Title	**Publishing problems of a CESSDA Common Data Portal (D5.2)**
Work Package	WP5
Authors	Atle Alvheim, Reiner Mauer
Date	19 August 2009
Dissemination Level	PU (Public)

Summary/abstract

The general aim of the future CESSDA portal is unified, flexible and useful access to data from the European Social Science data archives. This requires that the processes leading up to having data stored in the archival repositories are standardised and follow defined best practices, to facilitate development of user functionalities. Commonality in metadata standard and implementation will be fundamental for procedures and may facilitate use of common tools.

The data publishing process of the future CESSDA data infrastructure has to solve at least two important data problems beyond present status:

Complex data, i.e. datasets that consist of more than one square file put together in a collection have to be adequately described. Data are sometimes modified and we have to have ways of handling such situations in our data storage, data may be versioned.

The Data Documentation Initiative[1] (DDI) is a project run by and for the data archives in common. The aim of the DDI project is to develop a timely metadata standard that meets the needs of the data archives for data dissemination in the age of the Internet. Compared to version 2 (DDI2), version 3 (DDI3) of this metadata standard is a larger collection of documentation elements, it is based on a web-services paradigm and is implemented in XML. DDI3 aims to solve both problems mentioned above; this requires development of several supporting tools and technologies. Implementing DDI3 in all its aspects therefore becomes a large and long-term project. The present document discusses some practical problems posed by complex data collections and indicates simplified practical strategies and a somewhat stepwise implementation procedure for some specific cases.

[1] http://www.icpsr.com/DDI/

Publishing complex data

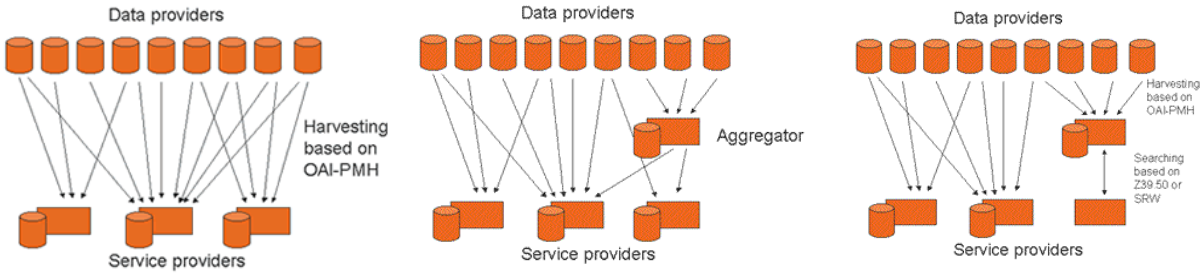
Data packages are supposed to be stored in a set of decentralised data repositories or archival storages.

These data repositories are bound together or standardised through support for a common metadata standard, the DDI2/3. Support of a DDI-based (meta)-data model is the backbone of communication- or interface standards-/protocols.

The data packages contain metadata, (but not necessarily data). The actual content of the data packages may be influenced by access conditions and access policies.

Task 5.2 of WP5 was to analyse the problems of bringing data into such a system, the data documentation/ingest/publishing problem related to bringing data into archival storage in such a way that it safeguards support for data location, data exploration and data download functionalities in addition to the safe long-term preservation of the data. It is difficult to separate this from general discussion of metadata organisation since functionality development is strongly based on explicit use of the metadata component.

The term "ingest" applies to the development and loading of documented data packages or instances into the archival repositories, the ingest process receiving Submission packages from a general collection/preparation stage and delivering it in a systematic way as Archival Information Packages into the archival repository. Data are stored in a distributed set of archival repositories; there are repositories in Cologne, Amsterdam, Tampere, etc. A repository cannot always be regarded as a simple unified storage. In a service-related situation it may be a layered hierarchy of different configurations representing the relationship between storage and use, and the dimension spanning the distance from storage to use may be regarded as orthogonal to the general preparation process. In the OAI[2] terminology there is a distinction between data providers and service providers, to illustrate the aggregation of services. To some degree this fits with the repository-internal processes of CESSDA repositories.



Held against the OAIS Reference Model[3], it is not clear which of the two levels illustrated above most adequately represents OAIS storage, but the distinction underlines the point that archives cover two basic aims: preservation of data and active scientific use of data. CESSDA data repositories are focused both on preservation and use, data are not only stored but are required to be actively used. Science to a large degree is about studies of relationships and this generates a need for additional functionalities in treatment and preparation of data, i.e. standardization,

[2] Open Archives Initiative Protocol for Metadata Harvesting
http://www.openarchives.org/OAI/openarchivesprotocol.html

[3] Reference Model for an Open Archival Information System http://public.ccsds.org/publications/archive/650x0b1.pdf

harmonization and extra documentation. Such functionalities are activated outside "the walls" of the archive and how to connect "versioning needs" back into archival storage then becomes a problem of its own. And it may be complicated to decide which level actually has to be versioned.

The OAIS model presents a process with a SIP, via AIP to DIP. In relation to CESSDA one important task has been to develop a Concepts, Classifications and Conversions Database, a tool which develops and stores data harmonization work. Research is a cumulative process and research is one of the few processes where it is legitimate to stand on the shoulders of others. This generates a need to feed the continuous work on data refinement back into the repositories and make them somehow available for future use by others. This of course involves extensive documentation and explanation. It is not intuitive where such standardization and harmonization processes come in: is it an activity at a stage between receiving a SIP and storing an AIP; is it more functionality related, between the AIP and the DIP, or maybe both? Neither is it exactly clear what its character will be. The main concern here is how we standardise the documentation, the metadata production process, as much as possible across CESSDA members while still allowing great flexibility in tools when we populate these repositories with data. Our concern here is therefore linked to all aspects of preparation and use of data.

The general idea/aim is that CESSDA archives should work towards producing DDI3-compatible XML as the standard transport (and storage) format into and out of the local data repositories. However, there are not yet any good implementations or tools available producing DDI3-compatible output of the required complexity. This does not change the ultimate aim but requires a well-reflected strategy for working towards that aim in the most efficient and convenient manner. In this report the strategy therefore will be to try to contrast intermediate solutions related to DDI2 or DDI3 respectively and to discuss what potential data complexity and portal needs can be covered. This implicates to some degree to play down importance of tools and instead focus on products.

