an obstacle to humanities research in the battle for resources to successfully compete with other sciences. The project initiators of AlfaLab aim to develop a digital infrastructure that may take the form of a digital portal and a virtual laboratory (research area) for alphas, or in short AlfaLab.

AlfaLab is an initiative of the KNAW (Royal Dutch Academy of Sciences), where five scientific institutes which have the objective is to cooperate and promote the use of digital methods within humanities research. AlfaLab disseminates knowledge about digital tools and data. AlfaLab creates and develops digital tools for the humanities community and investigates the use of digital tools in the humanities and how these (virtual) tools support and encourage partnerships.

The goal is to create a digital infrastructure for the humanities, consisting of the following components:

- laboratory: modifying, linking, providing and implementing innovative ICT tools for manipulating and analysing digital resources for the humanities;
- portal: developing a common access to these tools and to the available Dutch data files (a digital portal for the humanities);
- disseminarium: spreading knowledge about new research opportunities by using this infrastructure in larger groups of researchers.

The project focuses on three elements: (i) the Tekstlab (focusing on textual sources); (ii) the Spacelab (focused on geo-data); and (iii) Lifelab (focusing on lifelong population data).

The ultimate goal of AlfaLab is to develop a significant contribution to the humanities infrastructure within the Netherlands. The role of DANS in AlfaLab is to design, develop and implement of the portal environment (the ICT infrastructure where applications and data are being made accessible).

[16] http://www.dans.knaw.nl/en/content/categorieen/projecten/alfalab.

5.2 Connecting ARchaeology and ARchitecture to Europeana (CARARE)[17]

The ambition of CARARE is to ensure that digital content for Europe's unique archaeological monuments, architecturally important buildings, historic town centres and industrial monuments of heritage importance is interoperable with Europeana[18] and accessible alongside contents from national libraries, archives, museums and other content providers. CARARE aims to enable spatial and virtual reality content for heritage places to be brought together in Europeana and new services for users.

CARARE will add value to Europeana and its users by:

- promoting and enabling participation in Europeana by heritage agencies and organizations, archaeological museums and research institutions and specialist digital archives, and raising awareness of Europeana in the domain;
- establishing an aggregation services which contributes on a practical level to enabling interoperability, promoting best practices and standards to heritage organizations, taking account of the particular needs of content for archaeology and architecture. It will bring 2 million items (images, maps, plans, aerial photographs and 3D models) for Europe's unique archaeological monuments, historic buildings and heritage places into Europeana;
- implementing Europeana compatible infrastructures, standards and tools so as to make available millions of digital items for heritage places across Europe, thus contributing to the growth of Europeana;
- acting as a test bed for Europeana's APIs that are intended to make content available for other service providers to use, for example in the areas of tourism, education and humanities research;
- establishing the methodology for 3D and virtual reality content to be made accessible to Europeana's users.

CARARE runs from 1 February 2010 until January 2013 and is funded under the European Commission's ICT Policy Support Programme.

[17] http://www.carare.eu.

[18] The Europeana service offers access to millions of digital items provided by Europe's museums, galleries, archives, libraries and audio-visual organizations. Some of these are world famous; others are as yet hidden treasures. Europeana will deliver public access to over 15 million digital objects by 2011. http://www.europeana.eu.

5.3 Council of European Social Science Data Archives (CESSDA)[19]

The Consortium of European Social Science Data Archives (CESSDA) has been an umbrella organization for social science data archives across Europe since the 1970s. The member organizations have worked together to improve access to data for researchers and students. CESSDA research and development projects and Expert Seminars have enhanced exchange of data and technologies among data organizations.

In 2011, CESSDA is working to become a truly integrated European data infrastructure, with legal personality and full legal capacity, preferably with the legal status of a European Research Infrastructure Consortium. CESSDA is one of the 44 projects listed on the ESFRI Roadmap. The Netherlands is one of the countries taking part in CESSDA-ERIC, with DANS as the Dutch service provider.

CESSDA-ERIC will provide a distributed research infrastructure expressing the principal tasks as follows:

- facilitate access for European social science and humanities researchers to the data resources they require in order to conduct research of the highest quality, irrespective of the location of either researcher or data within the European Research Area (ERA), and beyond;
- coordinate and develop access practices, agreements, licensing models and any other legal and organizational measures that enable and extend such access to distributed data resources;
- coordinate and support the installation and maintenance of a technical infrastructure that allows such access to distributed data resources;
- actively contribute to the development, promotion and adoption of best practice for data distribution and data management, thereby enhancing the quality of infrastructural services;
- work continuously for the inclusion of further data sources, from Europe and beyond, into the infrastructure;
- provide training within the CESSDA-ERIC and beyond on best practise in operational processes and data management.

