The IASSIST QUARTERLY represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

Information for Authors
The QUARTERLY is normally published four times per year. Authors are encouraged to submit papers as word processing files. Hard copy submissions may be required in some instances. Manuscripts should be sent to Editor: Karsten Boye Rasmussen.

The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press Manual of Style or Kate L. Turabian's Manual for Writers, Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". Journal of the American Society for Information Science 30(2):77-82, March 1979. Announcements of conferences, training sessions, or the like, are welcomed and should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event.

Editor
Karsten Boye Rasmussen
Marketing & Management
SDU
University of Southern Denmark
Campusvej 55, DK-5230 Odense M, Denmark
Phone: +45 6550 2115
Email: kbr@sam.sdu.dk

Production
Walter Piovesan
Maps/Data/GI Library
Simon Fraser University
Burnaby, B.C.
Canada V5A 1S6
Phone: (778) 782-5869
Email: walter@sfu.ca

FEATURES
4 Editor's Notes
Karsten Boye Rasmussen

6 Data Documentation and Remote Computing at the International Data Service Center of IZA
N. Askitas

12 A Documentation Model for Comparative Research Based on Harmonization Strategies
John Kallas & Apostolos Linardis

26 Strengthening Data Security: an Holistic Approach
Sharon Bolton & Matthew Woollard

30 Swedish National Data Service’s Strategy for Sharing and Mediating Data
Carina Carlhed & Iris Alfredsson
Welcome to the IASSIST Quarterly (IQ) volume 32 2008. We have collected the 1, 2, 3, and 4 issues into a single issue for 2008 in order to catch up with our schedule.

In the text below - mostly by cutting and pasting - I am giving a short appetizer for the articles in this issue of the IQ. This type of editorial is among the few areas where plagiarism is actually welcomed.

Nikos Askitas is the head of the International Data Service Center of the Institute for the Study of Labor in Germany (IZA). At the 2008 IASSIST conference he presented what is here an article on the “Data Documentation and Remote Computing at the International Data Service Center”. The data documentation of the IDSC that started with translation of German metadata into English has developed into a detailed, in depth, searchable and standardized information service, especially helpful for comparative research. The datasets are in the areas: Employment and Wages, Education and Training, and Demographics and Migration. The documentation is available in HTML, PDF, and as DDI-files. This documentation production is explained in the first part of the article. In the second part of the article the IDSC experience with “remote computing” is described. Germany uses the concept of “factual anonymization” and the production of “scientific use files”. However, such files are not allowed for export. Instead IDSC supplies interfaces to scientists with both local and remote support for which IDSC has developed special software (JoSuA).

The article “A Documentation Model for Comparative Research Based on Harmonization Strategies” by John Kallas from University of the Aegean at Mytilene on Lesvos, Greece and Apostolos Linardis from the National Centre for Social Research in Athens, is proposing a documentation model for both longitudinal and cross-cultural studies. Different harmonization strategies are examined and three documentation models are proposed. The authors have chosen the term “cross-cultural” rather than “cross-national” as cultural discrepancies may exist within the same nation. The article underlines the importance of the data element, the concept, the universe, and the classification as they are study components where even small changes may affect the overall comparability. This is leading to looking at the stages for the different types of harmonization strategies: ex ante input, ex ante output, and ex post. Most documentation processes at data archives are ex post harmonization. The authors are aware that the proposed study documentation procedure is laborious for the researchers; however, the positive side is the benefits in searching and locating the data.

At the 2009 IASSIST conference in the session “Sharing Data: High Rewards, Formidable Barriers” Carina Carlhed and Iris Alfredsson from respectively Mälardalen University, Sweden and the Swedish National Data Service (SND) presented a report from an investigation carried out earlier in 2009. The report has been turned into an article for the IQ with the title: “Swedish National Data Service’s Strategy for Sharing and Mediating Data. Practices of Open Access to and Reuse of Research Data - The State of the Art in Sweden 2009”. The report is based upon a joint project between SND and four university libraries that carried out a national survey of existing databases and database research, as well as attitudes towards data sharing among researchers. This was carried out by email questionnaires sent to professors and doctoral students. In general the results show that doctoral students expressed great uncertainty about questions of amounts of reusable digital data, while professors emphasize lack of resources for researchers to document and make their data accessible for others. The groups consider the most effective interventions for enhancing accessibility to digital data to be that research grants should include funds for preparing the data for sharing and archiving, and that archiving data for use by the scientific community is acknowledged to be of scientific merit. A similar study was carried out in Finland and compared to this Swedish study. We hope to present the Finnish study in a later issue of the IQ. The Swedish Research Council founded in 2006 a Database Infrastructure Committee (DISC) to promote...
the development of an effective infrastructure for sharing research data. A product of this initiative has been the formation of the Swedish National Data Service (SND) that also is described in the article. The article further describes the procedures of the surveys and there might be followers for doing similar user investigations among other data organizations. The survey contains questions as to the knowledge of plans such as the roadmap “The Swedish Research Council’s Guide to Infrastructure” (2007) and the “OECD Guidelines on Open Access to Research Data from Public Funding” (2007). Answers to these questions exhibited a low level of knowledge, as did questions about making own data available. Read more in the article, and also about reasons given for not reusing digital data, and the seven suggested obstacles to sharing digital data.

Should you be interested in compiling issues for the IQ as guest editor(s) please contact me. If you don’t have anything to offer right now, then please prepare yourselves for the coming IASSIST 2010 if you are acting as chair for a session there. That is an obvious opportunity to bring quality sessions to more people than the session participants and also making the memory more sticky.


Articles for the IASSIST Quarterly are very welcome. Articles can be papers from IASSIST conferences, from other conferences, from local presentations, discussion input, etc. Contact the editor via e-mail: kbr@sam.sdu.dk.

Karsten Boye Rasmussen October 2009
Abstract:
The International Data Service Center (henceforth “IDSC”) of the Institute for the Study of Labor (henceforth the “Institute” or “IZA”) is the Institute’s organizational unit which serves its data needs as well as those of its various affiliated communities and the ambient associated research community at large. Since data are at the core of the Institute’s makeup, its IDSC (idsc.iza.org) is a major contributor to its vision of a virtual institute. Besides a short presentation of the IDSC we focus on two of its major work areas: data documentation and remote computing.

Keywords: metadata, documentation, remote computing, remote processing, data enclave

1. Introduction
The heritage of IDSC goes back to the 1990’s when a group of German economists, including IZA’s director K. F. Zimmermann, seeded an intense discussion in Germany about providing the scientific community with a better data infrastructure (R. Hauser, G. G. Wagner and K. F. Zimmermann, 1998). Commissioned by the Federal Ministry of Education and Research, the KVI made several recommendations on how to improve the scientific data infrastructure in Germany (KVI, 2001). These recommendations lead to the establishment of several research and service data centers, among which IdZA (IDSC’s predecessor) was the only partner representing labor economics. The German speaking reader may consult (H. Schneider and C. Wolf, 2008) for a good historical write up. After a successful evaluation of the first three-year pilot phase which started in 2003, a second phase of financing was awarded to IZA (2008-2010). The IDSC in its new form represents IZA’s continued commitment to the field.

The IDSC is currently undergoing a drastic restructuring of its assets as well as a broadening of its mission. The Center is now building partnerships with other players in the field in order to leverage complementary competence and build mutually beneficial long term relationships and alliances. The main components of the Center’s realignment are all contained in the Center’s acronym:

1. IDSC is International. It operates without national borders or other artificial frontiers, draws its know-how from the wide international arena and aims at catering to the international research community always in close alignment with the Institute’s vision of a virtual institute.

2. IDSC is about Data. This mostly means the technology of data but also involves ethics, legal, educational, and other aspects. IDSC develops, applies, and integrates know-how aimed at dealing with data in the context of operation of IZA. A core component of the work of the IDSC is about inventing, developing, integrating, and deploying/promoting solutions for computing with data and in particular with “difficult”, i.e. highly sensitive/confidential, data.

3. IDSC is about Service. IDSC services the data needs of the IZA resident research community, the various global and virtual IZA research communities (Fellow and Affiliate networks, etc.) and the research community at large in that order. The meaning of the order is dual: on the one hand it expresses priority in the sense that IDSC serves the local community first and the remote ones afterwards; on the other hand it expresses deployment order in the sense that local deployment is a preparation step for the large scale deployment. In that sense the IDSC finds itself in the privileged position of belonging to an ideal ambient environment in which to incubate ideas on technology applications.

4. IDSC is a Center. This means that besides being an organizational unit of IZA, it is an entity of its own with relationships to the other IZA units and the world. It is also a center in the sense that it aims to become the focal point of an International, Data-related, Service oriented network of economically-minded technologists and technologically-savvy economists. It also aims to become the ubiquitous place for scientists to look for data support, data access support and data services with emphasis on labor economics but also beyond.

This paper will focus on two important areas of work of the center: data documentation (Section 2) and remote computing (Section 3) and will close with a short mention
of future plans and challenges (Section 4).

2. Data Documentation
This work area is deeply rooted into the legacy of the IDSC and represents a significant area of activity in the Center’s body of work. Its original purpose was to translate (and along the way standardize) official German metadata in order to improve its usage in the scientific literature. The premise was that contrary to the country’s standing as a world-wide top exporting country, the amount of scientific literature based on German official data was rather poor. The IDSC took up the ambitious task of remedying this situation by improving on what it saw as its main cause: the lack of accessible documentation. The metadata offering of the IDSC, which by now goes well beyond translating German metadata into English, consists of a detailed, in depth, searchable and standardized information service, especially helpful for comparative research, which includes an ever growing number of datasets. There are currently datasets in the areas of:

- Employment and Wages
- Education and Training
- Demographics and Migration

For the end user the most important features of the metadata offering of the IDSC in its current implementation (http://idsc.iza.org/metadata/) are:

1. Every dataset has a searchable HTML presentation.
2. Every dataset has its metadata in PDF book form.
3. For every dataset its origin and how to reach it as well as the IZA discussion papers which have used it are included.

For a data professional, it is important to know that the metadata are saved in DDI form (currently version 2) and the DDI files are publicly available for download. Anyone is free to use these DDI files for their own presentation so long as the IDSC is properly cited and the DDI files are made available in the same way. So far as we know the DDI files of the IDSC are currently the only publicly available, variable-level DDI files. The reader may verify how difficult it is to google her/his way to real instances of DDI files anywhere outside IDSC!

Based on these DDI files and using open source and community tools such as the IHSN Microdata Management Toolkit a static HTML presentation is produced for each dataset. Since data documentation is a document like any other and since once compiled it remains largely unchanged the use of relational database supported metadata systems appears to be uncalled for in this context. The most essential element of dynamic implementations is recovered by indexing the static pages so as to make them searchable. This solution is very efficient and performs and scales very well. In what follows the main ingredients of this solution are described by discussing the work process involved in documenting a dataset.

The documentation work typically starts with a DDI file (version 2) which is produced using the Metadata Editor which comes with the IHSN Microdata Management Toolkit. The metadata which flows into such a document is collected in a variety of ways from a variety of sources and data formats all of which are case specific. By using a version of the CD-ROM builder from the IHSN toolkit, HTML and PDF presentations are produced. This version of the CD-ROM builder is modified to include the IDSC branding. Using the keyword attribute of DDI version 2, relevance and context keywords are attached to datasets and to their variables. These keywords belong to a concept hierarchy derived from the HASSET. The keywords are used to make the metadata searchable. The concept hierarchy has a HTML presentation in its own right which is based on its implementation in a relational database. This concept hierarchy module accompanies the searches and maybe used by the user in order to perturb the scope of a search. A search for variables on “Wages” for example will return in addition to the variables that match, the conceptual neighborhood for the concept: synonyms such as “earnings”, “pay” or “remuneration” broader concepts such as “income” but also narrower terms such as “low pay” as well as related concepts such “wages policy” all in the form of markup encoded searches accompanied by the number of hits behind the associated search. In effect presented this way the concept hierarchy becomes a kind of directory structure containing relevant results. The search may be restricted to variable pages, variable group pages, dataset overview pages, or dataset dictionary pages. The IHSN toolkit’s CD-ROM builder conveniently produces for each dataset: variable pages on which a full description of the variable is presented, group pages on which groups of related variables are presented as well as dictionary or overview pages which summarize the dataset’s focus. Making these datasets searchable separately is useful for achieving different ends. Searching dictionary or overview pages may be used to locate variables whereas search group and variable pages may be used to locate variables across datasets.