Presently we have a list of DDI-add-ons like ELSST[4] (here intended to be used to deliver standardised concepts or keywords across languages into the documentation process) and a large systematic collection of controlled vocabularies[5] directly related to the elements of DDI3. What potential use we can make of a gazetteer[6] has been questioned, while the CESSDA study classification is a useful simple controlled vocabulary. A Concept, Classification and Conversion database (3CDB) or a Question and Concepts (QDB) database are somewhat more complex controlled vocabularies, but in this connection, they are also basically functioning as controlled vocabularies. These two last mentioned databases may play different roles and support various functionalities in the overall infrastructure.

The data preparation process delivers data (AIP) to the local repository designated for that purpose, it could as an example be a Nesstar or a FEDORA-based server: The process goes:

1. Various resources are collected, among them the data matrix, the necessary metadata, questions, concepts, etc This is ordinary archival work;
2. Data are described, supplied with additional use-oriented metadata and loaded into locally maintained repositories;
3. Parts of the Metadata are then collected from these local repositories to build registries/indexes to support location and exploration functionality;

[4] European Languages Social Science Thesaurus http://gandalf.aksis.uib.no/lrec2002/pdf/3.pdf
[5] http://www.controlledvocabulary.com/
[6] http://en.wikipedia.org/wiki/Gazetteer

4. Portal functionality access repositories based on the index, the functionality may be supported by a QDB etc.

These four points are the essence of the CESSDA data infrastructure, and this report addresses the two first points, as basis for the last two points, which make up the "portal" as such. The main concern of this report is to understand problems related to how complex organised collections of data may pass through the two first stages so that they may be available for later parts of the process in ways similar to how simple data files are treated.

The report delivered by Metadata Technology on a technical specification for a European Question Database introduced and elaborated some general ideas for portal development based on the same basic philosophy as the Madiera[7] project and the Nesstar[8] tool. The (XML-coded) input data and metadata packages are basically used only for transport of content into the repository and not yet as the basis for functionality development. The actual metadata are broken up again in the data repositories, to make them more useful for functionality development. In the referred report there is talk about legacy databases, the internals of the Nesstar server, as an example, holds metadata in a relational database, modelled on the basis of DDI 1.2 with some adjustments (DDI1 or 2 cannot be completely expressed as a relational database). Reflections around the 3CDB and QDB databases here are very much in line with what is outlined in the MT report, although at the outset we tended to think more in terms of general harvesting as the basis for index development more than controlled publishing to registries. The MT report is obviously correct when holding that the publishing requirement and a non-deleteable status of elements make the process and products easier to control and administer and creates better persistence in a service-oriented setup. There might be a slight variety in vocabularies used. Publishing/ingest in the vocabulary here is to put data collections (AIPs) into the local repository, while in the QDB-report publishing means that some metadata are made available for input to a central registry through a registration process and not all metadata will automatically be published.

General background for the portal discussion

It is a requirement that data should be made available for storage in a standardised way that will support data location, exploration and data retrieval (i.e. the portal needs), even for complex comparative, over time repeated data and micro-macro integrated data (i.e. support a variety of more complex data models).

In the original application the problem had a simple formulation. A more detailed stepwise restatement summarizing history and arguments could be:

1. CESSDA presently has a general publishing strategy for data documented under the DDI 2 level metadata standard. This we have seen implemented in two general ways:

    - either through archive-internal developed database solutions holding large amounts of metadata often employed for several purposes, or;
    - via the Nesstar Publisher as a tailor-made tool for data documentation and publishing or through other more general XML-editors.

[7] http://www.madiera.org/
[8] http://www.nesstar.com/

Both strategies produce reasonably standardised DDI2 XML-files, in the CESSDA context presently most commonly published to the Nesstar server for Internet presentations. Several data archives are exploring the potential of Fedora as an alternative storage possibility.

2. Development of the documentation standard, DDI 3.0 from the DDI 2.0 generation was triggered by ambitions at several levels:

   - There is a need to handle more complex data structures. In Europe there are several sets of larger collections of data collected over time and many countries. ISSP and ESS are prime examples;
   - There is a need to handle "dynamic" data, i.e. versioning of data (the implementation of a life-cycle perspective). Data may change or accumulate additional content over time;
   - The need to run a more economic and efficient process focused on re-use of elements and distribution of work between producers and archives, i.e. the same question may be re-used, the same variable schema may be re-used;
   - The need to steer this development as a generic process and to develop and introduce a comprehensive standard for data archival metadata, with integration of processes and reuse of material, not only within one archive but across a whole community of data producers, data archives and data consumers.

3. However, the suggested implementation of DDI3 that we have seen realised so far represents two complimentary ways of thinking and the sub-points above do not carry the same weight for all interested parties. The major ambitions for the CESSDA part of the archival world were for the CESSDA-PPP an upgrade of the common data portal and the data documentation processes as a general expansion towards being able to handle complex files, the comparative problem and also the data versioning problem. However, this has as a more general activity been expanded by an important aim to develop a generalised common architecture for varieties of web-services based on an object-oriented architectural thinking. This should facilitate modularity in software and standard, general reuse of material, referencing of metadata objects to enhance efficiency and more economical work, in addition to adequate documentation of complex studies (tools & technical metadata standards to structure, capture, host substantial knowledge from concepts to harmonisation details) to extend, exchange and provide high quality data and metadata. This raises the ambitions formidably.

4. The ingest / publishing process should in such a context function server/ repository independent, for any potential repository solutions that may support DDI at the necessary level. The requirement is that publishing and storage technology meet in DDI-compatibility, and in this (5.2) specific connection it is groundwork for a functional specification that is being asked for. The ingest process works against single local repositories, while the user-oriented backend side might work via a common portal as an extra layer of the access function.