5.4 Common Language Resources and Technology Infrastructure (CLARIN)[20]

The project "Common LAnguage Resources and technology INfrastructure" (CLARIN) provides researchers access to language, text and speech resources

[19] http://www.cessda.org.
[20] http://www.clarin.eu.

and research tools across Europe. The services include archiving and re-use of data, as well as advice on metadata and standards. This way, CLARIN stimulates the exchange of knowledge and data between linguists, historians, speech technologists, communication researchers and many others. An important aim of CLARIN is the availability and usability of resources such as lexicons and text corpora for each language within the European Union.

DANS participates in a number of research projects within this framework. Once the partners in these projects have achieved all goals, they will become CLARIN centres. DANS is one of five Dutch organizations with this ambition.

Other CLARIN projects that DANS participates in have a linguistic nature. They aim at automatically extracting and/or curating linguistic data, as well as developing demonstrators. DANS is involved in the projects because of questions regarding software archiving and persistent identification of small text fragments.

Finally, the Netherlands Organisation for Scientific Research (NWO), CLARIN, and DANS have agreed that for certain applications for NWO grants for humanities research DANS informs the researchers about CLARIN standards that the research should live up to.

CLARIN stipulates that researchers and research groups participating in CLARIN make their resources freely available to other researchers. For this, CLARIN-NL refers to the *NWO Open Access Initiative* for publications but extends it to research data and tools. CLARIN demands from all projects that the deliverables will be accessible for the CLARIN community on a (future) CLARIN central server.

5.5 CLIO-INFRA[21]

CLIO-INFRA is embedded within the European Commission Initiative Digital Research Infrastructure for the Arts and Humanities (DARIAH). Within this NWO-funded project, the goal of the CLIO-INFRA project is: “On a global scale, bringing different and sometimes fragmented data sources together through alliances in an open access model disclosed with the use of a central portal”.

The purpose of CLIO-INFRA is the systematic mapping of the available quantitative information on the development of the world economy in the last 500 years. The goal is to provide a solid basis for the systematic study of the causes of global inequality. By using CLIO-INFRA as a foundation one is able to design and test crucial economic and social development theories. CLIO-INFRA shall interconnect a number of databases (hubs) consisting of

[21] http://www.clio-infra.eu.

data on global social, economic and institutional indicators over the past five centuries, with special attention for the past 200 years.

In CLIO-INFRA datasets will be created or improved, for example on living standards, human capital and cultural and political institutions. Economic and social historians from around the world will work together in thematic collaborations, increasing their knowledge of the relevant indicators of economic performance and its causes to collect and share.

The data sets are accessible via a central portal, which creates opportunities for the visualization of data. The long term goal of the project – as developed by the International Economic History Association – to the academic rules as to provide that international cooperation and to facilitate data exchange.

DANS will seek to establish and setup a long-term storage solution by creating a database archive consisting of aggregated data from Excel sheets.

5.6 Digital Collaboratory for Cultural Dendrochronology (DCCD)[22]

Within the dendrochronology field, the exchange of information about tree-ring analysis and samples is of the utmost importance. Dating of wood can only be made possible by exchanging finds and constructing benchmarking timelines based on pre-dated material.

DANS has built an international repository system for dendrochronological material as part of the Digital Collaboratory for Cultural Dendrochronology project. The repository facilitates dendro researchers with a system to safely store their measurements and object descriptions, and exchange this information with their colleagues.

The repository functions as the central hub of the dendro infrastructure, through which all European dendro materials can be uploaded, shared, searched, examined and downloaded. Its backbone is the TRiDaS data format, which is compatible with most of the laboratory software suites and has been defined especially for data exchange.

5.7 Digital Research Infrastructure for the Arts and Humanities (DARIAH)[23]

The mission of DARIAH is to enhance and support digitally enabled research across the humanities and arts. DARIAH aims to develop and maintain an infrastructure in support of ICT-based research practices. It brings together

[22] http://dendro.dans.knaw.nl/about.
[23] http://www.dariah.eu.

researchers, information managers and information providers and it gives them a technical framework that enables enhanced data-sharing among research communities.

DARIAH is working with communities of practice to:

- explore and apply ICT-based methods and tools to enable new research questions to be asked and old questions to be posed in new ways;
- improve research opportunities and outcomes through linking distributed digital source materials of many kinds;
- exchange knowledge, expertise, methodologies and practices across domains and disciplines.

DARIAH has now entered a transitional phase that will end in 2011. DANS was one of the initiators of DARIAH and coordinated the Preparing DARIAH project. DANS also contributed its expertise to technical and dissemination work packages.

In the current transition and future construction phases, DANS will be jointly leading the content element of DARIAH (along with Centre National de la Recherche Scientifique (CNRS), France) and will be contributing expertise in PIDs and some of its research infrastructure technologies to the e-Infrastructure.