The indexing and the search of these static pages are done using swish-e (swish-e.org). The indexing occurs once every night: once a dataset is staged by being written into the publishing area of the web service it becomes available to the indexer without further action necessary. A search API is built on top of swish-e which may be used by anyone wishing to integrate a search of the Center’s metadata inventory in their own presentations or other data products. An application of the search API is the Stata module of the author (Askitas, 2009) which may be installed in Stata by running “ssc install metadata”. The module integrates metadata right into Stata and attempts
to bring DDI into the realm of a widely used standard econometric application. An RSS feed enriched with Dublin Core elements is produced programmatically out of the DDI files. This feed is then used to produce a listing of the metadata using Stata’s own web capabilities by translating it to SMCL (Stata’s own markup language) on the fly. The search API is used to locate datasets relevant to a keyword. Subsequently Dublin Core files (derived from DDI) of these datasets are used to produce the SMCL presentation of the datasets we found. We are planning on integrating sample datasets right into this presentation so the user can search for and program against data in an integrated fashion.

The entire metadata offering of the IDSC is hooked to its own web analytics based on the open source Piwik (piwik.org) so that we can produce access and usage statistics of the metadata offered all the way to the variable level pages.

We plan to integrate metadata into a news aggregator for economics to allow, for example, researchers searching for grants to find the datasets to use in their project proposals. The idea here is to promote metadata right into one of the main activities during which a researcher looks for new data: projects proposals. In order to achieve the merging of news with metadata, news items are tagged with the same keywords as the metadata.

3. Remote Computing
Computing with data has never been as exciting and powerful as it is today. It has also never been as necessary or as ridden with issues and problems. As computing capacity is expanding and large amounts of data can be analyzed faster by empirical data analysts, research projects tend to become increasingly cross sectional and interdisciplinary and this results in more complex computing circumstances. In the past computing capacity and data were collocated in computing centers and there was hence basically one way to compute with it: on site. As computing devices proliferate, their mobility and capacity no longer excluding each other, constant connectivity is becoming the rule and generally computing is democratized this is no longer the case, hence: complex computing circumstances. Researchers can say more and more about the world by means of data based empirical research and they can, want or need to do it with more data from more locations.

On the other hand the demand for empirical research is increasing as the world gains in complexity and is increasingly thought of as a system expressed in equations and measured by variables. In this world privacy, disclosure and data protection acquire a new importance, complexity and perhaps interpretation.

Some of the main stakeholders in this world of data suitable for empirical research are collectors, producers, owners, data custodians and data analysts. Data collectors, producers and owners are usually the side which creates the data. This is where activities such as field work, data reorganization, quality control, definition of access rules etc take place. Data custodians are usually the owners of the data but this is not necessary. Custodianship is transferable by means for contractual agreements. Lastly data analysts are the people who use the data to create knowledge necessary for society to base policy decisions on. In an “ideal world” data is being produced and cared for, custodianship is smoothly and securely regulated and researchers get unlimited access to the data and are able to produce research results without friction. Friction in this setup is created due to the legal requirements, deficits or variability (across countries for example) thereof, the need for privacy protection and disclosure control and the complexity of the computing circumstances. This is the context in which the IDSC’s work on computing from afar is taking place.

In Germany the concept of “factual anonymization” is widely accepted since the 90s in effect enabling the creation of so called “scientific use files” which may sometimes be given to the researcher. These files are basically samples of the data which interest researchers and their being “factual anonymized” means that deanonymization is computationally sufficiently expensive. This came as a response to the increasing complexity of the “computing circumstances” mentioned above.

These files however may not be sent abroad and do not always entirely cover a project’s data needs. In these cases on site computation is still necessary and this is where the work of IDSC comes into the picture. The Center runs its own data enclave which both conforms to the strictest data security standards and yet strives to achieve the highest possible degree of scientific freedom. To achieve this, IDSC applies a properly stratified way to interface with the scientists working in it:

- Locally through a contained “ultra thin” network segment
- Remotely via several tools in its remote computing portfolio.

Several Research Projects based on highly sensitive datasets are currently hosted within the IDSC Data Enclave. Some of these are:

The Center offers the possibility for computing from afar whenever it is a lawful data custodian and there are no regulations prohibiting it from doing so. One of the tools used by the center for enabling remote computing is JoSuA which was conceived and developed at the IDSC of IZA. Originally designed in order to grant international researchers access to German labor market data, JoSuA has matured into a flexible data analysis instrument with a configurable degree of automation and is designed to fit the needs and specifications of each individual data provider.

JoSuA is therefore suitable for use by data providers who own such data and wish to make it available to a larger research community without jeopardizing the security of their data. Data providers may use JoSuA in one of the following ways:

- Install own JoSuA instance
- Host the data at the IDSC Data Enclave
- Use Hosted JoSuA

In all three flavors JoSuA allows the data provider to maintain full control of the output’s censoring. JoSuA is used by several partners of the IDSC: the IAB (http://fdz.iab.de) which has its own installed instance, currently in test phase; the IQB (http://www.iqb.hu-berlin.de/), which hosts its data at the IDSC Enclave. Discussions with a number of other partners are in progress regarding adoption of JoSuA for their own data and/or research communities. As of this writing JoSuA is undergoing a production streamlining in order to prepare it for a larger install base: versioning, Service Level Agreements, feature management are some of the issues which need to be solved. In the remainder of this section we would like to give a brief summary of JoSuA aimed at both data analysts and data custodians hoping to have more product material to show for soon.

A typical scenario for using JoSuA at the IDSC is as follows. The researcher becomes aware of a dataset in a number of ways: by browsing the Center’s metadata offering, because he/she worked with it during a visit at the institute or through the institute’s fellow network etc. If the IDSC is a rightful host and data custodian of the dataset, it will take up the task of assisting with remote access via JoSuA otherwise (and provided it cannot become a custodian of the data) the Center will mediate between the user and a lawful owner/custodian of the data. The IDSC has a dedicated budget available on a first come first serve basis which covers access costs imposed by the data providers. In case of IDSC custodianship the researcher(s) get JoSuA accounts and are enabled to compute against the dataset(s). The code is submitted via a web interface by either file upload, cut-and-paste, or by email to a specific address. The output is censored to comply with any rules imposed by the data producer and is then released to the user.

JoSuA is agnostic to statistical anonymization or non-disclosure procedures, although it allows for attaching automatic censoring based on either a black list or a white list of incoming commands or outgoing results. Actual control remains with the custodian; JoSuA just facilitates the effective enforcement of existing rules and offers an interface for doing so. JoSuA automates all other aspects of running a service, such as keeping a record of user’s projects and jobs, producing business reports, monitoring performance, and managing the logistics of multiple incoming jobs.

JoSuA is designed to prevent data loss. Malicious user action, malfunction of the product, or other such situations may lead to, at most, contained, non-primary data loss: disclosure of one user’s code or output to other users or to a third party may occur (if for example the user forgets a logged in browser on a foreign workstation) but it will never lead to disclosure of primary data. This is due to the backend architecture of JoSuA which pulls submitted jobs inwards and may not be made to push data outwards. Results may be pushed outwards (manually or programmatically depending on configuration) but these are non-critical in the sense of data disclosure since they are censored and hence publishable.

Generally there are two types of approaches regarding computing with data from afar: remote computing and remote processing. The two types of approaches differ in many ways the most important of which are the degree of interactivity and the degree of exposure of the data. Typically interactivity and data protection are inversely proportional. In Germany the display of data on the screen is regarded as a data transfer (to the location of the screen) and is hence not a viable means of computing with sensitive data from afar if the transfer of the data to the location of the researcher is forbidden. Examples of remote computing tools are products based on the VNC or ICA protocol (owned by RealVNC and Citrix respectively). By remote processing one understands a process in which a remote data analyst submits (e.g. per email) analysis code which is run by a local operator who then returns the possibly censored results to the remote user. JoSuA is neither a remote computing nor a remote processing tool although it has elements of both: it allows more interaction...
than remote processing and less than a remote computing tool. It support however several concepts such as that of a user community, community operator, project co-ownerships etc.

It is important to note that the researcher never gets any direct access to the data. The researcher is allowed to compute against the data and gets to see only publishable results conformal to data owner regulations. This makes JoSuA the only tool applicable in cases commercial packages (which are more interactive and allow the display of data on the screen) are not allowed because the display of the data on the screen is considered “data transfer”. JoSuA is suitable even in case where the researcher periodically visits the IDSC or any center running a copy of JoSuA since it allows the researcher to be able to continue working whether locally or from afar.

For researchers who think JoSuA may be helpful for their research and data custodians who would like to look into the possibility of using JoSuA to serve their own data to remote researchers a good way to start is to contact the IDSC at idsc@iza.org or via the IDSC help desk on http://idsc.iza.org/.

5. Future Plans and Challenges
We plan to find an organic, functional way to connect data documentation and remote computing. To that end the work in Askitas, 2009 represents only the beginning and more output is to be expected in this direction. The IDSC is actively working in the areas of data visualization, other forms of data presentation, and new forms of metadata discovery and presentation while actively developing JoSuA further.

More generally the IDSC is involved in a wide array of new and exciting projects and partnerships as its activities are diversifying and its output picking up in both volume and outreach. One of its main challenges will therefore be to integrate a well thought out, commonly accepted framework of legal, ethical, and educational guidelines which will routinely be part of its daily work and operation. This will be done with the assistance of the Institute’s interdisciplinary data committee but also of its partners and communities.

References
N. Askitas, The IDSC of IZA, an international data service center, IASSIST08 Technology of Data: Collection, Communication, Access and Preservation, Conference at Stanford University, Stanford, CA, USA May 27-30 2008.

N. Askitas (2009). METADATATA: Stata module to enable access to metadata, Statistical Software Components S456988, Boston College Department of Economics.


Notes
1 N. Askitas, Head IDSC of IZA, Info: http://www.iza.org/
home/askitas, email: askitas@iza.org. This paper is based on the author’s talk at the IASSIST08 (N. Askitas 2008).

2 Commission to improve the informational infrastructure by co-operation of the scientific community and official statistics: http://www.ratswd.de/

3 The HASSET thesaurus was developed by the UK Data Archive at the University of Essex. Neither the UK Data Archive nor the University of Essex may be held responsible for any errors in this material. We are currently considering other concept hierarchies and taxonomies such as the EUROVOC.

4 JoSuA is designed conservatively in order to avoid making it vulnerable to recurring internet attacks. For an account of the latest scare with DNS cache poisoning see David Schneider, 2008.
A Documentation Model for Comparative Research Based on Harmonization Strategies

Abstract
This paper deals with studies of fixed design, where comparability is feasible either for culture or for time and culture. For the time dimension, studies are divided into cross-sectional and longitudinal, and for the cultural dimension they are divided into monocultural and cross-cultural. Since modern societies are mainly organised into nation-states, cross-cultural studies are carried out more often on country level. However this does not necessarily mean that the organisation of cross-cultural studies in the same country is not possible. Comparative cross-cultural studies follow strategies of data harmonization such as the «ex ante input harmonization», the «ex ante output harmonization», and the «ex post harmonization», as well as mixed strategies. These strategies of data harmonization are complex procedures for which success or failure is reflected in the final study product.

In this paper, a documentation model is proposed for both longitudinal and cross-cultural studies. Three documentation models are proposed according to the different harmonization strategies examined. Finally, the different documentation models are integrated into one.