5. Ingest and repository storage development is in this context often dominated by an archival perspective, which always slightly provocatively represents the danger of becoming data graveyards. To counter that, data have to be brought further, from storage to analytic software and the portal component explicitly make this more instrumental. This whole endeavour is intended as an integrated production line where elements are dependent upon each other and the decisive ultimate aim and criterion is to make data potentially available

for analytic use by external users. Documented data are of limited use as such without a repository technology that supports development of the necessary functionality of data location, data exploration and data access. A repository is of little use without efficient data documentation, and that all data have to be carried further to analytic use, it is not storage alone that is of interest. Even if it is outside the immediate aim of this report, we could add some further comments: the analytical needs are so diversified that it is difficult to envision a repository-handling technology that also covers the analytical needs for all kinds of data-structures, so this most likely has to be a task for more specialised analytical software. R could be a good candidate for the final analytical functions, as flexibility is also here a keyword. Users are not expected to be interested in DDI-versions, or may be downright negative, because of complexity and the amount of resources it takes away from potential substantive research. Only a limited proportion of the users would also be interested in R, simply because flexibility and potential are very demanding on users. Few users may see the long term benefits of an elaborate metadata standard, so decisions on implementation and use of metadata standards will be a political question. And as users are not always analytically sophisticated data repositories have to deliver data to the most common and comprehensive user tools: these users have to inform the guidelines.

This point needs to be discussed. The data archives are in practice spanning two major projects, run by different authorities and with different agendas. The CESSDA-PPP project could be seen as quite flexible, focused on solutions for immediate European problems, whereas the DDI Alliance[9] promotion of DDI3 operates with absolute givens. For that reason we try here to contrast differences in potential solutions and the degree to which intermediate solutions will function as steps in the right direction.

6. The technical architectural considerations for this whole complex are related to how we break up the various content components of DDI into collections of objects and develop an information model to facilitate the modularity and reuse. In the DDI 3.0 specification, this is defined as an explicit aim and is carried relatively far based on XML state-of-the-art and object-oriented web-services thinking. However, it becomes very complicated technically to develop good tools given the complexity and magnitude of this objects system so it is legitimate to investigate what alternative strategies may be possible or necessary for development of such Lego-systems until they are better established and been evaluated in practice. Break-up of the traditional single sequential XML-file and use of the smaller components for building processes and products that have to be put into the system, stored somehow in the system and found and transported out of the system seems to be a great idea. From an input perspective we could put together an archival product that is flexibly stored in some kind of relational or legacy database. XML-files could be transport vehicles in, and to some degree stored and potentially used for transport out, of this server core. It is an extensive amount of work to develop good tools that cover the needed functionality and some specific elements seem to create problems, in particular versioning or data dynamics. Generally, the possibilities opened by DDI 3.0 seem formidable. But there are also a lot of unresolved questions. The object model is very large and complicated and the creation, administration and maintenance of the identifier systems across a large user community is a formidable problem. Some of the problems could be worked on, if not solved, through alternative implementations, but generally with less computer actionability. The actual implementation facet of DDI is the one factor that may be played around with; few are questioning the content or the components.

[9] http://www.icpsr.umich.edu/DDI/org/index.html

7. Here we will mostly concentrate on one problem: how do we develop descriptions of complex data structures? The other side of the coin, how we develop functionality to use these products for constructive purposes, should not be under-estimated, but is not our concern here. The interface technology we already have available for the development process, exemplified by Nesstar Publisher or the GESIS-developed DSDM or CBE tools, could potentially include more group-level information and deliver the functionality outlined by DDI3 based on a step-by-step extension of DDI2 that is further processed. The DDI 3.0 -specified grouping/comparative scheme could be taken as a specification of required metadata elements and the recording/measurement of fundamental relationships, and a solution could be based on that. However:

    a) the solutions may be too descriptive and impractical and;
    b) this is only intended to cover the transport of data into archival repositories until more appropriate and full DDI3-based tools take over the task.

    The question is whether we should build this up pragmatically and demonstrate usefulness through example, or take the DDI3 standard, both in content and technical architecture as a given. In the first instance, it is important to develop this as an argument for further development, not against.

Clarification of the specific problems

The ultimate aim of the CESSDA-PPP is that the distributed set of European data repositories should be bound together by a common (virtual) data catalogue and portal mechanism for data discovery, exploration and retrieval. We have a defined general aim for the work.

However, the major justification for developing CESSDA as a set of decentralised nodes for (national) data collection and preparation is that this is the structure, maintenance and ownership solution that will generate the most data of relevance for social science research; it is not triggered by technological concerns.

To be able to specify the documentation input process that every participant has to follow, we could structure our discussion along two lines:

A: What problems are we supposed/trying to solve? We chose to work backwards from the functionalities we needed to supply for the requested services, and that way tried to establish the associated metadata needs that a CESSDA portal would require. In addition the data documentation problem also had to take into account our various metadata add-ons

- There is underway development of a partly separate Questions and Concepts Database, this is intended to be used at different point in the overall scheme to enrich and rationalise the data documentation and support functionalities for data exploration;
- There is also underway development of a separate Concepts, Conversions and Data Harmonisation Database, for similar, maybe more, use- and user-oriented purposes;
- CESSDA has over the years contributed substantially to a Multilingual thesaurus, potentially functioning as an important hierarchical controlled vocabulary of substantive concepts;

- DDI implementation work and CESSDA have developed a row of other lists of Controlled Vocabularies and classifications that may be merged into the process via a DDI profile/common template, or similar.

Some background:

The report delivered by Metadata Technology on technical specifications for a European Question Data Bank outlines ideas for a (virtual) QDB as a registry[10], a common catalogue of specific pointers, developed on top of what is made available and stored in distributed CESSDA nodes. The service function of the nodes, the CESSDA distributed repositories, are pictured as grouped sets of logical metadata objects, "banks" of specific types of content-related data elements that are grouped together for use. Development of create, update, retrieve and delete operations are all crucial to these banks functioning correctly. Likewise, these banks should have functions for grouping and comparability, functions that bring together or link elements into a wider system.

The build-up of the repositories and the integrating registry is outlined as a triggered and very controlled process, and inserted elements (pointers) should be non-deletable to protect future integrity of the system. Inserting new versions of objects is therefore also a decentralised operation. This seems to be a realistic and implementable strategy that makes it possible to administer processes and maintain such a decentralised setup, although with some administrative overhead. This suggested setup bears similarities to the architecture behind some of the tools we already have, and in this report these ideas will function as an important foundation. However, the MT report takes DDI 3.0 as a given, while it might be expected that the implementation problems related to DDI 3.0 are so formidable that it should more be regarded as a longer term aim.