5.8 European Social Survey (ESS)[24]

The European Social Survey (ESS) is an international database that has become a prime instrument for innovative comparative research on social and political attitudes. Up until now ESS has collected four biannual rounds of survey data (2002, 2004, 2006, 2008). For each round, the project interviews approximately 50,000 people in Europe (about 1800 in each country) on a broad scope of their social and political attitudes and social backgrounds in a strictly replicated format, which creates rich opportunities for internationally comparative research with a longitudinal perspective. Currently available download statistics reveal over 17,000 users of ESS data in more than 170 countries: users in the Netherlands rank in the top ten. In the first 4 years of their availability (2003–2007), research based on ESS data has yielded more than 400 scientific publications.[25] The data of ESS are disseminated by the Norwegian data archive NSD.[26] The data are available free of charge and without restrictions, for not-for-profit purposes.

ESS may be depicted as a research infrastructure. The infrastructural nature of ESS lies first of all in the rich and high-quality data resource it con-

[24] See http://www.europeansocialsurvey.org. The information in this paragraph is abstracted from the ESS website.
[25] See http://www.europeansocialsurvey.org. ESS Bibliography.
[26] See http://ess.nsd.uib.no.

stitutes for comparative research on social and political attitudes. Secondly, ESS also collects a rich array of social background variables, and thirdly, ESS constitutes an important methodological infrastructure, due to its innovative procedures for question formulation, translation, measurement, sampling and data access.

In the Netherlands, ESS is a project on the National roadmap and is funded by the National funding of the European Strategy Forum on Research Infrastructures (ESFRI).

DANS is responsible within the ESS-NL Task Group together with the consecutive National Coordinators to disseminate the use of the ESS data among both junior and senior researchers in the Netherlands.

5.9 European Holocaust Research Infrastructure (EHRI)[27]

The aim of European Holocaust Research Infrastructure (EHRI) is to "create a sustainable world-class Holocaust Research Infrastructure of European dimensions, which will bring together virtual resources from dispersed archives". Archives containing Holocaust-related materials are fragmented and scattered across the world, therefore making access to resources both complicated and time-consuming.

EHRI was launched in Brussels in November 2010 and will run from October 2010 for four years. This project is financed by the European Union under the 7th Framework Programme for Research and Technology Development. EHRI's main objective is to support the European Holocaust research community by providing an online portal that will give access as open as possible to dispersed sources relating to the Holocaust from all over Europe and Israel, making data available for Holocaust research around Europe and elsewhere into a cohesive body of resources. EHRI will also be encouraging collaborative research through the development of tools.

DANS role in EHRI is to contribute to the Standards and Guidelines[28] for participating archives and for EHRI itself, and to the Data Integration Infrastructure of tools, metadata and multilingual thesauri.

5.10 PersID[29]

The PersID initiative provides PIDs as well as a transparent policy and technical framework for using all kinds of scientific, cultural and other resources in the internet.

[27] http://www.ehri-project.eu.
[28] http://www.ehri-project.eu/partners-organisation.
[29] http://www.persid.org.

Eight national libraries and research institutions in six European countries have worked successfully together in the PersID project (from October 2009 to March 2011). The project was funded by the Dutch SURFShare programme, with a grant from the Knowledge Exchange programme.

The identifier system chosen in the PersID initiative is that of uniform resource names (URNs). The URN system encompasses the traditional bibliographic identifiers such as ISBN and ISSN, but also national bibliographic number (NBN). National libraries administrate the identifiers, which may be assigned to a wide variety of digital objects.

The PersID project partners agree in a letter of intent to use the results from the project for future cooperation, aiming minimally to keep the achieved situation up and running.

DANS is responsible for developing and building the Dutch URN:NBN resolver (http://www.persistent-identifier.nl/). A recent update of the resolver for the German-speaking countries (http://www.nbn-resolving.org/) successfully demonstrated how a URN:NBN identifier entered there was redirected to and resolved by the Dutch resolver.

5.11 Verteld Verleden, spoken testimonies online[30]

The goal of the Verteld Verleden (spoken testimonies) project is to make a start with a distributed approach for shared access to oral history collections and, in relation to that, formulating clear guidelines and best practices for owners of collections in order to get started with new technology.

As a basic principle, the participating institutions offer the metadata in XML format according to the Dublin Core metadata model. The metadata have been made accessible to the harvester of Verteld Verleden via the Open Archives Initiative (OAI) Protocol for Metadata Harvesting.

In the field of technology, existing standards are used within Verteld Verleden and open source components are re-used which have been developed within national and European research programmes (among which Catch, MultimediaN, MultiMATCH). The components include voice recognition, search features, thesaurus audiovisual archives and an interface component for visualizing words from a transcript in a cloud.

The project, which gets its funding from the regulation *Digitalisering met Beleid* (Digitizing with Policy), has a 2-year span. A web portal will be launched in 2012 which initially will make the oral history collections of the project partners accessible for the general audience.