1. Cross-national studies as a sub-case of cross-cultural studies
The original pattern of comparative research brings together at least two social formations. Some researchers identify cross-national with cross-cultural studies. Thus, for Hantrais (1995), comparative research is a research pattern of the social sciences that aims to conduct comparisons between representations that result from two or more social formations. Globalisation and the revolution in communication technology on one side and the developments in Europe and the course of its unification on the other side, bring in question our current understanding of a «nation-state» as well as the convention to consider comparative research simply as cross-national research.

Globalisation began in the economic sector and next swept across sectors of policy, culture, and knowledge production. The consequence of this evolution was the gradual delimitation of relations, and of the role of the nation-state (Albrow 1998). This change became evident in social science literature as the loss of territoriality followed by the reduction in the sovereignty of nation-states and denationalization (Zurn 1998).

Despite the controversial nature of the «nation-state», most cross-cultural studies are still organised as cross-national ones. However, in order to also cover the case of cross-cultural studies that are not cross-national, we will use the term «cultures» instead of the term «countries» which is commonly used in cross-national studies.

Moreover, as previously stated, cultural discrepancies may exist in the same «nation-state» either on a regional or local level. A basic cultural difference is language. Countries such as Belgium, Finland, and Luxembourg are obligated to carry out any national study in multiple languages. The action plan of the European Science Foundation reports explicitly: «Translations should be made into any language
which is used as a first language by five percent or more of a country’s population» (1999, 10).

This leads us to consider cross-cultural research to have a wider nature than cross-national research and cross-national research as a sub-case of cross-cultural research with a determined culture, namely the nation–state.

2. Formal description of the data element in cross-cultural research

In empirical fixed-design studies (Robson 2007), data production is organised based on a data schema. This data schema is constructed on the basis of statistical ontology, which predicts that the examined population is constructed of similar units of observation, each of which is described by concrete, distinguishable attributes, each of which is represented by a variable. The adoption of statistical ontology as an organisational model of the data schema of fixed-design studies has two basic advantages. The first advantage is that it allows the application of statistics as a method of data analysis. The second advantage is that it allows for the development of a documentation initiative for studies of fixed design. Each population attribute is formally described by a data element that is defined by one concept and by one pattern of value determination (Kallas and Linardis 2009). Each concept is defined by one term and one definition. Each value determination pattern is defined as a) a mono-dimensional classification, b) a number that results from the direct measurement of a concept, or c) text.

The unit of observation, as it is introduced by statistical ontology, is a mathematical schema that does not always correspond to real objects of observation. That is, to social objects that are presented in social practices independently from the observer. The relation of a unit of observation with a real object of observation occurs in the context of each concrete study. In certain cases where the examined social phenomenon consists of the relationships between more than one object of observation (for example in the case of a household in which we usually have at least two objects: the household and the members of the household), the unit of observation does not correspond to real objects of observation but simply represents the total set of all attributes of individual objects (Kallas 2005).

In cross-cultural studies where partial recordings describe different societies (the objects of observation based on which the formal descriptions of social phenomena are constructed) it is possible to differentiate from society to society and consequently from recording to recording. This difference means that either the corresponding objects between two recordings cannot be described by the same data elements, or that certain data elements are not precisely the same.

Consequently, in cross-cultural research, a data element should be described in reference to the object of observation that it describes. The object of observation in each study is related to one unit of observation, which should be documented based on the following:

a) the definition of objects of observation that make up the unit of observation;

b) the social system in the context of which the objects of observation are constructed. Each object of observation is notionally defined in the context of a concrete social system. When it is used in the context of another social system then it is differentiated notionally, and consequently it is described via a different pattern. These differences also concern the data elements whereby the object is defined;

c) the territory where the recording is organized.

The universe is defined by a) a territory, b) a unit of observation and, c) other partial determinations (such as age, sex, marital status, income etc.). The universe refers to both the data elements as well as to the objects of observation. Vice versa, each data element or object of observation refers to a universe. A documentation schema of a data element suitable for the cross-cultural research is shown in figure 1 (see next page).

It is possible to define a territory by a multilingual, controlled vocabulary that includes country and region names (for example the Nomenclature of Territorial Units for Statistics or NUTS classification) as well as other predefined values such as international, European, etc. It is also possible to define the object of observation and social system by a controlled vocabulary. This standardisation further helps in comparative research since the objects of observation are often used by researchers as basic search and comparability criteria.

The documentation schema of a data element includes three basic structural components: the concept, the universe, and the classification. It is possible to reuse the same data element for the documentation of one or more studies. It is also probable for some other data elements to be identified partly by reusing a subset of the structural components that compose a data element. In addition, components such as the universe or category schema can also be reused for the determination of other entities. For example, the category schema can be reused either for the determination of a classification, for the determination of a question, or for the determination of a variable. Additionally, the universe can also be reused for the determination of study «wave instance». However, the universe of each data element usually constitutes a subset of the general universe of a wave instance. In most cases, both universes are similarly identified at the territory and unit of observation level but differ in their other determiners.
The data element, the concept, the universe, and the classification are study components for which even small changes in the content should be documented since they may affect the overall comparability. This is the reason why they are defined as versionable objects. These objects are identified by one code but also by a version number (complex key). Both major and minor changes in the lower level objects simultaneously alter the version of parent objects. The version change documentation should include the following: a) the new version date, b) who made the change, c) why there was a change, so that the users can comprehend if the version change influences the analysis of data (DDI Alliance).

3. Approaches in data harmonization
Harmonized data can be achieved either by using strategies for collecting harmonized data from the beginning (in the study design) or by using harmonization strategies for existing data (Granda, Hadorn and Wolf 2008).

3.1 Strategies for the production of harmonized data in cross-cultural studies
The following is a short description of strategies followed for the production of harmonized data in cross-cultural studies.

• Ex ante input harmonization

Ex ante input harmonization means that the institutions that participate in the study have agreed on common concepts, common measurement patterns of the concepts and also on common questions based on a common source questionnaire. The ex ante input harmonization is used mostly in cross-national studies but is also applied in studies that are conducted in the same country. For example, a question relative to the underground in Greece, such as «How many times do you use the underground per week?» is asked only to Athenians but not to all Greeks. Consequently, an agreement is required for the concepts, the measurement patterns, and the questions even in the same country in order for the questions to have meaning for all participants. In cross-cultural studies where ex ante input is applied, no country-specific variations are allowed except the ones that are absolutely essential such as the language used in the questionnaires (Ehling 2003).

• Ex ante output harmonization

In ex ante output harmonization the institutions that participate in the study have agreed on common concepts and common measurement patterns. The objective is fixed and the choice of suitable questions is left to participating research groups who adapt the questions to the cultural particularities of the universe that they study. Each research group determines its own concepts and measurement patterns; however, they must correspond to the common concept via transformation routines. For example, let us assume that in a cross-national study the participants are asked to indicate their highest level of education. A
common measurement pattern for education level is the use of an international classification such as the International Standard Classification of Education (ISCED). The measurement of education level via ISCED may serve international needs but not national ones, since a country would serve itself more if its national data were detailed enough. Another reason to follow this strategy is that the same data can be collected simultaneously for different studies. More concretely Lene Mejer reports:

«Output harmonization means to give a common internationally agreed definition for a variable and then leave to each single Member State to decide on its implementation. Each Member State decides what is the best national source for the variable (for example from already-existing surveys and/or registers) » (2003, 69).

This strategy is used mostly for cross-national studies; however, it can be used as methodology in cross-cultural studies.

- Mixed Strategy

Some studies, while they follow the strategy of ex ante input harmonization (such as the European Social Survey\(^2\) or ESS and the International Social Survey Programme\(^1\) or ISSP), in certain selected data elements they apply the ex ante output harmonicization strategy. For example the highest education level in ESS is ex ante output harmonized while the study is ex ante input harmonized. In this case, the harmonization strategy should be placed at data element level per wave and not at study level. There is another case of «mixed strategies» where the literal question is agreed upon by the harmonization committee but the category schema is fixed by each country separately.

3.2 Harmonization strategy for existing data: ex post harmonization

Ex post harmonicization is a harmonization strategy where the total study results from already-existing studies. In ex post harmonicization, the institutions that participate in the study agree on common concepts, on common measurement patterns, and on common universes (common data elements). They also agree on already-existing studies that have to be ex post harmonized using transformation routines. The achievement of harmonized data via this process is not guaranteed, even if it has been optimally designed, because of the diversity of concepts and measurement patterns in existing studies. Since no new questions are created, the basic structural elements of these studies are the data elements. New data elements are created that reference already-existing ones. For the implementation of such studies (that resemble research programs more than studies) transformation routines are required, which are written with statistical software. The difficulty of implementing studies following the strategy of ex post harmonization lies in the localisation of common concepts and measurement patterns between the universes. A very useful tool for such studies would be a bank of concepts, classifications, and universes for the localisation of similar data elements.

4. Data archives and the documentation process: a comparative perspective

In recent years, new organizations have been created in Europe called data archives (DA). DA deal with the accumulation, documentation, and dissemination of data. These organizations support secondary analysis and comparative research, and act as mediators between the producers and analysts. The European council is called the Council of European Social Science Data Archives (CESSDA). Each DA must document its own studies based on a common strategy that ensures the following:

- Reuse of common structural study components of a simple study in the same DA

Each study component can be constructed from other structural components. For example, study components such as classification, question, and variable use some common structural components such as codes and categories. Often the category schema of classification, question, and variable coincide. For example, the codes and categories that are used for ISCED classification (for the corresponding question but also for the corresponding variable in a statistical data file) may coincide. In this case the study components’ common structural components should be imported just once and then reused (DDI Alliance) even in the case of a simple study that is neither longitudinal, nor cross-cultural. Each person who documents a study should follow these rules so that double entries are avoided. To aid in this laborious documentation work, certain processes for localisation of common structural components can be automated.

- Comparability of a longitudinal study in the same DA.

- The comparability of a longitudinal study in the same DA lies in the reuse of study components between waves. Components such as concepts, classifications, universes, questions, and variables are principal components for comparability between waves of longitudinal studies and they should be reused in the various waves. Consequently, each DA should maintain local banks of all these study components.

- Comparability of different studies in the same DA

Comparability of different studies is also based on the reusability of the same principal study components.
between the various studies, as in a longitudinal study.

- Comparability of a cross-cultural study in the same DA

The proper documentation of a cross-cultural study involves all the participating organizations and it differs depending on the harmonization strategy that has been followed. The documentation of a cross-cultural study based on the harmonization strategy followed is developed analytically in section five. While the documentation of a cross-cultural study often occurs in different DA’s, in this work we will deal with the documentation of a cross-cultural study in the same DA (or data-metadata repository).

- Study comparability in different DA’s

Study comparability in different DA’s is a very critical process for wider comparative research but this will be analyzed in a later work.

Summarizing the above, it is immediately evident that the documentation procedure is a difficult and laborious process. On the other hand, the result of this procedure will be useful for the wider research community, particularly for those researchers who want to carry out comparative research and secondary analysis. The comparative documentation further strengthens the role of DA’s. The documentation process is best carried out in collaboration with the primary data producers as well as with the statistical institutes.

5. The documentation of a cross-cultural, longitudinal study

The documentation process is rendered particularly difficult and laborious in the case of cross-cultural, longitudinal research. The collaborating institutions should follow the documentation of the coordinating institution, since the resulting documentation will be based on common agreed concepts, measurement patterns, questions, and universes. The documentation completed by the coordinating institution should not be changed by the participating institutions. The documentation language of the coordinator is the common agreed language (usually English).

In the cases that follow, the model is presented first and then a description of how the model should be used by participants based on the agreed-upon harmonization strategy. It takes into consideration the most complex study type, the cross-cultural, longitudinal study, since all other studies can be documented based on this. The diachronism relies on the reuse of study components for each wave. The multiculturalism lies in the creation of references between the source and universe study components.

It should be noted that the models that follow concern the most complex study type – the cross-cultural, longitudinal study – but they document just one study. Another limitation is that the documentation takes place in the same DA and not in distributed documentation systems.

Below is a short description of the main entities used in the models:

- Study: the entity that is used to store the general study information such as title, more general objectives, summary etc. The main purpose of adopting such an entity, beyond the storage of general information, is that it aims to unify all study waves. The study level documentation is completed by the coordinating institution in the common agreed language. Translation into other languages occurs only for dissemination reasons. This entity is not reusable but can be referenced by other studies or by other study components.