Because we are functioning in a world of limited resources: How generic a solution do we need or can we afford, given the implementation costs? Do/can we risk/afford to stop development by not going straight ahead for DDI 3.0 and take that as a given framework?

For the Questions and Concepts database and a Harmonisation database it is also a question of how to generate/administer/update/integrate these components within a larger structure.

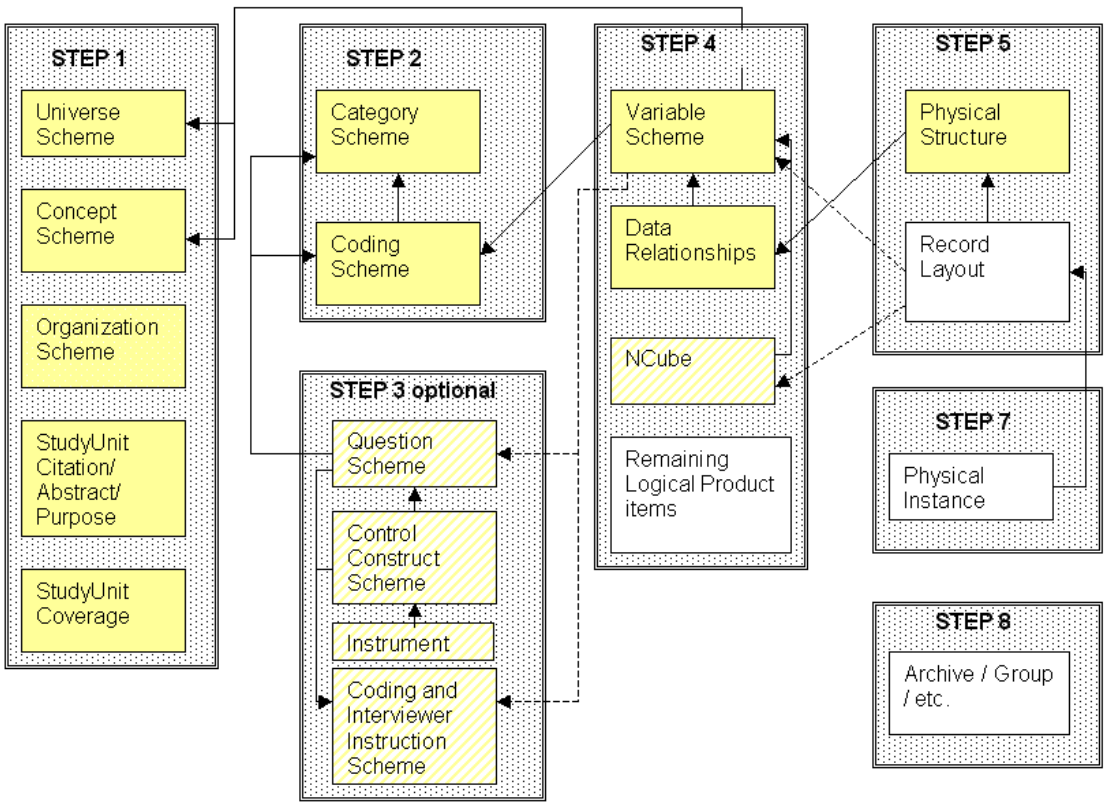
Web-services and a service oriented architecture are not an absolute requirement for the ability to solve these problems, but it may well be that it is the best. It is claimed that it is easier and more flexible to integrate components in a decentralised and networked structure that way. Certainly it seems like a more efficient way of integrating different centralised functionalities on top of a very decentralised structure. But it is also more development work and there are other open questions.

For the CESSDA common portal, the OAIS Reference Model functions as a general overall framework. A further good description and background for data-types, functionalities and metadata requirements from a process perspective can be found in a conceptual paper developed as part of WP8's Task 1, and which details central objects and core processes.

[10] See Wikipedia
http://en.wikipedia.org/wiki/Metadata_registry

Conception	Production			Repurposing	
Research project concept & study design	Data production/ collection	Data & Metadata processing	Data & metadata archiving, Publishing	Data & Metadata Exploration & Discovery	Data Analysis
Researchers	Data producers	Data providers/archivists		Researchers	
Translation	Q-DB Instruments	Q-DB H-DB	Concepts Thesaurus Translation	Concepts Categories Classifications	H-DB
		Ingest	Store Repository	Discovery Exploration Download	Compare Data use
Context	Instruments	Harmonisation Controlled Vocs			

The illustration above is a tentative summary in tabular format. This high level conceptualisation of the use-oriented problems above could be contrasted against the DDI3 phasing model for data documentation work copied in below:



By contrasting these two tables it is possible to outline a detailed DDI profile or selection of metadata elements, indicating status in a common CESSDA metadata recommendation, with a distinction between what is mandatory, what is recommended and what are potentially optional

elements. In tools this may be implemented as a configurable setup or template for different types of data. The simplest practical way of doing this might be to:

1. Start from the present CESSDA Common template, published at http://www.ddialliance.org/DDI/related/cessda-rec.pdf;
2. A mapping of metadata elements between DDI2 and DDI3 is available as a spreadsheet at http://www.ddialliance.org/DDI/ddi3/mapping-spreadsheet.pdf or as a tree-structure at: http://www.ddialliance.org/DDI/ddi3/variable-fields.txt.

The DDI3 model above also clearly indicates functionality needs related to data documentation activities.

Researchers have already expressed (in the evaluation of the CESSDA-PPP application) that they value access to a database or a thesaurus giving overviews of concepts, categories, classifications and a tool like a harmonisation procedure to find and explore potential comparable research data and to establish and enhance the value of the data, more than anything else we had suggested. In a CESSDA data portal such a tool has to be located at the exploration and user end of the process and not as part of the production and archiving process.