[30] http://www.verteldverleden.org.

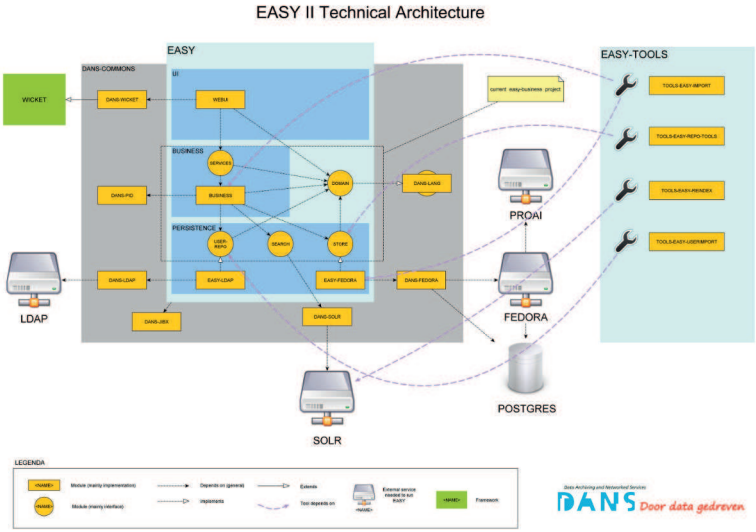

Figure E.1 The EASY system

6 DANS data research infrastructure

The data infrastructure at DANS is centred on its digital repository plus tools and additional support services that help researchers deposit and maintain research data. Self-archiving services targeted to specific communities of researchers, such as archaeologists, are built on top of the of the repository infrastructure.

6.1 EASY[31]

The Electronic Archiving SYstem (EASY) is a pivotal component within the repository infrastructure of DANS (Figure E.1). EASY is based on the repository system Fedora and it has been developed to:

- facilitate self-archiving by researchers;
- enable data curation and management by in-house data experts;
- facilitate publication of data.

Registration is mandatory, but open to anyone interested. When registered, users can archive their data by entering metadata and uploading the accompanying raw data-files. After review by one of the data experts, the data are published on the EASY website from where they can be downloaded.

[31] https://easy.dans.knaw.nl.

Although EASY advocates Open Access, one has to keep in mind that researchers may have reservations with making their data available to just anyone. For this reason, those who are hesitant to release their data into the public domain can either impose an embargo period, during which access is restricted, or take full control over who can download their data by granting access based on individual permission requests.

Metadata are always publicly available, even to those who have not registered as a user in EASY. They are published through the search and browse interface available in the web application, as well as through an OAI interface. Anyone is free to harvest this OAI data and do with the metadata whatever he likes. In order to enable more control of what is actually harvested, service providers can limit their queries to specific disciplines, collections or metadata formats.

Future plans with regard to the dissemination of data include providing (programmatic) access to the data itself, for instance by publishing an API to inspect the data as RDF triples and connecting EASY to Open Linked Data initiatives. This would require a more format/discipline-specific set of content models, that will tell the system how a dataset is actually structured.

A new version of EASY (EASY-II) has been released in September 2011.

6.2 e-Depot Dutch Archaeology (EDNA)[32]

The e-Depot Nederlandse Archeologie (EDNA) pilot project was initiated in September 2004 and ran until February 2006 with funding from SURFfoundation.[33] In 2007, the setting up of EDNA was followed up by the retrospective archiving project EDNA II, which is collaboration between DANS and the Rijksdienst voor het Cultureel Erfgoed.[34] The Dutch archaeology e-depot of digital grey literature and research data is located at DANS, using the EASY self-archiving system. There are currently over 15,000 datasets deposited in EDNA, with some 12,000 being publications only.

The aim of EDNA is to highlight, for Dutch archaeologists, the importance of durable archiving of digital data generated through archaeological research. There is a legal requirement for all archaeological finds and analogue documentation to be deposited with a provincial depot after the completion of the research.

The deposition of data and literature is primarily a process of self-deposition with the depositors adding the metadata themselves. However, there are

[32] http://www.dans.knaw.nl/en/content/categorieen/projecten/edna-e-depot-dutch-archeology.
[33] http://www.surffoundation.nl/en.
[34] http://www.racm.nl.

additional checks made by archivists at DANS. Additional documentation about the methodology used in the research is also deposited.

A PID is assigned to the research project as a whole, but currently no PIDs are being assigned to each individual document.[35]

EDNA has more levels of restriction to access the data archived than is normally part of EASY. These levels of restriction, from the most open to the most restricted are:

- Open Access, but not anonymous as the data user must log on. The registration to use EDNA is minimal and consists of name, password and e-mail address;
- professional archaeologists, including archaeologists working for companies, government agencies and students (for educational use only) and validity of membership is checked;
- personal access, for a researcher first requesting access, which must be agreed and confirmed by the depositor. This provides the depositor with control over their data and who might have access to it. The professional archaeology community is small in the Netherlands and therefore this control over access is easy to manage. It is believed that this control may help to limit access to important information about archaeological monuments to treasure seekers;
- no access (to private data), for example data that have been deposited during an excavation but the literature has yet to be published.