- Wave: a longitudinal, cross-cultural study takes place in many time and universe instances. The time instance of a study is called a wave while the universe instance of a study wave is called a wave instance. While documenting, but also while a study is conducted, it is common practice to establish the time and, for each time period, to receive snapshots for the various universes that participate in the study. In addition, at this level, the general wave title, any special objectives per-wave, and the total duration of the study wave are all recorded. Information such as the universes or institutions that participate in the study may be recovered automatically from the wave instance documentation level so that no differences in the aggregated fields of wave level exist. The most crucial documentation at wave level has to do with the determination of the study harmonization strategy. It is also crucial this be selected from a controlled vocabulary where the user chooses between the following options: a) ex ante input harmonization, b) ex ante output harmonization, c) ex post harmonization, or d) mixed strategies. The harmonization strategy is determined at wave level, not at study level, because it is possible (although rare) that the harmonization strategy may change from wave to wave. The wave entity is also used for the grouping of source data elements, of common source questionnaires (when they exist), and of the harmonized statistical data files. The documentation at wave level should be completed by the coordinating institution in the common agreed language. Translation into other languages is done only for dissemination reasons. The wave entity cannot be reused but can be referenced by other study components.
• Wave instance: the entity that is used for storing information concerning wave snapshots per universe. The documentation of wave instance level is completed by all the research groups that participate in the study, in the languages that have been decided per group but also in the common agreed language for dissemination purposes. The wave instance includes extensive information such as universe, sampling methods, participating institutions by role (local coordinator, financiers, data producers, organizations responsible for data dissemination), researchers, sampling frame, data collection method, time of data collection, and description of weights (accompanied by weighting methodology). The wave instance is not a reusable object but can be referenced by other study components.

• Source data element and universe-specific data element: entities used to store information concerning the data elements that were introduced in section two. The implementation of these two entities in a database does not necessarily require the creation of two tables for the two types of data elements; however, both data elements are presented as separate entities in the entity relationship diagrams in order that the required relationships are evident. The same holds for the questionnaires, the questions, the data files, and the variables. The data element and its structural components are reusable entities for different studies or study waves.

• Source questionnaire and universe-specific questionnaire: entities used to store general information concerning the questionnaire such as the number of questions, type of questionnaire (standardized versus non-standardized), abstract, and link to the questionnaire file. This is also a grouping entity for questions. Questionnaires are not reusable entities but have to be defined again in each wave or wave instance.

• Source question and universe-specific question: according to Kallas and Linardis (2009), the questions are composed of some or all structural elements presented in figure two. The question is also a reusable object.

• Harmonized data file and universe-specific data file: entities used to store general information about the statistical data files such as the number of variables, the number of cases, and likely a link to the statistical data file. Data files are also used as grouping entities for the variables. The data files are not reusable entities and have to be defined again for each wave or wave instance.

• Harmonized variable and universe-specific variable: the variables consist of structural elements such as name, description, type, measurement level, and category schema. The variable entity, as it is described here, consists only of metadata and not of data. The same variable may have a number of data depictions but in different statistical data files. The variable is a reusable entity.

• Finally, the transformation routine is the process that describes the necessary transformations of the universe-specific data element to source data element.

5.1. Case 1: study documentation following ex ante input harmonization strategy
The documentation process for ex ante input harmonization is portrayed in figure 3 (on next page). The left parallelogram portrays the documentation that should be completed by the coordinating institution while the right one portrays the documentation that should be completed by the participating institutions.

Figure 2: Documentation schema of a question
The documentation may be published in intermediary stages (indicated below). The basic principle of ex ante input harmonization is the standardization of data elements and questions between the research groups. The stages of documentation are as follows:

Stage 1: Documentation of the general context of the study and wave, and documentation of data elements (documentation provided by coordinating institution).

1. Documentation of the general context of the study: Documentation is completed by the coordinator at the beginning of the study.

2. Documentation of the general context of the study wave: Study waves have to reference the corresponding study.

3. Documentation of common data elements (common concepts, measurement patterns, and universes): Data elements should reference the study wave and not the study because the data elements may differ from wave to wave. For example in the ESS, there are some data elements that are used in all study waves while others are added or removed periodically.

Fundamental practices to ensure comparability: dissemination of stage one documentation to the participating institutions.

Stage 2: Translation of the context of study and study wave; translation of the data elements; determination of each wave instance (documentation provided by participating institutions).

1. After the dissemination of stage one documentation to the participating institutions, each participating institution should translate the general context of the study and of the study wave, and the data elements, for dissemination reasons.

2. Each participating institution should then document the wave instance based on the universe it represents. Each wave instance has to reference the corresponding wave.

(First intermediate phase of study publication)

Stage 3: Documentation of source questionnaire/s (documentation provided by coordinating institution).

1. Documentation of source questionnaire/s: Each questionnaire should reference the corresponding wave.

2. Documentation of source questions: The reference between source questions and the corresponding data elements as well as between source questions and the source questionnaire is required.

Fundamental practices to ensure comparability: a) dissemination of stage three documentation to the participating institutions, b) creation of a statistical file template with common variable names (based on the data elements) for all participating institutions, c) dissemination of the template to the participating institutions.

Stage 4: Translation of source questionnaire/s, leading to universe-specific questionnaire/s (documentation provided by participating institutions).

1. Documentation of universe-specific questionnaire/s: The reference between the universe-specific questionnaire and the corresponding wave instance is required.

2. Documentation of universe-specific questions: The questions specific to each universe are created via translation of the source questions in language or languages decided by each research group. The reference between universe-specific questions and source questions as well as with the corresponding universe-specific questionnaire is required.

Fundamental practices to ensure comparability: a) each institution conducts the research, collects the data, and submits the statistical data files to the coordinator, according to the template already sent by the coordinator, b) at the same time, each institution preserves the data files for the stage six documentation procedure.

(Second intermediate phase of study publication)

Stage 5: Documentation of harmonized statistical data file/s (documentation provided by coordinating institution).

1. Documentation of harmonized statistical data files that have come from the merging of universe-specific data files: The reference between harmonized statistical data file/s and wave is required.

2. Documentation of harmonized variables: The reference between harmonized variables, the harmonized statistical file, and the corresponding source questions is required.

Fundamental practices to ensure comparability: dissemination of stage five documentation to the participating institutions.
Stage 6: Documentation of universe-specific data files (documentation provided by participating institutions).

1. Documentation of universe-specific statistical data files: The reference between universe specific statistical data files and the corresponding wave instances is required.

2. Documentation of universe-specific variables: The reference between universe-specific variables and the statistical data file they belong to, universe-specific variables and the corresponding universe-specific questions, as well as reference between universe-specific and harmonized variables is required.

(Final phase of study publication)

5.2. Case 2: study documentation following ex ante output harmonization strategy.
A basic difference between ex ante input harmonization strategy and ex ante output harmonization strategy is that the second presupposes the determination of universe-specific data elements by the participating institutions. Consequently, the participating institutions should document the universe-specific data elements and reference them to the common agreed data elements via transformation routines. Also, there is no source question, just a source data element. On the other hand, there are universe-specific questions but these are considered mostly as additional documentation of universe-specific data elements not as fundamental structural study components. The documentation process for ex ante output harmonization strategy is portrayed in figure 4. For simplistic reasons we have not drawn the structural elements of the data element again (concept, measurement pattern, and universe). The left parallelogram portrays the documentation that should be completed by the coordinating institution while the right one portrays the documentation that should be completed by the participating institutions.

Following this strategy, there are five study documentation stages instead of six. This occurs because stage three of ex ante input harmonization does not make sense here since there are no source questionnaires or source questions. The

Figure 4: Study documentation model following ex ante output harmonization strategy
differences between the documentation procedures of the two strategies are summarised as follows:

- Stage 2. This stage includes two additional documentation actions to be performed by the participating institutions: 2.3) documentation of universe-specific data elements; and 2.4) documentation of transformation routines of source to universe-specific data elements.

- Stage 3. As was mentioned before, stage three does not exist. Nevertheless, the coordinating institution can establish some fundamental practices to ensure comparability such as: a) the creation of a statistical file template with common variable names for all participating institutions, based on the data elements; and b) the dissemination of the template to the participating institutions.

- Stage 4. Phase 4.2 is different because there are no source questions. Consequently, the universe-specific questions are developed from scratch instead of being produced as translations of the source questions. Reference between universe-specific questions and questionnaire as well as between universe-specific questions and data elements is required.

- Stage 5. Phase 5.2 is different because there is no reference between harmonized variables and source questions. Reference between harmonized variables and harmonized statistical file is required.

5.3. Case 3: study documentation following ex post harmonization strategy

The basic difference between this strategy and ex ante output harmonization lies in its relationship to the common concept. In ex ante output harmonization, its relationship to the common concept is guaranteed because the researchers design the wave instances from scratch, keeping in mind the common data elements. In ex post harmonization, the individual studies have been designed autonomously by the researchers without adhering to a common concept. Thus, the relationship to the common concept is not guaranteed.

On the other hand, the two harmonization strategies have a lot of similarities related to methodological issues. The documentation process is similar to the one that was described based on figure 4. The basic difference is that in ex post harmonization, the new study derives from already-existing studies. Consequently, it should initially be documented using the already-existing study waves or wave instances from which the new study derives. It would provide great relief from the excessive documenting load for researchers working on an ex post harmonized study if documentation of existing study waves or wave instances was already available.

Let us assume that a new study is designed following the ex post harmonization strategy. The new study concerns attitudes for a set of countries (two of which are: Cyprus and Russia), for the time period 2004-2005. The coordinating committee decides to harmonize ex post the second wave of the ESS. According to Jowell et al. (2007), Cyprus and Russia did not participate in the second wave of the ESS. Nevertheless, the coordinating committee is aware of the existence of other national attitude studies for Cyprus and Russia during 2004-2005 and decides to harmonize them ex post. At the same time, the coordinating committee has to decide on the common data elements of the new study. After the common data elements have been defined, the data elements of the existing surveys have to be transformed via routines to the common ones. It is common for different groups to undertake the transformations of different studies. In our example, three groups will undertake the burden of transformations of source data elements to the common agreed data elements: one group for the ESS, one for the Cyprian study, and a group for the Russian one. The documentation process of the three groups includes: a) the documentation of each new wave instance; b) the reference of already-documented data elements (ideally) to the wave instance; or c) the documentation from scratch of all previously conducted studies in the research program, if their documentation does not exist; and d) the application of transformation routines to universe-specific data elements and source data elements.

The documentation process in ex post harmonization is portrayed in figure 5. The left parallelogram portrays the documentation that should be completed by the coordinating institution of the research project, while the right one portrays the documentation that should be completed by other institutions that participate in the research project. These institutions will have undertaken the documentation of concrete wave instances from already-existing studies.

Figure 5 differs from figure 4 in the following ways:

- Each new study wave can be designed based on existing study waves and/or existing wave instances. Consequently, suitable documentation at wave level is required.

- The relationship between source data element and universe-specific data element is a “many to many” relationship, since the same universe-specific data element may correspond to more than one source data element in the same system. For example, a universe-specific data element may correspond both to the source data element of the
initial study and to a new source data element that was created during the design of a study following ex post harmonization.

5.4. Common documentation model for all harmonization strategies
As already mentioned, the harmonization strategy followed should be determined at wave level. In the case where the strategy followed is ex ante input harmonization, the documentation model is the one in figure 3. The documentation models of ex ante output harmonization and ex post harmonization are depicted in figure 4 and figure 5, respectively. In the case where the harmonization method is determined as “mixed strategies”, the harmonization strategy has to be determined for each data element per-wave. The entity relationship diagrams 3, 4 and 5 are unified in figure 6. In using this model, the documentation process will differ depending on the choice of harmonization strategy.