In this present outline the ISSP Role of Government data are used as our prime use case demonstrating both the comparative problem and relationships over time, two of the more complex problems encountered. The first premise is that we seek a solution that allows us to publish one relatively integrated product (one instance?) to a repository, one product that should incorporate, allow us to generate on-the-fly or reference all those pre-processed additional documentation elements presently made available from the GESIS ISSP website. There are various ways we may construct the product or set it up as a loosely connected network. The second starting point is that DDI 3.0 in the GROUP and COMPARISON modules delivers a reasonably well specified list of the necessary metadata elements, for most specific situations we do not have to repeat lists that are specified there. This has to be verified, we have seen no real analysis of the match between potential functionality needs and the rather descriptive elements incorporated in the COMPARATIVE module. A useful test is under way in the WP8.2.2 deliverable, and the conclusions there will be important for what elements we incorporate in our analysis. Another very specific test is being carried out by ICPSR.

It is important what kind of data repository we aim to develop at the decentralised node. The present understanding is that the step from DDI 2.0 to 3.0 represents a qualitative difference, is fundamental. But DDI3 is several things, it is both a list of elements and it is an object-oriented services-based architecture, and the qualitative jump lies in the architecture/the implementation. The list is just data, or a gradual scale facilitating more detailed descriptions and more complex organised data. So, if there is no necessary connection between list and architecture, employing only the list as a basis for going from DDI2 to DDI3 could well be viewed as a fine-graded scale in terms of how it supports our functionality needs. We could do 90% of the functionality and settle with an extended version of DDI2, or try to go 100% and try to implement DDI3 in all its aspects. It may be that we could go to v.2.9 and cover most data complexity; however it may also be we have to go to v.3.0 to cover the full versioning potential. If we settle for less than the last fundamental jump, we probably have to introduce some other mechanisms, like more institutional rigour. However, there is a real danger that if we take out some of the value of DDI3 by thinking in these terms, metadata elements will not behave in the same way in the two versions.

Whatever instrument is used to develop the AIP, it needs to be coded in DDI compatible XML. Presently there are no good tools available for production of DDI3 XML from scratch with all the DDI3 capabilities. In the report from Metadata Technology on technology for a Question database, use case 6 argues that this could to some degree be remedied by conversion of DDI2 XML to DDI3 XML. This is the same strategy that has been used in the Dutch DataplUS project.[11]

Nesstar Publisher v4 operates with the possibility of building simple files together in complex collections and also aggregating data and producing cubes. As mentioned on p.10, a detailed mapping of DDI2.1 to DDI3.0 is available as a spreadsheet at http://www.ddialliance.org/DDI/ddi3/mapping-spreadsheet.pdf and as a tree-structure at http://www.ddialliance.org/DDI/ddi3/variable-fields.txt. Further there are very specific translation notes available: http://www.ddialliance.org/DDI/ddi3/translation-instructions.pdf. A relatively easy, implementable solution to generate DDI 3.0 XML, also for the GROUP module, could be the use of Nesstar Publisher v4 or a comparable product, since their file format holds much of the necessary information for writing out the XML. To use and develop functionality based on the COMPARATIVE module is substantially more user application oriented.

This naturally generates questions of a strategic/political character that the CESSDA-PPP should not ignore:

1. Is it worthwhile to tinker with the solutions noted below prior to full DDI3 implementation? (even if it is not significantly influencing architectures/implementations)
2. Do our resources allow us to go for a full DDI3 implementation?
   (Does it really take more resources?)
3. Does DDI 3.0 really deliver all that is needed? Are the Comparative/Group modules well enough developed for our functionality needs?
4. Is it constructive to think this developed in stages?
5. Is writing out DDI3 without a tool to process it on the other side just building bridges into nowhere, and creating data graveyards?

The implications of the repository bank functionalities outlined in the QDB report on grouping and comparability is not included in this analysis.

Concerning technology

The simple view is that in DDI2 we think related to a standard square file: our data are one product and our metadata another, finite, product, where we have some technical possibilities to integrate them into one common product, what we could call an Archival Information Package (AIP) or something like that (the vocabulary of the OAIS reference model). However, in most implementations the DDI standard for all practical purposes is broken down and stored in a variety of database structures on the basis of an explicit metadata model.

In DDI3 we explicitly start by breaking metadata up in its smallest pieces, most of them have an ID, may be manageable and versionable (by an authority/owner). We think in objects and object hierarchies in the individual case, but objects may also be available in groups as in a database table or through referencing each other via URNs into more complex constructs that are maintainable. Our archival product (a documented file/collection) becomes a building that we put together on

[11] http://www.surffoundation.nl/en/projecten/Pages/Dataplus.aspx

request, we select the relevant bricks and put them together. Many bricks may exist in more than one version and not everything is necessary for every purpose, so products may differ. They have the potential of being purpose-specific. And if persistence is required in referencing, it forces us to make published material read-only/non-deletable.

The main justification for this technology progress in DDI is the intention to develop the life-cycle perspective on data (versioning/dynamic data) and to cover more complex data structures: we give data greater possibilities to flexibly live and develop. The complexity of data collections and data dynamics could be regarded as orthogonal dimensions, and the break-up in elementary particles is triggered more frequently by dynamics than by data complexity. To some degree dynamics here may become a problem for the development of simpler solutions for complexity.

The "life-cycle" concept is not always precise. What it usually means institutionally is that data may be re-used for several purposes. That is not very complicated, it gets complicated when such a process fosters data dynamics, when it results in changes, corrections or updates of data that need to be recorded. This could be as:

- Added/corrected metadata, the part that is usually free, searched and only presented;
- Added/corrected data, the part that is restricted, explored and processed;
- More files in a collection, where "files" have to be linked in a systematic scheme.

These points represent several levels/along several dimensions, values, variables, etc.

Our life-cycle concept is partly built on the need for practical changes to the data product (the AIP changes) and partly an institutional re-use ideology, a reuse that only creates problems if it results in updates.

If we build a system for full DDI3 implementation, it will presumably allow greater flexibility and reduced double storage, where the most decisive improvement is that we can version the single elements down to a very detailed level. The general principle of building bottom-up seems to give more flexibility, but still most comparisons are bi-directional maps, comparing schemes. If the bi-directional maps solve our functionality problems we have not seen this analysed and documented. This problem will be covered in more detail by D8.2.