6.3 Persistent identifier services[36]

As organizations and researcher more and more tend to add PIDs to datasets and publications they have produced, the need for services based on these PIDs are growing. The main service, developed for the Netherlands by DANS, is the availability of a so-called resolver (http://persistent-identifier.nl/), which can be used to detect the actual URI of a resource with a specific PI. Of course, combining the national resolver on a European level can augment the value of this service. DANS is cooperating with both an Italian and a German meta-resolver.

[35] This level of granularity of PIDs will occur in the imminent migration to the next release of EASY.

[36] http://www.persistent-identifier.nl.

6.4 Migration to Intermediate XML for Electronic Data (MIXED)[37]

Migration to Intermediate XML for Electronic Data (MIXED) contributes to digital preservation by dealing with the problem of file formats. Over time, file formats become obsolete. When that happens, the information in such file types is no longer accessible. MIXED follows the strategy of converting files to XML as soon as possible, preferably when data are ingested into an archive, such as EASY. As a service, MIXED can convert files with tabular data (spread sheets and databases) to softwareindependent XML for long-term preservation. The XML can be converted back to the original software dependent format, or to formats of other suppliers of software, or to new formats in the future. This approach solves or alleviates two problems of ordinary migration: (i) it diminishes the need for repeated migrations considerably, because it migrates out of the version sequence of application-bound file formats; and (ii) it facilitates interoperability of data that have been created in different file formats, because they all will be translated into application independent XML.

MIXED consists of a framework plus plug ins. Plug ins take care of the conversions between application file formats and application independent XML formats. At present MIXED can handle these file formats:

- Data Perfect;
- Access 2000 and 2002;
- dBase III and IV;
- Excel 2003.

MIXED is used in the DANS ingest and dissemination workflow, but it is public software. Parts of it are already published in open source repositories and other parts will follow.

6.5 National Academic Research and Collaborations Information System (NARCIS)[38]

National Academic Research and Collaborations Information System (NARCIS) is the national Dutch portal for information about researchers and their work. NARCIS has been a service of DANS as from February 2011. Based on a user survey conducted in 2009,[39] most of its users come from universities and scientific institutions. The number of users is about 1 million per year.

[37] https://sites.google.com/a/datanetworkservice.nl/mixed.
[38] http://www.narcis.nl.
[39] http://depot.knaw.nl/5662/2/What_are_your_information_needs_Elpub_2010.pdf.

As a portal, NARCIS is collecting information from different types of data providers: metadata from repositories, metadata from EASY and descriptions of research institutions, researchers with their expertise and research projects. Besides, NARCIS acts as an access point to scientific news feeds.

Harvesting is the main aspect of the NARCIS system. NARCIS acts as a service provider but can also act as *a data provider to internationally operating services providers like DRIVER and WorldScientific (links!). In fact, you could see NARCIS like a kind of a national aggregator*

One of the most important developments in NARCIS is the implementation of identifiers for objects and researchers. In the Netherlands, digital author identifiers (DAIs) are assigned to researchers (authors). OCLC maintains the central database with DAIs. The Dutch scientific institutions are using a special name space for the DAIs: the eu-repo namespace, to identify information assets used in the European Research Libraries. Information on the eu-repo name space can be found at http://info-uri.info/registry/OAIHandler?verb=GetRecord&metadataPrefix=reg&identifier=info:eu-repo/

Apart from the DAI, NARCIS is also showing PIDs of the objects (publications and datasets) in the repositories of the scientific institutions. Although the institutions are free in choosing the system to add PIDs to objects, the decision has made that at least the URNs will be used (see http://tools.ietf.org/html/rfc3188 and 5.10 PersID).

An elaboration of the existing model is the inclusion of description of enhanced publications in NARCIS. An enhanced publication is a composed object, for instance a publication enhanced with other information like the dataset that has been used in writing this very publication. The metadata of these enhanced publications are described in so-called resource maps, using OAI Object Reuse and Exchange.[40]

6.6 DANS EASY online analysis tool[41]

DANS EASY online analysis tool offers its users additional features in comparison with the standard EASY. The tool is to be used within the social sciences. This service allows the searching, browsing, analysing and downloading of social science data. Major difference with EASY is the possibility to create online tables based on the well-documented datasets. With DANS EASY online analysis tool, data may be analysed online, for instance using regression analysis. This feature is typical for this service that as a matter of facts has been derived from Nesstar[42]

[40] http://www.openarchives.org/ore.
[41] http://194.171.144.69/webview.
[42] http://www.nesstar.com/

DANS EASY online analysis tool is available to a small subset of the social sciences datasets in EASY, namely:

- Cultural Changes in the Netherlands Studies (CV);
- Dutch Parliamentary Election Studies (NKO);
- Social and cultural trends in the Netherlands (SOCON);
- National survey pupils secondary schools (NSO);
- Facilities use survey (AVO);
- Time-budget survey (TBO).