The documentation model in figure 6 also serves the needs for the documentation of different studies based on a comparative perspective. This is feasible because entities such as concept, classification, universe, data element, question, and variable, that constitute the basic study components for comparative research, are reusable entities for all studies. The reusability of these study components, in a documentation system of a specialized architecture, aims at comparative documentation between different studies. Another useful outcome of such a documentation
model is that it is feasible for a researcher to locate universe-specific study components derived from source study components.

When they are referenced, the study components referred to above can never be deleted or changed. These components are identified by Persistent Identifiers (PIDs). If one of these components has to change then a new version of this component has to be created.

Multilingualism of study components is applied in two cases: a) when a component is translated by an institution in order for its translation to be an “active component” of the study (for example translation of the source questions to universe-specific questions); and b) just for dissemination reasons (for example translation of a study’s abstract). The reasons for translation of a study component should be declared in the documentation. In the first case, both major and minor changes may lead to version change of the study component, not so in the second one.

6. Conclusions
The general documentation process in a DA or in a metadata and data repository is based mostly on ex post harmonization procedures. The institutions that document a longitudinal, cross-cultural study should do so based on already-existing documentation. In the case of longitudinal studies, this should be done by repeating study components from other waves, and in the case of cross-cultural studies, by referencing source and universe-specific objects.

Figure 6: Study documentation model used for every harmonization strategy
Consequently, the documentation procedure not only involves the typical description of the components derived in the context of a single research project, but also all of the a posteriori references (additional documentation) between independently-designed studies.

The proposed study documentation procedure is laborious for the researchers that are making the documentation. Additionally, s/he should know the specifics of each study, which presupposes a close collaboration with the primary investigators. Moreover, the “golden super rule” of METANET research project states: «Metadata are as important as data, and metadata need as much work as data» (Sundgren, 2003, 129). The study documentation with suitable metadata can provide particular advantages in the identification of «equivalent» or «equal» study components either for longitudinal or cross-cultural studies, or even studies with similar subjects.

Documenting studies based on the proposed documentation model, allows the researcher to:

- search and locate questions that use either common concepts, common classifications, questions that are addressed in common universes, or even questions that use common data elements;
- search all the variables that are derived from the same questions;
- locate data element transformation routines, so that the sequence from one data element to the other is clear;
- locate all the translations of a source question, likely accompanied by qualitative criteria such as validity and reliability but also non-response rates (Sarris et al. 2007);
- locate data according to the following criteria: a) the concepts that the data imply, b) the universes the data refer to, c) the time period the data refer to or the time the survey was conducted, d) concrete classifications based on which the data have been produced.

This model can also be extended and modified so that it may cover the documentation of comparative research in distributed environments. In this case, not only the model would be of particular interest, but also the flexible and functional architecture of the overall documentation system. The extension of the model in distributed environments as well as the architecture of such a system will be analyzed in a future paper.

References


Sarris, Willem E. and Irmtraud N. Gallhofer. 2007. Can questions travel successfully? In Measuring attitudes cross-nationally: Lessons from the European Social...


Notes
1 Kallas, Ioannis ,University of the Aegean, Department of Sociology , Academic Field: Methods & Information Techniques of the Social Sciences,Address: Tertseti & Mikras Asias str., 81 100 Mytilene, Lesvos, Greece./ Tel. (+30) 22510 36559. email: i.kallas@soc.aegean. Linardis, Apostolos , National Centre for Social Research, Address: 14-18 Messoghion Av., GR-115 27, P.O.B 142 32, Athens, Greece.Tel. (+30) 210 7491656.email: alinardis@ekke.gr

2 http://www.europeansocialsurvey.org/

3 http://www.issp.org/

4 The documentation of all universe-specific entities is completed by all participating research groups, in the languages that have been decided by each group but also in the common agreed language.
Strengthening Data Security: an Holistic Approach

Abstract
In the light of heightened concern around data security, this paper highlights some of the measures that can be used to develop and strengthen security in data archiving. The paper includes discussion of the different approaches that can be taken towards the construction of firm and resilient data and information security policies within the social science data archiving communities. While international standards can provide theoretical guidelines for the construction of such a policy, procedures need to be informed by more practical considerations. Attention is drawn to the necessity of following a holistic approach to data security, which includes the education of data creators in the reduction of disclosure risk, the integration of robust and appropriate data processing, handling and management procedures, the value of emerging technological solutions, the training of data users in data security, and the importance of management control, as well as the need to be informed by emerging government security and digital preservation standards.

Keywords
Data security; data archiving; information security; data handling; user training.

New legislation and the data security ‘climate’
During 2007, a series of high-profile data losses by UK government and associated organisations took place, involving reputedly 37 million items of personal data. In a climate of increasing concern over data security and identity theft, legislation in the form of the Statistics and Registration Services Act 2007 (SRSA) was at that time already making its way through the UK Parliament, its provisions intended to become effective from 1 April 2008. As a result of media furore over the data losses, in November 2007, the Cabinet Office was instructed to review data handling within government departments and make recommendations for their improvement where needed. This review was to be conducted under the direction of Robert Hannigan, Head of Security, Intelligence and Resilience. An interim report was produced in December 2007, followed by the final report in November 2008. Together, the SRSA and Cabinet reports had a marked effect on how data are now handled across UK government departments and culminated in the ‘Mandatory Minimum Measures’ for data handling. The SRSA also established the UK Statistics Authority, and for the first time introduced criminal penalties for the unlawful disclosure of confidential information (i.e. data relating to, or identifying, a particular person or business held or disclosed by the Statistics Authority).

The UK Data Archive (UKDA) is the curator of a large collection of UK government social science research data, held for use by the academic community. It is also licensed as a legal Place of Deposit by The National Archives, allowing the UKDA to ingest and preserve public records. Therefore, while not strictly covered by the Statistics and Registration Services Act, by the nature of its business the UKDA is intimately concerned with data integrity and security. While the UKDA has developed and maintained robust security practices over the years since its inception, it was felt that the time was right to review practices in response to the challenge of new legislation.

During the course of its work, the UKDA acquires data and associated materials from data creators, conducts ingest processing to prepare those data for secondary use, and supports users once they have received the data. Therefore, a similar holistic approach to the audit and refinement of data security was taken, to ensure that all the UKDA’s data acquisition, ingest, access and support activities are supported by coherent and consistent procedures that work at all stages of the data archiving life-cycle.

Data creators and security enhancement
Firstly, the initial part of the process was reviewed – the work that the UKDA undertakes with data creators – and an assessment was made of how the new UK legislation on data security may affect practice. The UKDA will advise data creators at all stages of their project: from the planning stage, throughout the data gathering process and after completion and deposit of the data. This helps to ensure respondent confidentiality while maintaining sufficient detail within the data to enable effective research. The work is wide in scope, as data are acquired from a range of sources including large, well-funded government organisations, established research centres, and independent...
small-scale academic research projects.

The first tangible effect of the new UK legislation that the UKDA experienced was a marked tightening of physical data transfer process from government, including the increased use of file encryption, more secure methods of data delivery, use of courier services, etc. This was unsurprising given the recommendations of the Cabinet Office report. The UKDA accordingly streamlined and secured acquisition and internal data transfer procedures to accommodate these developments.

Beyond the security of data transfer, new developments in the nature of data released from the Office for National Statistics (ONS) also quickly became apparent. Since before 2006, the ONS Microdata Release Panel (MRP) have overseen the release of ONS datasets by providing advice and testing for statistical disclosure, and as a result of the SRSA, the Panel updated their policy for the release of data at certain access-controlled levels. The UKDA were primarily concerned with two ‘levels’ covered by this data access control strategy, which have been developed by negotiation between the ONS and the UKDA: the End User Licence and the Special Licence. Datasets are compiled by ONS to these respective levels of detail prior to deposit with the UKDA.

To be able to download and use data from the UKDA, each user must register an account with the Economic and Social Data Service (ESDS - for which the UKDA is a service provider), via the UKDA website, and agree to the terms of the End User Licence. This includes an undertaking that the user must preserve at all times the confidentiality of information pertaining to, and not to attempt to identify, individuals and/or households in the data collections, nor must they share data with others who are not registered users. In accordance with the terms of how the SRSA defines levels of information, ONS data made available for research access at End User Licence level must not reveal or have the potential to reveal the identity of an individual. However, it is also recognised that researchers may sometimes need access to more finely-detailed data to conduct their research, and for this reason the Special Licence was set up with ONS. To gain access to data held under a Special Licence (which may include data designated as ‘personal information’ under the SRSA, but must have had some degree of protection applied), the prospective user must apply via the UKDA to ONS for Approved Researcher status. The candidate must provide extensive details of their academic background, status, and prospective research, and prove that they are a ‘fit and proper’ person to receive Special Licence data.

There are significant differences between the End User Licence and Special Licence versions of a dataset. As a practical example, the lowest geographic level permitted for End User Licence data is generally Government Office Region (GOR, or Nomenclature of Territorial Units for Region (NUTS) level 1) and education or employment and other demographic data may be banded or aggregated. Special Licence data, by contrast, may include geographic data at a finer resolution, such as Unitary Authority and NUTS2 or NUTS3 geographies, and more detailed education, employment and demographic data. Of course, the advent of the Special Licence does bring an added administrative burden; separate holdings of Special Licence and End User Licence versions of a dataset require doubled ingest processing work to be undertaken at the UKDA, and the administration of Approved Researcher applications is resource-intensive for both the UKDA and ONS. However, these steps must be taken to ensure that in a climate of heightened concern around security, data of a sufficient level of detail remain potentially available to the research community while respondent identity is still protected.

Further to this (whilst it lies outside the remit of this paper) the UKDA will launch the Secure Data Service (SDS) in late 2009, where more potentially disclosive data will be available to selected Approved Researchers, who will then undergo further levels of security and system training.

However, not all government data acquired by the UKDA originate from ONS. While government organisations primarily concerned with data have robust procedures in place, other departments, especially those who have only recently begun to release data to researchers, are still coming to grips with the provisions of the SRSA, and do sometimes need guidance. As a result of consultations with ONS over End User Licence and Special Licence data, the UKDA are uniquely placed to offer such guidance, should potential confidentiality issues arise, and often do so. As well as offering practical advice on how to balance disclosure risk while maintaining useful detail within data, by means of data edits, access control or a mixture of both, the UKDA (alongside ESDS Government) are currently involved in the process of facilitating discussions between the ONS MRP and other government departments, so that all may benefit from the MRP’s expertise. Additional negotiation and guidance is likely to occur as a result of the Secure Data Service with new licence agreements between the UKDA and the ONS and between the UKDA and researchers.

In the UKDA’s experience, academic researchers also vary widely in data security expertise. Some are attached to large research centres with established data managers and procedures, and others may undertake small-scale team or solo projects. While the data UKDA receive from government are largely quantitative, the data generated by academic projects may be quantitative, qualitative or a mixture of both. A considerable amount of work is undertaken by the UKDA to educate and inform the academic community on all aspects of data management, including obtaining consent; maintaining confidentiality and security; holding well-publicised regular workshops.
covering both quantitative and qualitative data are held around the UK; and providing advice to individuals and organisations at all stages of the research process. The data deposited at the UKDA as a result of unique academic projects may present unusual challenges, and any potential edits or levels of access control that may be needed to reduce the risk of disclosure are discussed with researchers prior to the commencement of full ingest processing. The recent Managing and Sharing Data guide gives straightforward and plain English advice on security issues that affect researchers.