What data complexity do we have (to support)? DDI3 distinguishes between two types: GROUPS, a relatively technical split up in groups that may contain or cover very great complexity, and COMPARATIVE, a somewhat simpler and more substance-based recording/mapping of similarities at different levels of the unit dimension. A GROUP can be comprised of StudyUnits and SubGroups. A standard set of attributes describes the following dimensions for grouping: Time, Instrument, Panel, Geography, Datasets and Language. The setup for the COMPARATIVE module is based on three components: measurement along six different dimensions: universe, concepts, question, variable, category scheme and code scheme, a component that organises these measurements as a relationship between a source and a target in a bi-directional relation, and an actual measure component that is partly text-descriptions of similarities and differences and partly machine-actionable codes.

WP 8 has done a formidable job of working through a formal description of data complexity. We need to be able to handle such data, simply because they are quite common as research data. Here we may focus on three basic classes/types of complex data:

1. Comparative data ➔ data across space / systems / universes (= the unit dimension) Reflected in DDI3 as the UniverseMap or Universe Scheme;

2. Data collected over time, as independent samples or as dependent panels (varieties of adding more variables/attributes). Reflected as several substance maps and also making extensive use of GROUPing. The GROUP module is in most cases discussed as the tool to record such relationships;

3. Micro-macro linkages, a different more technical-oriented problem of linking information at different aggregation or registration levels.

The combination of 1 and 2 is also quite common, well-known examples are the data from the International Social Survey Program (ISSP) or the European Social Survey (ESS). These two use-cases could be distinguished by the indication that ESS is more rigorously organised and that ISSP shows larger internal diversification. Point 3 above is not very explicitly related to DDI, which is focused on documentation as description of data. Micro-macro linkage is more of a user-responsibility as an application question.

How does the GROUP module in DDI3 actually work?

The grouping structure consists of several hierarchical levels. The Group (top level) contains common metadata which are inherited down the hierarchy of the grouping structure. Inheritance is a benefit of using XML. Subgroups can be created at one or more lower levels. For ISSP we could first have module and next time-point. Finally a Study Unit represents a single study on which all the lower-level modules depend, in ISSP and similar comparative data collections we would have several country Study Units at the lowest level.

Groups and study units **both** contain a cluster of modules which describe a collection, and the processes of developing the metadata and data content. These are 'Concept', 'DataCollection', 'LogicalProduct', 'PhysicalDataProduct' and 'PhysicalDataInstance'. The concept of inheritance means that classes of specific information always, and at any level, inherit from their ancestor classes. The specified metadata at the top of the hierarchy is valid for all studies in this group. If information is not valid for a member of the group, on a lower level the local mechanism overrides, allowing this information to be replaced.

The purpose of groups is described using the attributes which summarise relationships using dimensions of time, panel, geography, instrument and language. These attributes allow the purpose to be machine-actionable, while the group also includes an element for describing the purpose in human-readable format. For example, **TimeGroupCodeType** indicates how all members of the group are related along the dimension of time. All relationships are inferred by the markup author, and should be considered as her/his own interpretation of the data:

Code: T0 - No specified relationship;
Code: T1 - Single Occurrence;
Code: T2 - Multiple Occurrence: Regular Occurrence: Continuing;
Code: T3 - Multiple Occurrence: Regular Occurrence: Limited time;
Code: T4 - Multiple Occurrence: Irregular Occurrence: Continuing;
Code: T5 - Multiple Occurrence: Irregular Occurrence: Limited time.

DataSetGroupCodeType indicates how all members of the group are related in terms of physical data products in relation to data collection efforts:

Code: D0 - No specified relationship;
Code: D1 - Single data file from a data collection;
Code: D2 - Multiple data products from a single data collection;
Code: D3 - Integration of multiple data sets into a single integrated structure;
Code: D4 - Multiple data files originating from different data collections.

Evaluation of this set-up has indicated some pros and cons:

- It is possible to document coherence and variation over time;
- It leads to improved efficiency of the data documentation process;
- However, it is complex to migrate data;
- It is difficult to administer complex grouping structures;
- There is limited flexibility once a standard is defined.

This is not specific to DDI3, it is more of an illustration of the complexity of problems; but it indicates that DDI3 does not really give us a radically different solution, what it gives us is a systematic, serious treatment of the problem within a larger integrated solution, and that it needs a very good flexible interface on top.

So, what are the alternative, available interface technologies?

The present Nesstar version 4 format operates with data instances or projects as file hierarchies, similar to DDI3 but referring back to DDI2 for metadata specification, with relative descriptive metadata as the common, higher level information. Nesstar is used to illustrate alternative procedures here; we could probably have illustrated the same problems with reference to the GESIS tools DSDM or CBE. Nesstar solves the problem of defining a Universe scheme by asking for the file/-subfile key specifications and validating them against the data by linking up files according to these specified keys, as we would do in a database. The outcome is comparable, but more stringent than the DDI3 suggestion, and could presumably be expressed in DDI-XML, but is not as application oriented and flexible and is located differently in the work process. Whatever interface we prefer for doing the work, it is not very difficult to use the DDI3 based metadata specifications at higher levels than the Study Unit. But what could be more interesting to develop in such an interface is GROUPing defined as higher levels variables, i.e. over time defined as a trend. Such a strategy would not limit the flexibility of documentation possibilities in DDI3.

Strategies for writing DDI3 XML require a tool that makes it possible to build both the structure and an XML-writing component. A totally DDI3-based solution will take a long time to develop. DDI3 is a very ambitious project and requires the inclusion of an identifier system in a service-oriented architecture – few people have worked on this, or tried to implement it in practice. Possible solutions include experimenting with the level of portal functionality for complex cases using the Metadata Technologies report solution for the production of XML. Such an intermediate solution for publishing DDI3 XML complex data could be as follows.[12]

[12] The following section has been developed from the report on a CESSDA Question Bank, to illustrate an intermediate strategy for production of the necessary XML, but also with some ideas of how to optimise this.