6.7 Netherlands' Geographical Information System (NLGis)[43]

Netherlands' Geographical Information System (NLGis) is a DANS service in which historians can reproduce and visualize regional variation in Dutch historical municipal data, based on data from the last two centuries.

NLGis is a web application that supports the spatial component in historical research. GIS plays an important role in this kind of research. Researchers may upload, display and download the map with historical municipal data.

6.8 DANS data support services and policies, standards and guidelines

DANS promotes the permanent storage and traceability of research data. To this end it provides, among others, practical services to researchers and research groups. Data from numerous resources is made freely available by or through the mediation of DANS. Besides, DANS organizes on request symposia, subsidizes small data projects and carries out ICT activities on behalf of various research projects.

DANS can guarantee permanent access for all standard and preferred formats; in the case of irregular data formats, access is dependent on future developments in software.

Researchers may use consultancy services (data consultancy) with regard to scientific data processing. Merely offering an archiving system is not enough. Therefore, consultancy services have been developed in order to encourage the use of EASY or improve the quality of research data. Examples are:

- data deposit guidebooks for Archaeology, Social Sciences And History;
- help texts in EASY;
- courses and presentations (text material and powerpoints stored on internal DANS website, accessible to DANS employees only);

[43] http://www.dans.knaw.nl/en/content/categorieen/diensten/dutch-geographic-information-system-nlgis.

– communication with depositors.

DANS puts a lot of care and effort into ensuring that data deposits come with good metadata. We discern between three types of metadata: (i) project-specific metadata (Dublin Core); (ii) file-specific metadata; and (iii) metadata on the level of variables (codebooks).

File-specific metadata needs to be sent as a file list together with the dataset. The data deposit guidebook contains a reference table for other variables the depositor can choose to describe the files with. The guidebook texts will be revised in 2011.

With regard to archaeological files, we have agreed on a hybrid division of tasks. We have made arrangements with the archaeological field that archaeological data must be made accessible at DANS.

The guidebooks state that DANS archivists do not change the contents of the file. Archivists only convert files to preferred formats, check the (meta)-data and assign file and dataset rights (publish files). There is no general policy for archivists at DANS on converting files; however, these are common migrations that DANS currently performs:

– word processor documents to PDF/A;
– images to jpeg and tiff;
– vector images to PDF/A and SVG;
– Geographical Information System files to Mid/Mif (MapInfo export format);
– Computer Aided Design (CAD) to DXF version r12 (AutoCAD);
– spreadsheets to CSV (datatables) or PDF/A (reports);
– databases and data tables (dbf) to CSV;
– video to MPEG-4.

6.9 Data seal of approval[44]

Within DANS, a seal of approval for data has been developed to ensure that archived data can still be found, understood and used in the future.[45] In 2008, the first edition of the Data Seal of Approval, written by Laurents Sesink, René van Horik and Henk Harmsen, was presented in an international conference. In spring 2009, the Data Seal of Approval was handed over to an international Board.

[44] http://www.datasealofapproval.org.

[45] DANS – Data Archiving and Networked Services – is an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW), and is also supported by the Netherlands Organisation for Scientific Research (NWO). Since its establishment in 2005, DANS has been providing storage of and continuous access to research data in the social sciences and humanities.

The Data Seal of Approval and its quality guidelines are of interest to research institutions, organizations that archive data and to users of that data. It can be granted to any repository that applies for it via the online assessment procedure.

The criteria for assigning the Data Seal of Approval to data repositories are in accordance with and fit in with national and international guidelines for digital data archiving such as Kriterienkatalog vertrauenswürdige digitale Langzeitarchive, as developed by NESTOR;[46] Digital Repository Audit Method Based on Risk Assessment (DRAMBORA), published by the Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE);[47] and Trustworthy Repositories Audit and Certification (TRAC): Criteria and Checklist of the Research Library Group (RLG).[48] Furthermore the following has been taken into account: Foundations of Modern Language Resource Archives of the Max Planck Institute[49] and Stewardship of Digital Research Data: A Framework of Principles and Guidelines published by the Research Information Network.[50] The guidelines in this document can be seen as a minimum set distilled from the above proposals.

Fundamental to the guidelines are five criteria for digital research data, which together determine whether or not it may be qualified as sustainably archived:

- available on the internet;
- accessible, while taking into account relevant legislation with regard to personal information and intellectual property of the data;
- available in a usable format;
- reliable;
- citable.