**Internal data handling**
The second part of the holistic approach the UKDA took to strengthen data security was to audit all aspects of ‘in-house’ data handling. Members of staff at the UKDA have a wide range of expertise and experience, and work on diverse data tasks. Therefore, it was essential as part of the audit process to examine and scrutinise internal procedures, both human and technological, for the handling and storage of dataset files and associated administrative materials. As a result of this, existing good practice was identified and additional methods developed. These were collated into a comprehensive set of data security procedures, which were then distributed to all UKDA staff. In addition to the requirement that all staff register as UKDA users and are thus bound by the same conditions of the End User Licence, a Confidentiality Agreement has been also introduced. This details the responsibilities of staff with regard to data confidentiality, and requires the signature of the individual staff member. The smooth introduction of such stringent security measures needs to be carefully handled, and staff professionalism, awareness and expertise must be respected and acknowledged. Explanations as to why security measures were to be strengthened were given to staff, and all background information on the SRSA and the Cabinet reports have been made available, including legal aspects regarding differential treatment of data at the End User Licence and Special Licence levels. Staff members were encouraged to ask questions and actively feed into discussions throughout the process, and where needed, training was provided. Feedback from staff at all levels has resulted in some excellent suggestions that have since been incorporated into the security procedures. At the same time, a UKDA Security Plan has also been developed in conjunction with the planning of the Secure Data Service. This plan brings together all aspects of information security within a single document and is based on the ISO 27001 standard. While the Plan has been developed explicitly for the Secure Data Service, it has ramifications which extend across the whole of the UKDA's procedures. It is within physical and environmental security and operations management that the key changes are likely to impact, but there are also technical changes that will be necessary and will influence the way in which staff carry out their daily work, even if their work is entirely unrelated to the Secure Data Service. The Plan will also be revised to take account of the more recent Her Majesty’s Government (HMG) Security Policy Framework and the provisions made in that Framework will have to be applied to the UKDA's interactions with government departments as well as researchers.

While this paper does not cover technological solutions in detail, it must be emphasised that the UKDA servers and system architecture and associated infrastructure are maintained to broadly conform to ISO 27001/2, and are updated periodically in step with technological advances. Regular systems testing is undertaken to identify any vulnerabilities, and other practical measures include the inclusion of integrated checksums in the names of downloaded files to prevent SQL injection, and processes put in place to prevent cross-site scripting and OS Command Injections. These ‘development’ and technical solutions are are under continuous development, and there will also be additional changes before the commencement of the Secure Data Service. The Security Plan also dovetails with other UKDA Policies and Procedures including the established UKDA Preservation Policy, which covers data authenticity/integrity arrangements.

All plans, policies and procedures need regular review and updating. All the external factors that influence internal UKDA procedures and policies will continue to develop, and the UKDA must keep pace with these developments to both maintain internal standards and promote external confidence in the work of the UKDA. This ongoing process of review and revision will remain collaborative; all staff will continue to be encouraged to reflect on issues relating to data security that affect their roles and to use that knowledge to improve working practice. The successful introduction and implementation of new and sometimes tiresome procedures related to security needs should have considerable staff support, and it is only by allowing staff to contribute to the process that this support can be effectively harnessed.

**Data security: The user’s perspective**
The UKDA’s work with data users comprises the third major area for data security review and development. As there are currently over 50,000 registered UKDA/ESDS users, providing data security advice and support to users is a considerable task. The UKDA alongside its ESDS partners, acts to represent the interests of data users and accordingly facilitates dialogue with data creators to ensure users are given access to sufficiently-detailed data to permit useful research to take place. However, this brings an associated responsibility to promote safe data practices and train users in effective data security. The UKDA have developed measures to facilitate this, including regular workshops and training, and a web-based guide to good practice on micro data handling and security, which has been available for some years and is regularly updated to take account of new developments. These measures will be
extended considerably for the Secure Data Service. It is the UKDA's opinion that mandatory training courses for users of this service will provide one of the main planks in the strategy of allowing researchers desk-top access to secure data. The combination of detailed training, implementation of secure technologies, strong penalties and conformance with relevant standards will provide the necessary checks and balances to provide the most viable user experience.

Training in best practices in data management leads to training for researchers about the risks of disclosure. For the Secure Data Service it will be imperative that researchers are aware of what constitutes disclosure, how to identify disclosure risks in their own analyses, and what the legal and practical consequences of disclosure are. Training users will reduce the risk of security breaches. As part of the audit of data security, further measures are currently under consideration to improve the range of tools to educate data users on robust data security practice.

However, an effective policy has to be in place to deal with any sanctions and breaches of data confidentiality by users, who are made aware of potential measures that can be taken against them or their host organisation in event of a breach. Again, the planning for the Secure Data Service has identified new risks which will be covered by a new multi-tier procedure.

**Conclusion**

To summarise, it must again be emphasized that the work undertaken by the UKDA to strengthen data security has of necessity taken a holistic approach. The three fronts on which data security work is of prime concern (data creators, internal practices, and data users) are all interdependent. Work with data creators, including government departments at the beginning of the process, inculcates strong internal practices and data security awareness at the UKDA, which in turn leads to the better safeguarding of data and excellent educational work with data users. This paper has shown how data security at the UKDA has been strengthened in response to a changing external environment, and how the work will continue to develop as that landscape changes further.

**Notes**

1. Dr. Sharon Bolton, Data Services Manager, UK Data Archive: email sharonb@essex.ac.uk. Dr. Matthew Woollard, Associate Director, Head of Digital Preservation and Systems, UK Data Archive: email matthew@essex.ac.uk


8 The UKDA Security Plan is currently an internal document only.


Swedish National Data Service’s Strategy for Sharing and Mediating Data

Practices of Open Access to and Reuse of Research Data

Abstract
This paper begins with a description of the current key actors in Sweden, which are promoting research infrastructure and accessibility to research data, put into context. The Swedish National Data Service’s (SND) organization, mission, and strategy to promote data sharing is also described. SND’s strategy is a combination of top-down and bottom-up activities. An example of a top-down activity is to influence research funders to put higher demands on future open access data when studies are completed or to support researchers through the whole research process by providing guidelines on ethical and legal issues. Examples of bottom-up activities are to be present in different research contexts and to inform about the benefits of sharing data. One example of this is a joint project with SND and four university libraries. SND has conducted a national inventory survey, initiated in the fall of 2008, of existing databases and database research, as well as attitudes towards data sharing among researchers and university managements within social sciences and humanities departments at Swedish universities and university colleges.

In addition to the inventory process, two survey studies have been carried out in spring 2009, one targeting professors and the other doctoral students in the same domains of disciplines at Swedish universities and universities colleges. The questionnaire contained 80 items covering the researchers’ affiliations; domain of discipline; gender; age; familiarities with research policies and ventures; and use, reuse, and archiving practices of digital research data. Furthermore, there were questions about possible reasons for not using digital data, interventions and barriers to enhanced reuse and accessibility to data, possible agents in overcoming barriers, and willingness to share their digital research data. The surveys were carried out through email questionnaires sent to professors (N=549) and doctoral students (N=1147). The results from the surveys show that doctoral students in general expressed great uncertainty about questions of amounts of reusable digital data and effective interventions to enhance accessibility to digital research data. They identified research ethical aspects as important barriers to sharing digital research data, while professors emphasize lack of resources for researchers to document and make their data accessible for others as the most important obstacle. Concerning interventions to enhancing reuse of digital data, the majority of the doctoral students and the professors thought it should be effective to get more information about accessible research data in data archives or databases. Nearly 100% in both groups reported that more training in research methods, digital research databases, and information about accessible e-tools would be effective interventions. The most effective interventions for enhancing accessibility to digital data were that research grants should include funds for preparing the data for sharing and archiving and that archiving data for use by the scientific community is acknowledged to be of scientific merit. Surprisingly, when it comes to the degree of urgency in sharing their own data, the professors seem to be a bit more eager to share data than the doctoral students. The results are compared with the results from the parallel study of the professors and from a recent survey targeted at professors in various social sciences and humanities disciplines at Finnish universities (Kuula and Borg, 2008).

1. Introduction
1.1 Building a Swedish research infrastructure
The Swedish Research Council (VR) has, since its start in 2001, been focusing on the need to build a research infrastructure. As a part of this work the Committee for Research Infrastructure (KFI) was set up in 2004. The main purpose of KFI is to formulate long-term strategies and handle resource allocation for expensive scientific equipment, large research facilities, and extensive databases. The committee also deals with Swedish interests in, and funding of, various national and international research infrastructures. The overall aim is to provide better conditions for Swedish researchers by ensuring access to high-quality infrastructures.

The committee is also the producer of the Swedish roadmaps for research facilities to meet future scientific demands. The first Swedish Research Council’s Guide to Infrastructure was published in 2006 and the second by the end of 2007 (The Swedish Research Council, 2007).

As part of the Swedish Research Council’s major

Carina Carlhed and Iris Alfredsson
not an easy task. Only a small proportion of data produced by Swedish social science researchers were deposited at SSD. SND’s conditions are, however, better than SSD’s: increased economic resources and an organization placed within a bigger network of infrastructural resources. However, an important factor for the result is the general attitudes towards data sharing among researchers. Is there simply no culture of data sharing and reuse of data among researchers in Sweden? Or does sharing and reuse exist, but not via a data service?

We have identified two major barriers for reaching our goals: legal barriers and possessive barriers. The legal barriers are obstacles in Swedish current laws and regulations. The possessive barriers are thresholds connected to unconscious attitudes of researchers.

1.3 Legal issues

Issues surrounding shared data infrastructures have important legal implications. For this reason, DISC has surveyed the legal regulations that apply. The survey includes an inventory of relevant regulations in the areas of integrity protection, copyright, and archiving (DISC, forthcoming). The report will provide a basis for determining the actions needed to facilitate the creation of a common data infrastructure. A working hypothesis at DISC is that the issues involved are so fundamental that they require a public investigation.

An example of legal obstacles pointed out by DISC is the regulation concerning the use of the Swedish person-identified population registers on health and social conditions. This very important source for Swedish research gives unique opportunities to create research databases for longitudinal research in medicine and social sciences. However, the current ban on creating a common research infrastructure with personalized data limits the use of these resources. The Personal Data Act, the Secrecy Act, and other regulations allow the use of research material only for specific projects. Universities may not collect and store data intended to serve a wide number of researchers in the same scientific area.

DISC also calls attention to the fact that the central Swedish administrative agencies, mainly Statistics Sweden, are not given the basic instruction to provide the research community with data from registers. Instead they sell research data to individual research projects as the need arises. This results in the fact that research funders over and over allocate funds to purchase the same research data. It also takes responsibility for a broader area. According to Section 3 of the agreement, SND “shall meet the needs of the research community for data on empirical research in the areas of social science, humanities, and medicine. Actions include providing technical, legal, educational, and other administrative resources for collecting, storing, and distributing data for research.”

SND is governed by a steering board and by a national reference group. The board of SND consists of five national representatives for the above sciences, appointed by the Swedish Research Council, the national reference group, and the University of Gothenburg. Formally the new organization started 1 January 2008. However, SSD performed the operational tasks during the first six months of 2008. On 1 July 2008 most of the SSD staff was transferred to the new organization. At the same time SSD was closed down, and the SSD data collection and equipment were taken over by SND.

1.2 The Swedish National Data Service (SND)

According to the guiding principles of SND, the main purposes of the data service are to mediate information on databases and other digital material collections for research, to facilitate access to research databases, and to serve as a knowledge node for documenting and managing research data in several knowledge fields. Thus, a very important task for SND is to strengthen the altruistic approach of the importance of data sharing and open access among researchers.

Experiences from SND’s predecessor SSD, show that this is not an easy task. Only a small proportion of data produced by Swedish social science researchers were deposited at SSD. SND’s conditions are, however, better than SSD’s: increased economic resources and an organization placed within a bigger network of infrastructural resources. However, an important factor for the result is the general attitudes towards data sharing among researchers. Is there simply no culture of data sharing and reuse of data among researchers in Sweden? Or does sharing and reuse exist, but not via a data service?
concerning the collection of research data, as well as the use and reuse of data. The jurist will also represent SND in the cooperation between DISC and SND on legal matters.

1.4. Possessive barriers
Data collected by a researcher or a research team are often considered to belong to the original investigator(s). This is not the case, as the ownership of the data most often is connected to the university where the researcher is employed. Nevertheless Swedish researchers often bring the data with them when changing workplaces.

Experience shows that data not used and taken care of rapidly get obsolete. Documentation gets lost and old data formats become unreadable.

When asked if they would consider depositing their data at SND, researchers often doubt that their data are of interest for other researchers. Other reasons for not sharing data with others are that data are not properly documented and organized or that reuse of data needs a lot of information from the principal investigator.