To illustrate how a repository publication could work, here is an example from a hypothetical survey documented in DDI2 using the Nesstar Publisher, Version 4 as a user interface that allows specification of internal relationships relevant for the GROUP module. We are assuming that we want to publish ISSP as one collection of simple surveys (12 modules, spanning more than 20 years givse several hundred single files). Variable level documentation should include universe, question text and interviewer instructions. Concepts have been captured in the study description.

Aimed at developing CESSDA XML based on the DDI3 model, the metadata are imported into a CESSDA Toolkit and broken into several components:

- One or several Study Unit(s) (docDscr + stdyDscr);
- Parallel Logical Product(s) (dataDscr);
- Variable Schemes (one per file) also holding variable groups (fileDscr);
- Several Category and Code schemes containing categorical variables code & labels (one per categorical variable);
- Question Schemes and Instruction Scheme (likely one per fileDscr);
- Appropriate Concept /and Universe Schemes (depending on how survey and variable level universes and concepts are merged);
- Given that DDI2 does not provide string mechanisms to capture the questionnaire flow, a simple linear Control Structure Scheme can be created to associate the questions with Logical Record (in LogicalProduct, one per file), Physical Data Product (one per file) defining the file characteristics, Physical Date Instance (pointing to the actual data files). These can be ASCII or SPSS, Stata, SAS files. This is where the summary statistics (min, max, mean, frequencies, etc.) are stored;
- If cubes are present in the DDI 2, they will generate various NCubePhysical DataProducts. Various other materials can be generated.

A CESSDA Tool-kit Publisher should then perform some initial integrity test to make sure that enough information is available to comply with the conceptual model requirements. The only required element in DDI2 is the survey title. This is clearly insufficient in a metadata rich environment. The toolkit will also require an agency, survey ID and possibly other metadata elements. These can be extracted from the DDI metadata if available or taken from local application preferences.

At this stage the user has the option of storing the information “as is” in the repository but this would not be taking advantage of the reusability features of the conceptual model. Once the initial metadata have been validated, various optimization steps can take place, including:

- Code and categories used by more than one variable can be merged into a single scheme;
- Questions and Instructions reused by more than one variable can be aggregated;
- Concepts and universes can likewise be aggregated (if applicable);
- Variables used in multiple files could also be aggregated into a common variable scheme and reused by reference.

These metadata import/optimisation/curation procedures should be accompanied by relevant quality assurance procedures (such as metadata reports) to facilitate the process. At any time, the various objects can be saved and uploaded into the repository for storage. Note that all of the above metadata are under the umbrella of a StudyUnit so it remains a coherent package (no loose objects). Once the optimisation and quality assurance processes are completed, the various

metadata elements can be registered and become searchable and retrievable by CESSDA applications. They remain part of the original study but can be searched at the "Bank" level (variables, questions, classifications, etc.) Note that this entire process can potentially be automated or semi-automated through batch processing.

In our two most relevant use cases we could list versions of hierarchies. Actually, we could have them in many versions, thus developing a case for preferring DDI3 to DDI2. The present Nesstar implementation does not have the same reshuffling potential as a full-scale DDI3 version.

ISSP ESS
Role of Gov't Wave (2002, 2004, 2006, 2008)
1985, 1990, 1996, 2006 Countries
Module/Topics Topic
Question Question
Variables Variables

We want to include/combine all relevant descriptive information in a comprehensive package, and we want to develop the functionality a user has access to so that the end product can be analytically investigated. So it is actionable information versus inactive descriptive info. Nesstar presently has:

Project
File (is the actual physical "file" = package/instance)
Study = groups of datasets, with potential for very detailed description.
External resources in a study could be Dublin Core, DDI2, Photos, etc.
Dataset = an actual matrix or set of matrices

- Comparative: the most typical use of this term is when we contrast universes/populations.
- A sample is a collection of individual cases, more or less representative, representing a population or a universe, which is the most common, but not the only one break level.
- Comparisons within universes are what we could probably regard as analytic breakdowns and not conceptually "comparative".

For practical end-use functionality purposes we probably do not need to keep the two types of pre-defined comparative data and post-defined comparable data separate. The big difference is how we construct the matrix, not how we analyse it, the difference here is (to simply select between the selection of predefined data and constructed data, using a procedure in which some data may be post-defined as comparable. We might as well expect users to find the process complex (or difficult), with too much flexibility, if it becomes too complicated to generate analysis-ready data. As data users we analyse by traditional statistical methods, summarising within samples and contrasting aggregate sample figures. Our basic practical need is to bring data into a format like the square matrix below. Both dimensions count, the problem of comparable information partly existing along the unit dimension and partly making variables/attributes comparable (harmonised) or to measure/describe similarities and differences. However, most often the unit information is implicit, we do not bother to describe countries; we take them as default well-defined separate contexts.

Abstract	Comparison	England	Universe 1	Sample 1	File 1	Concept		Concept	
						Question		Question	
						V 1	V 2	V3	V4
		Germany	Universe 2	Sample 2	File 2	V 1	V 2		
		Norway	Universe 3	Sample 3	File 3	V 1	V 2		

We would normally regard universe/context as another (qualitative/nominal) variable in the analytic model. Comparative research may be regarded as an effort to incorporate another important system level variable into an unspecified model, just as in multi-level research.

If we have a comparative problem like the one illustrated above, three different universes give rise to three different samples represented in three files that may be added together, and across them we have (identical?) questions that result in variables. What is needed on the metadata side and what functionalities do usage and users require?

In a DDI2 related setup we may describe separately every line in the matrix, with separate abstracts per line, and in each national language for every line. However, the analytically interesting topic is to relate lines to each other, over columns. To have a summary abstract over all three lines and comparisons of lines pairwise, or for every line against a common standard, we would need an explicit hierarchy where common information is raised one level. Alternatively we may make it one cumulative file, but then we take away one level and have difficulties developing the pairwise comparisons except through analysis.

Are pairwise comparisons necessary, given that we have here files representing the universes directly, not put together from an object-collection of some magnitude? In DDI3 such information, even though it is carrying a lot of descriptive information, is mainly justified by being computer actionable. What do we get from developing the same kind of pairwise descriptions for a file of thirty ESS participants?