The associated guidelines relate to the implementation of these criteria and focus on three stakeholders:

- the data producer is responsible for the quality of the digital research data;
- the data repository is responsible for the quality of storage and availability of the data: data management;
- the data consumer is responsible for the quality of use of the digital research data.

[46] http://edoc.hu-berlin.de/docviews/abstract.php?id=27249.
[47] http://www.digitalpreservationeurope.eu/announcements/drambora.
[48] http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91.
[49] Peter Wittenburg, Daan Broeder, Wolfgang Klein, Stephen Levinson and Laurent Romary. http://www.lat-mpi.eu/papers/papers-2006/general-archive-paper-v4.pdf.
[50] http://www.rin.ac.uk/data-principles.

More information can be found on the Data Seal of Approval website: www.datasealofapproval.org.

6.10 Repository audit and certification (trustworthy digital repository)

On a European level there many data repositories that can cooperate in a network. But these data repositories may differ in technical level. However, users need common guidelines. Important developments are going on. Apart from the ESA European LTDP Common Guidelines (earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_Issue1.1.pdf), a similar approach is proposed in the European Framework for Audit and Certification of Digital Repositories, which was outlined in a Memorandum of Understanding between CCSDS, DANS and DIN.[51] This framework defines three levels of trustworthiness:

- **basic certification:** granted to repositories which obtain Data Seal of Approval (DSA) certification;
- **extended certification:** granted to Basic Certification repositories which in addition perform a structured, externally reviewed and publicly available self-audit based on ISO 16363 or DIN 31644;
- **formal certification:** granted to repositories which in addition to Basic Certification obtain full external audit and certification based on ISO 16363 or equivalent DIN 31644.

The granting of these certificates will allow repositories to show one of three symbols (to be agreed) on their web pages and other documentation, in addition to any other DSA, DIN or ISO certification marks.

6.11 DANS literature publishing infrastructure

DANS participates in the electronic newsletters "Archive Letter" and the quarterly journal "e-Data and Research". Notable recent acquisitions/published datasets are brought to extra attention in these newsletters. Links to these newsletters and their archives are given on the DANS homepage.

[51] Giaretta, D, Harmsen, H, and Keitel, C. "Memorandum of understanding to create a european framework for audit and certification of digital repositories". 2010. Available at http://www.datasealofapproval.org/sites/default/files/20100709_020_signedMoUtocreate aEuropeanFrameworkforAuditandCertificationofDigitalRepositories.pdf.

7 Lifecycles and scholarly primitives in the humanities and social sciences

DANS is involved in a wide variety of research projects in the humanities and social sciences. Therefore, to identify points of commonality in disparate research practices and methodologies where Open Access infrastructures are potentially utilized by researchers to support their practice, we reviewed a number of data and research lifecycles to identify one that could be utilized. The aim was to develop a framework for structured interviews of scholars and identify phases within the lifecycle where data and literature are produced and consumed. These phases may appear obvious until one considers that in some humanities disciplines research literature is the source material for further research questions, and data from Social Science surveys in one discipline can be source data in another.

The Scholarly Communication Life Cycle[52] (Microsoft, 2008) identifies four phases to a research lifecycle, plus two activities common to all. The phases are:

- data collection, research and analysis;
- authoring;
- publication and dissemination;
- storage, archiving and preservation.

Collaboration and discoverability, additional cross-cutting activities, augment all phases of the lifecycle. The addition of collaboration as a feature and need across the lifecycle is not seen in other data or research lifecycles. Although the concept of the "lone humanities scholar" is often quoted,[53][54] the activities of collaboration and communication, throughout the whole of the research process, must be considered central to any Open Access infrastructure.

The British Library also present a similar four-phase research lifecycle, consisting of:

- idea, discovery, design;
- obtain funding;
- experiment, collaborate, analyse;
- disseminate findings.[55]

Although essential, "obtain funding" maybe considered as out-of-scope for the purpose of an Open Access infrastructure for the humanities and social sciences. Furthermore it seems somewhat incongruous that "collaborate" is

[52] http://research.microsoft.com/en-us/about/msr_scholarlycom.pdf.
[53] http://openreflections.files.wordpress.com/2008/10/talk-communia-20102.doc.
[54] www.ahds.ac.uk/e-science/documents/Robinson-report.pdf.
[55] Newbold, 2008,

only in a single phase of the lifecycle and that there is no mention of archiving, preservation or publication.[56]

The DARIAH research lifecycle model (DARIAH, 2010), based primarily upon the previous two examples but also others, provides a simplified combination of both research and data lifecycles (Figure E.2). Search and discovery for research resources is a key feature of this lifecycle as is gathering (or collecting) these resources into an environment where analysis and experimentation can take place (DARIAH, 2010). The addition of "share" at the centre of the lifecycle, in addition to collaborate, implies that data and literature is made available (publicly or to a select group during early research phases) in every phase. This should be considered as an important feature in an Open Access infrastructure. Although *archiving* of a research data and literature should be an integral phase in the lifecycle, occurring as it does after *publishing,* non-permanent *storing* of collected research material and data created occurs at all stages of the lifecycle.