1.5. Activities to promote data sharing
The SND strategy to promote data sharing is a combination of top-down and bottom-up activities. An example of a top-down activity is to influence research funders to put higher demands on future open access data when studies are completed. To make the researcher aware of the complete life-cycle of data, the research plan always should include a plan for how to preserve and share the data. Another way of encouraging data sharing is to regard it as a merit to make your research data available for other researchers.

Another activity is to support researchers through the whole research process by providing guidelines on ethical and legal issues, on how to store and document data, etc. The SND web site will be the central place for this information, but it will also be published in different printed versions.

Examples of bottom-up activities are to be present in different research contexts and to inform about the benefits of sharing data. One example of this is a joint project with SND and the university libraries in Gothenburg, Lund, Linköping, and Malmö. Financed by the Royal Library’s Open Access Program, the aim of the project is to look into open access within the humanities and arts. The one-year project, starting in September 2009, will try to answer the following questions: Where and how to store research data? Which parts can be published as open access? How to connect the open archives and the Swedish National Data Service? How to connect research data and publications?

1.6. Feedback from the research community
When working on a strategy to promote data sharing, you need to know the opinion of the target group. Inspired by our colleagues at the Finnish Social Science Data Archive (FSD), we decided to ask the professors within the humanities and social sciences about their opinion on open access and data sharing. To compare with another target group within the research community, we also asked the same questions of the Ph.D. students within the humanities and social sciences.

1.7. Outline
The Swedish National Data archive (SND) is currently an operative key actor in conducting a national inventory survey, initiated in the fall of 2008, of existing databases and database research as well as attitudes towards data sharing among researchers and university managements within social sciences and humanities departments at Swedish universities and university colleges. The aim of this ongoing inventory survey is twofold: first, to identify and coordinate existing data resources; and second, to identify barriers and enablers to using and depositing data to open repositories. Some preliminary findings from this inventory survey and follow-up interviews with researchers based at the participating departments have revealed a number of important issues that need further investigation: the general unwillingness towards sharing information about research data with coordinating institutions (such as SND); the reported time scarcity preventing researchers from collecting, coordinating and delivering information about research data; and the ethical concerns of how to handle commitments to research subjects and how to protect sensitive information. A number of issues also clearly related to the fact that there was a wide distribution among social sciences and humanities disciplines represented in the inventory survey. There were, for example, quite different views among the respondents on fundamental issues such as the nature of and purpose of research, research ethics, ownership of research data and research results, and how to best enhance research infrastructures. In addition to the inventory process, two survey studies have recently been carried out – one targeting professors and the other doctoral students at Swedish universities and university colleges with input from the above mentioned national inventory study by SND, and from a Finnish survey, which was carried out 2006 by the Finnish Social Science Data Archive (FSD) targeting professors in various social sciences and humanities disciplines at Finnish universities and the practices related to open access to research data (Kuula and Borg, 2008).

The empirical part of this paper is based on two recently conducted survey studies, targeted at professors and doctoral students within humanities and social sciences at Swedish universities and university colleges, with the broader aim of investigating existing practices and attitudes when it comes to availability and reuse of research data. The results are tentative and descriptive and are discussed in a theoretical context in another conference paper (Axelsson & Carlhed, forthcoming).
2. Procedures of the surveys

The two surveys, one directed to Swedish professors in humanities and social sciences and the other directed to Swedish doctoral students in the same domains of disciplines, contained 80 items covering the researchers’ affiliations; domain of discipline; gender; age; and familiarities with research policies and ventures, and use, reuse, archiving practices of digital research data. Furthermore, there were questions about possible reasons for not using digital data, interventions and barriers to enhanced reuse and accessibility of data, possible agents in overcoming barriers, and willingness to engage in promoting changes in this area and to share their digital research data. The surveys were carried out through e-mail questionnaires and with lists of respondents based on retrievals from databases at the universities’ offices for IT or personnel administration. In some of these lists it was easy to recognize respondents’ disciplines; others were sorted by thematic or interdisciplinary departments and no information about discipline was accessible. Therefore, even departments that were within science and technology, educational sciences, and social medicine were included, but only departments that described themselves as interdisciplinary on their websites. Nevertheless, most of the departments were within humanities and social sciences. Because the population was broad and had somewhat non-distinct boundaries, we asked respondents to reply to us if they did not use perspectives of social science or humanities in their research. In those cases they were cancelled from the survey.

Initially, the survey was sent to 1589 professors from 35 universities/university colleges, and after the cancelling procedure of non-social science or non-humanities researchers (by the definition above) there were 1436 professors. The response rate was 38%, with 549 responses. The same procedure was carried out with the population of doctoral students. However, the lists from the universities that formed the respondent list had minor inaccuracy problems, due to some “natural” conditions, namely doctoral students becoming doctors. This affected the update status on information in the university personnel information systems, which had in some cases inaccurate information about the doctoral students. In addition, doctoral students at university colleges could also appear at a list from another university, hence with double mail addresses. A check up was made before the distribution of the e-mail questionnaire in order to avoid obvious doubles; however in some cases the e-mail addresses were abbreviated and impossible to relate to the names of the doctoral students. Similar to the professor survey, the population was broadly defined, which called for a similar procedure for cancelling, by respondents’ reply stating their non-social sciences or non-humanities affiliation. Initially, the doctoral student population included 4697 potential subjects and after the cancelling procedure (mentioned above), 4065 remained. The response rate was 28%, with 1147 responses. When comparing how the professors’ response rate patterns related to the distribution among a selection of the universities that received the largest proportion of questionnaires, we can conclude that the response rate from the larger respondent groups’ universities alternated between 22 to 44 %. See table 1 below.

Table 1 shows response rates based on the initial number of questionnaires sent before the cancelling procedure of non-social sciences or non-humanities affiliation. Because our method of selection was somewhat unstable, we found it necessary to investigate our precision further. The Swedish National Agency for Higher Education produces statistics about the universities and university colleges. Comparing statistics of professors and doctoral students and their affiliation to university and disciplinary domain from 2008 and our response patterns gives a view of how our survey succeeded in targeting the population. It seems that the population of professors (constructed from statistics, i.e., number of professors in different domains of disciplines and university), is well-covered by our group of professors who have participated in the survey. In concordance with this one can argue that our response frequencies are higher in reality (see table 1).

Table 2 shows the doctoral students’ response rates
related to the distribution among a selection of the universities that received the largest proportion of questionnaires.

Like the professors’ response rate patterns, the table above shows response rates based on the initial number of questionnaires sent before the cancelling procedure of non-social sciences or non-humanities affiliation. As argued above, the actual response rate is higher when comparing it with the statistical population, which we constructed for comparison reasons. For some universities, however, the response rate was lower in this comparison. It signals distortion in our precision about the doctoral students. In conclusion, our generalization opportunities are limited due to these aspects that have been discussed above. It seems that the ground for conclusion about the group of professors is more stable than the group of doctoral students. Nevertheless, a large number of professors (N=549) and doctoral students (N=1147) have participated in our studies, which implies considerable opportunities to assume valid conclusions.

### 2.1 Generalizability

In the professors’ group, a majority of men answered the questionnaire, 73% compared to 27% women. This reflects the demographics of the larger population, whereas 23% of the professors in social sciences, humanities (and law) are women. In the group of doctoral students the conditions were opposite; 61% of the doctoral students in our survey were women. In comparing with the statistics from The Swedish National Agency for Higher Education, the larger population consisted of 56% women. In both cases we can conclude that women were slightly more inclined to participate in our surveys than men.

Considering age, with our survey we seem to engage a larger part (25%) of the younger doctoral students (younger than 29 years old), than expected (16%). The same counts for the group of professors, but there were only a minor difference. Two percent more of professors participated that were younger than 50 years old (18%), compared to statistics from The Swedish National Agency for Higher Education (16%).

According to domains of disciplines, it seems that our groups of professors and doctoral students reflect the structure of the larger population (table 3).

Based on the discussion above, our conclusion is that the results from our surveys could be treated as fairly valid, in spite of the relatively low response rate. The amount of responses from professors and doctoral students in different domains of disciplines, age, and gender corresponds to the official statistics that have been described and discussed.

### 3. Results

The Swedish Research Council has in a current venture made a long-term strategic plan - a roadmap *The Swedish Research Council’s Guide to Infrastructure* (The Swedish Research Council, 2007). In the questionnaire we asked the researchers about their knowledge about this venture and their opinions about it. Eleven percent of the professors were familiar with the venture and the guide and only 1% of the doctoral students. Half of the professors’ group did know about the venture but not its details and 40% did not have knowledge about it at all. This was also true for...
the majority of the doctoral students (82%). Professors were more inclined to express positive opinions about the venture and the doctoral students followed the same pattern (Figure 1).

The knowledge about the OECD Guidelines on Open Access to Research Data from Public Funding (2007) was generally low; 75% of the researchers (both groups) did not know about it at all. Surprisingly, 61% of the professors were not aware of its existence (Figure 2). Breaking down results by domains of disciplines, it seems that professors within social sciences are the most informed about the OECD guidelines, and the group which was least informed was the doctoral students within law. Considering the situation of being a doctoral student, we are not surprised at the large amount of them not having knowledge about the guidelines and/or the research venture mentioned above.

3.1. Archiving practices and reuse of digital research data
The primary condition of archiving and reusing digital research data is that data are collected and compiled in some way. 73% of the professors stated that digital empirical data are used in research and 16% stated that the use of digital empirical data is unusual or are never used. The major part that did not use empirical digital data was the professors within humanities (56%). Among the doctoral students, they declared that digital empirical data was used (42%), but they expressed a great uncertainty about these questions generally.

According to the professors, the digital data are often kept by the researcher after analysis and reporting, without any actions to documentation (46%) but according to 15% of the researchers, they occasionally keep the digital data without any further documentation. Archiving practices where digital data always are kept and documented in a catalogue/database at the university is quite rare (11%). Almost half of the professors’ group stated that these practices were unusual or never occurred. The same tendency showed concerning facilitating availability of digital research data at a data archive, that is, unusual practices. However, it seems that the research data are not regularly destroyed after analysis and reporting; at least 49% stated that destruction is rare.
and only 3% reported that it was common. The reuse of
digital data are relatively common; 59% of the professors
stated that data are reused in Ph.D. works or other research
projects and only 3% reported that it never happened.
The use of reused digital data in teaching is also quite
frequent according to 59% of the professors. Reusing
all kinds of empirical data is most common in situations
when researchers use the data themselves; approximately
one-third stated that they pass data on to other researchers
who are studying similar kinds of areas. Five percent
of the professors reported that this never occurred.
Regarding the amount of the digital data that are reusable,
professors are more optimistic in general than the doctoral
students, who seemed very uncertain and had difficulties
expressing opinions of estimates (Figure 3). There were
small differences between both groups and the domains of
disciplines in these issues.

The wide range of empirical research data within social

and sciences and humanities is shown in figures 4 and 5.

In both domains of disciplines researchers use a broad
empirical base, especially when it comes to use of non-
digital empirical materials, where 74% of social sciences
researchers and 86% of humanities always use several
empirical sources. When considering the use of digital data
it seems that the empirical data are less varied, according
to 66% of the researchers in social science and 68% of the
researchers within humanities.

Important reasons for not reusing digital data are mentioned
by the professors as uncertainty about the quality of data
(62%), ethical aspects (57%), technical issues (53%) and
juridical issues (49%). However, the professors’ group is
divided in opinions and the other part does not think that
these factors are crucial (38%-50%). According to those
who think ethical aspects are important, we found that these
professors were mainly from social sciences. That is true
also for doctoral students in the same domain of discipline. The importance of juridical aspects is represented by the doctoral students in law, but not the law professors. Both professors and doctoral students in humanities deviated in general from the others in these issues, i.e., the technical issues were considered important. They also report other reasons for not reusing digital data, such as not using empirical or/and digital data at all, lack of knowledge and routines, decontextualized data having weak relevance for others, etc.