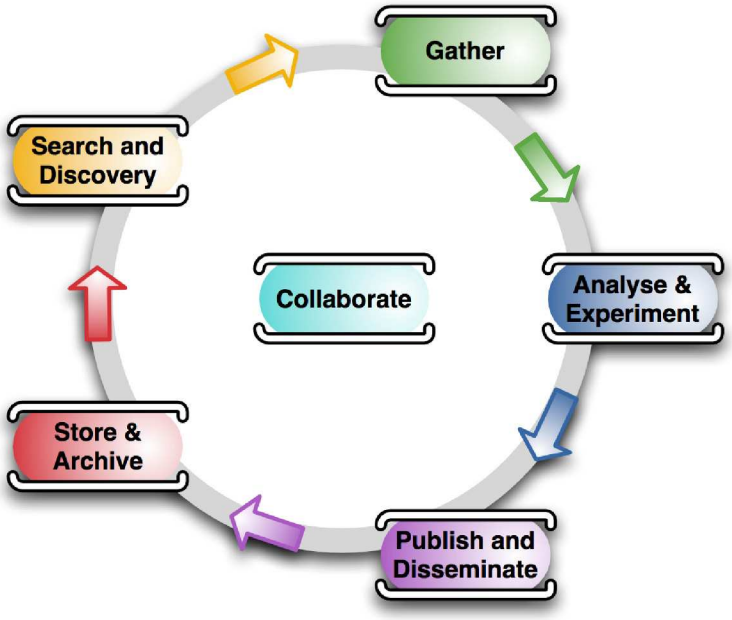

Figure E.2 The DARIAH research and data lifecycle

[56] Publication can be inferred as a form of dissemination.

8 Challenges, opportunities and trends

As an organization promoting storage, curation and access to datasets, DANS will be confronted with some major changes in the next decade.

8.1 E-research

One of the biggest challenges is the rise of e-research: data-intensive research. In this type of research, researchers will rely on the quality of and the access to data. The role of DANS is clear: on the one hand it may start to serve as a central access point to datasets from all kinds of disciplines. On the other hand it plays an important role in the development of regulation of the deposit of data, so that researchers are confident that the data accessed via DANS are valuable. Here one can see a close relationship with data curation.

In e-research, the boundaries between the different disciplines tend to blur. e-Scholars often want to combine data from the humanities and social sciences with data from for instance biology and geography. It certainly will be a major challenge to DANS to realize data curation and data access to (for DANS) non-traditional disciplines (such as the natural sciences) as well. By close cooperation with the Dutch technical universities, the scope may be broadened to the technical disciplines as well.

8.2 DANS as a research organization

DANS could continue to act as a pure service-oriented organization. By doing so, it will take the risk to miss important development in the field of e-research. It would be better to change the scope of the institute in such a way that participating in research will become possible. In other words, for the improvement of the data infrastructure, it will be advantageous to have (e-)scholars working in the institute. Such a group could for instance be involved in research in standardization of (meta-)data and in a kind of trend watching in the data-intensive research field.

8.3 Grid

Apart from this, within DANS a research focus will have to be on giving access to data. Developments in Grid and Cloud computing have hardly been studied with this respect. What is for instance the influence of Cloud computing on costs, trust and quality of data? How will researchers couple data from different sources to each other using the cloud and will these couplings be sustainable?

8.4 Software for data access

Another aspect of data preservation is the fact that more and more datasets cannot be used without special applications that have been developed for these very datasets. What could be the role of DANS in giving access to these data? Would it be necessary to preserve the application software as well or would it be possible to develop an application independent archival system?

8.5 Enhanced publications

An institute as DANS has to be prepared to cope with the developments in the fields of semantic web, "deep access", linked data and RDF. Unlike the examples above, these development will make it possible to combine traditional publications with datasets, audio fragments, software and so on: enhanced publications. Enhanced publications will influence the scholarly communication process in a dramatic way, on the condition that all their components will be Open Access. Otherwise, a frustrating, partially closed, data infrastructure will be developed.

8.6 Scientometrics

Giving access to datasets will also change impact measurements for institutes and researchers. Until now, scientometrics was only based on traditional publications. When re-use of data(sets) will be measured, this will give an additional impact to measure the impact of a specific research.

8.7 Linking of datasets

As already is common in a field like astronomy, linking of datasets available in different data centres (in different countries) will give scholars the opportunity to combine these datasets in, for instance, secondary analysis, data mining and visualization.

9 List of figures

10 List of tables