Concerning interventions to increase reuse of digital data, 95% of the doctoral students and 93% of the professors thought it should be effective to get more information about accessible research data in data archives or databases. Nearly 100% in both groups reported that also more of training in research methods, digital research databases, and information about accessible e-tools would be effective interventions (89%-95%). It seems that professors and doctoral students in humanities are most positive towards more education interventions and researchers in social sciences are the least positive, but all groups are generally positive to the interventions proposed.

### 3.2 Obstacles to sharing digital data

Our seven suggested obstacles to sharing digital data have been ordered in degree of perceived difficulty by the respondents. The professors regard deficiency of resources for researchers to document and arrange their data for reuse as the most difficult obstacle to sharing digital data. They also ranked lack of other resources like guidelines and directions for documentation as an important issue. Another obstacle highly ranked by the professors was doubt about the correct use of their data, i.e., risks of mistakes and misuses. An additional impediment was the fact that their respondents were not informed that their contributions should be used in the research society in general, only for a particular study. Juridical obstacles and loss of one’s own advantage of competition in keeping data to oneself were not considered as crucial. The least important obstacle, according to the professors, was ethical aspects such as threats to confidentiality and delicate information. The doctoral students, however, thought that the ethical aspects mentioned above were the most difficult obstacles of all. After that, they considered the information given to the respondents and the use of their contributions to the research society in general was an important issue. Deficiency of resources for researchers to document and arrange their data to for reuse were also ranked as important, followed by juridical aspects. The least important obstacles according to the doctoral students were lack of other resources like guidelines and directions for documentation, loss of one’s own advantage of competition in keeping data to oneself and doubt about the correct use of their data, i.e., risks of mistakes and misuses of data. The response pattern did not change depending on the researchers’ use of digital data or not.

On the other hand, researchers in social sciences and women were more concerned with research ethical aspects and threats to confidentiality, etc., while researchers in humanities and men tend to stress lack of resources to document and arrange their data for reuse. According to age, older researchers tend to emphasize lack of resources and juridical issues. Younger researchers pointed out ethical aspects, threats to confidentiality, and doubts about incorrect use of their data. Thirty-eight percent of the professors and 34% of the doctoral students stated that these obstacles did not prevent them from sharing data with The Swedish National Data Service (SND). However, 65% of the researchers indicated that these obstacles did prevent them from sharing data to SND. When we asked if, for example, SND could help them to overcome these obstacles, 43% (of the 65%) responded positively. The research funders were also regarded as important agents in overcoming the obstacles by 43% of those who expressed that the obstacles prevented them to share their data. Researchers who did not feel prevented to share data believed to a greater extent that SND could be of help (55%) and research funders as well (59%). There were however, minor differences, where the older researchers and the researchers in humanities were more optimistic about overcoming the suggested obstacles. There were very small differences between professors and doctoral students. On the other hand, if the researchers would consider engaging in promoting alterations in these areas, the doctoral students tend to look forward to them, while professors are inclined to issues of jurisdiction.

Surprisingly, when it comes to the degree of urgency in sharing their own data, the professors are a bit more eager to share data (30%) than the doctoral students (24%). In total, there were 53% of the researchers who thought it was urgent to share data (55% of the professors and 52% of the doctoral students), but only 26% in the total group reported that they intended to share their data. A large proportion of the total group expressed doubts about sharing data (40%). Researchers in law were the least keen on doing it and thought it was not so urgent. Researchers in humanities however, were those who distinguished themselves as potential “sharers.” According to gender and age, men and older researchers expressed more willingness to share than others.

### 3.3 Promoting accessibility to digital research data

The results show which authorities the researchers point out as important key actors in promoting accessibility to publicly-funded digital research data and also to actively participate in shaping guidelines. The universities and university colleges were considered as the most important key actors according to 82% of the researchers and the second was the two largest research funders for social sciences and humanities, The Swedish Research Council.
and The Swedish Council for Working Life and Social Research, with 80% of the researchers’ responses. Fifty percent stated that Statistics Sweden would also be participating in shaping such guidelines. No significant differences between doctoral students and professors, or age, in this matter were observed. According to domains of disciplines there were very small differences, except for the opinion about participation of Statistics Sweden, where researchers within social sciences emphasized this as a key actor to a larger extent than the others. It seems also that male researchers are a bit more pessimistic about the importance and role the proposed key actors could play.

Furthermore, the most effective interventions for enhancing accessibility to digital data were that research grants should include funds for preparing the data for sharing and archiving (88% of the doctoral students and 83% of the professors) and that making data accessible for the use by the scientific community is acknowledged to be scientific merit (87% of the doctoral students and 83% of the professors). Generally, the doctoral students were more optimistic about the efficiency of interventions proposed, especially the issue of acknowledgement of promoting accessible data to be of scientific merit and except for those mentioned as top-ranked above, where the professors expressed beliefs in their efficiency in a higher degree. There were no or very small differences in response pattern among the domains of disciplines in these issues. According to gender it seems that women researchers believe in more education about life cycles of digital data (74%) and research ethics (77%) to a higher degree, than men. 59% of the male group thought that more education about life cycles of digital data is needed and 63% of the male researchers believed that more education about research ethics is necessary.

4. Discussion
Our results are descriptive and have been presented tentatively in this article. Further statistical analyses are needed concerning impact of differences and in addition, scrutinized examination of all opened questions, where a lot of interesting comments are made by the researchers.

Overall we interpret the researchers’ attitudes towards current ventures and strivings in research infrastructures as predominantly positive. A key actor is The Swedish Research Council that has, in a current venture made a long-term strategic plan - a roadmap The Swedish Research Council’s Guide to Infrastructure (2007). The researchers’ knowledge about this venture was minor. Professors were more inclined to express positive opinions about the venture and the doctoral students followed the same pattern.

The knowledge about OECD Guidelines on Open Access to Research Data from Public Funding was generally low and somewhat discouraging; professors within social sciences were the most informed, however. The least informed were the doctoral students within law. In comparison with the Finnish survey (Kuala & Borg 2008), where 81% of the professors did not know about the OECD recommendations, compared to 61% of the Swedish professors, it could be encouraging from the Swedish perspective. Nevertheless, time has passed between these surveys and perhaps have the Finnish professors got more informed than they were in 2006.

A conclusion based on our results is that it seems important to raise issues of guidelines in social sciences and humanities concerning accessibility to digital research data and to engage researchers and relevant authorities in creating arenas for discussing and shaping research infrastructure for the future. According to the researchers, the universities, university colleges, The Swedish Research Council, and The Swedish Council for Working Life and Social Research are the most important key actors in promoting accessibility to digital research data from public funding through participation in shaping guiding principles.

The most effective interventions for enhancing accessibility to digital data that were identified were that research grants should include funds for preparing the data for sharing and archiving and that making data accessible for the use by the scientific community is acknowledged to be of scientific merit. More education about life cycles of digital data and research ethics were expressed as needs.

Considering archiving practices, use and reuse of digital research data, 16% of the Swedish professors stated that the use of digital empirical data is unusual or are never used. Eighteen percent of the Finnish professors reported a similar amount of digital data non-use. When comparing between the countries what happens to digital data after analysis and reporting, it seems that it was more common for Finnish professors to keep digital data without any further actions to documentation (56%) compared to Swedish professors (46%). Data are destroyed to a larger extent in Finland (20%) than in Sweden (3%). However, the saved data is reused by the researchers themselves to a greater extent in Finland (94%) than in Sweden (54%). The opinions about amounts of reusable digital data differ also; 50% of the Swedish professors stated that more than half the amount of produced digital data is reusable, compared to 21% of the Finnish professors. In analyzing responses to important reasons for not reusing digital data, it appears that the Swedish researchers emphasize ethical, juridical, technical aspects, and quality of data as more problematic than the Finnish researchers. Again, it seems that there are a considerable amount of issues that need to be clarified and solved in order to develop a well-functioning research infrastructure with a high degree of re-using practices within social sciences and humanities. As information of importance, we think that the researchers’ beliefs that promoting accessibility of their own data to
be acknowledged as a scientific merit and that research grants should include funds for preparing the data for sharing and archiving, points out legitimate measures with both force and enticement - like the stick and the carrot. The last mentioned intervention was also one ranked high by the Finnish professors (80%), but their top priority of effective interventions was establishment of guidelines and principles by the Finnish universities together (84%).

The Swedish professors also point out other obstacles to sharing digital data, and regard deficiency of resources for researchers to document and arrange their data for reuse as the most difficult obstacle to sharing digital data together with lacking guidelines to documentation, while the Finnish professors reported that it was the situation when the respondents were not informed that their contributions should be used in the research society generally. They share this concern with the Swedish doctoral students. These results relate to the results mentioned in the former paragraphs, namely the need of different types of guidelines (ethical, technical, and juridical), earmarked resources to documentation and education in this area.

As we mentioned in the Results section we found that the professors seems to be more eager to share data than the doctoral students. A large proportion of the total group also expressed doubts in sharing data, probably because of uncertainty and lack of sufficient guidelines.

The Finnish questionnaire did not have a pushing question like we had, but the Finnish professors were asked their attitude to open access to digital research data collected in their own research and 76% of them expressed positive attitudes. One might conclude that professors in social sciences and humanities in Sweden and the Finnish professors differ a lot in opinions about digital research data. However, two years have passed with increasing focus on open access issues in research policies in these countries. It would be interesting to see if the Finnish professors have changed their minds since 2006.

About our own results, it is always interesting when the research is surprising. We were surprised that the professors were more positive and humble towards sharing and promoting accessibility to digital research data than were the doctoral students. But on the other hand, being a doctoral student means a lot in terms of keeping on one’s own’s track, concentrating on the Ph.D. work, and having little time to orient oneself to ventures, research policies, and university practices. In conclusion and in spite of many prejudices about “conservative” professors, it seems that one has to acknowledge their positive orientation about e-science and put forward these survey results of barriers and opinions to be able to support and realize sharing of digital research data in the future. At least, the results of these surveys have to be acknowledged and seriously taken care of in understanding the obstacles and challenges we face, in order to achieve a sufficient and approved research infrastructure, adapted for the distinctive features of social sciences and humanities with their wide range of empirical materials.

5. Acknowledgments
The authors wish to thank the doctoral students and professors who participated in the surveys.

References

Database Infrastructure Committee (DISC). (forthcoming). Juridical conditions for a research infrastructure.

Högskoleverket. (Swedish National Agency for Higher Education). http://www.hsv.se/statistik/statistikomhogsklan/personal.4.6df71dcd1157e43051580001770.html


Notes
1 Carina Carlhed, Mälardalen University, Sweden, Email: carina.carlhed@mdh.se, and Iris Alfredsson, Swedish National Data Service, Sweden, Email: iris.alfredsson@snd.gu.se.

2 http://www.hsv.se/statistik/statistikomhogskolan/personal.4.6df71dcd1157e43051580001770.html. In this paper, comparisons have consistently been made between statistics from The Swedish National Agency for Higher Education for year 2008 and the background information about the participants in our surveys.
The **International Association for Social Science Information Service and Technology** (IASSIST) is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the **IASSIST QUARTERLY**. They also benefit from reduced fees for attendance at regional and international conferences sponsored by IASSIST.

Membership fees are:
- Regular Membership: $50.00 per calendar year.
- Student Membership: $25.00 per calendar year.
- Institutional Subscription: $75.00 per calendar year (includes one volume of the Quarterly)

Please make checks payable, in US funds, to IASSIST and Mail to:

**IASSIST**,  
Assistant Treasurer US  
Shelagh Mackay  
University of California  
Davis, CA, USA  
95616-8617

---

**Membership form**

Please see my choice below:

- [ ] $50 (US) Regular Member
- [ ] $25 Student Member
- [ ] $75 Subscription (payment must be made in US$)
- [ ] List me in the membership directory
- [ ] Add me to the IASSIST listserv

---

**I would like to become a member of IASSIST. Please see my choice below:**

Options for payment in Canadian Dollars and by Major Credit Card are available. See the following web site for details:

[http://www.iassistdata.org/membership/membership.html](http://www.iassistdata.org/membership/membership.html)