DDI2 does not have specified information elements for comparing pairs of lines, as it was developed for the single square file. However, with a file format incorporating a hierarchy, it is possible to develop these kinds of information elements in a relative, descriptive way. Even if it is not possible to refer directly to a source scheme or target scheme at the level of detail in DDI3, it is possible to, for example, regard a common standard as a source- or target-scheme and to compare sources and targets. This could work across all generic maps of DDI3 and open up possibilities to compare universes/samples in terms of sampling, concepts, questions, variables, etc. Beyond establishing and documenting equalities or differences, it would also allow for functionality for empirical harmonisation.

If the lowest level is a cumulative file then the most relevant technology, the comparison and recoding of frequencies or similar, which is not a very complex process, is the essence of analytic use. If the lowest level is country files, it is probably easier to incorporate paired maps; at first glance it seems intuitive to think in terms of deviations from a common standard, at least we are

then avoiding all the permutations. This is similar to the distinction between *comparative* (which has a common standard) and *comparable* (which lacks an explicit common standard)

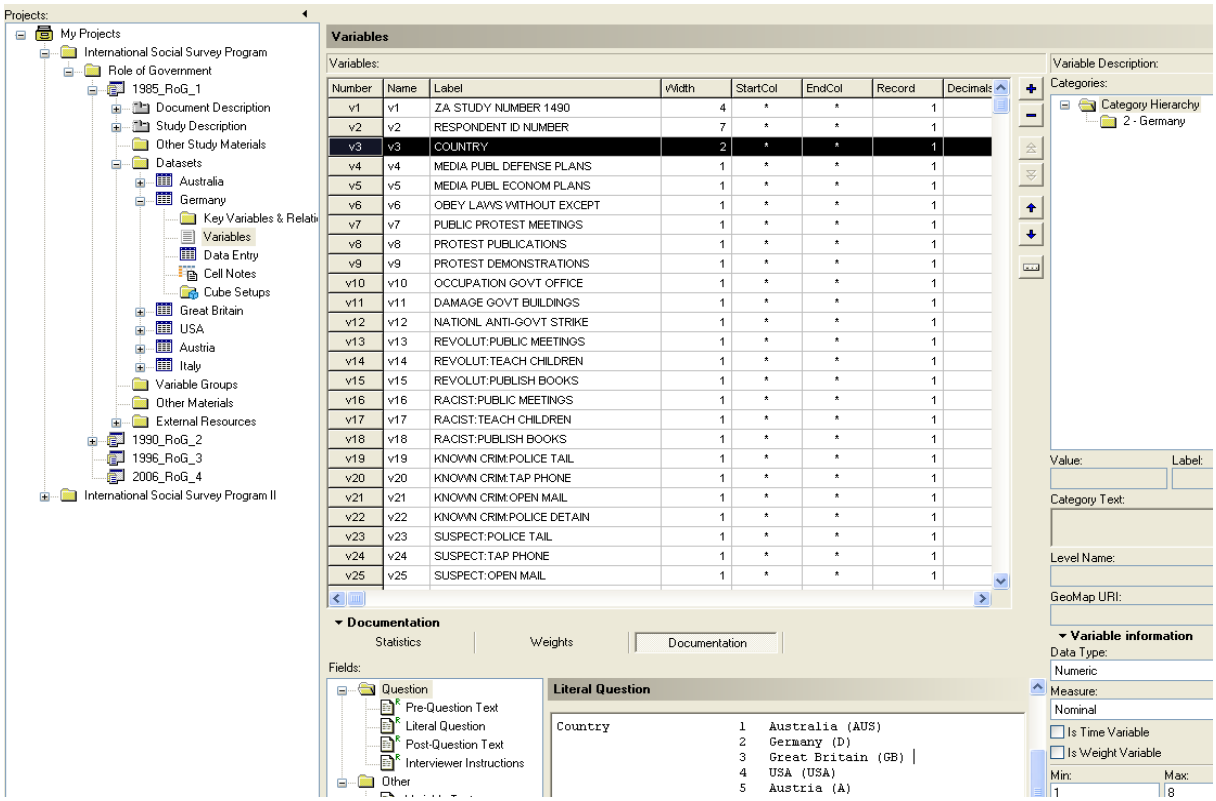
All this requires is that files are standardised. Nesstar presently requires all the languages that are used on the menu; then the Austrian and German 1986 Role of Government could be documented in German and English, while the Australian, British and American files only use English (of course they could also be documented in German!).

Units need to be comparable and attributes carry comparable content or meaning. This is the case for all situations. In DDI3 there are six elements singled out for measurement of comparability: universe, concept, question, category, codes and variables.

We focus on metadata needs to develop the functionality described: locate, explore, analyse. In the scheme above we need metadata that document variables/attributes to establish (degrees of) functional equivalence and grouping possibilities, and descriptions of universes, samples and files (abstract, methodology and technical practical info). From the analyses of WP8 this is specified as metadata needed to document or facilitate the development of:

- Context: the project, the temporal, the spatial;
- Instrument: comparison of measures, variables, questions;
- Data harmonisation (status and procedures);
- Discovery-related substance.

A DDI2-based solution with ISSP Role of Government 1986 as its use case looks like this:



The COMPARATIVE module specifies how to record comparability in social science data. Our concern here is how we go about actually doing it. The example above illustrates documentation at a project level (ISSP), module-level, wave (time) level and dataset (single country) level, in

addition to the dataset-internals. Concept as a measurement scale could be used at several levels, universe is at a StudyUnit level, question, variable, category and code are 'dataset-internals'. The conclusion so far is that it is possible to include explicit elements to describe deviations from a common standard procedure measured in terms of universe or sample, concepts, question, category, codes and variables, but these metadata have to be distributed around the whole metadata setup in the appropriate location with some allowance for an explicit hierarchy. It would be possible to use the structuring power of the six generic maps drawn up by DDI3, but documentation of differences should be more closely linked to the general question or variable metadata. That would then make it possible to generate or write out the XML of the Comparative module if DDI3-compatible XML is required.

Time: The practicalities of the GROUP problem.

In a research dimension this requires the linking of observations in higher order "variables", like trends, changes, differences, etc. In the example below, we could specify where sample 1 and 4 are drawn from, i.e. they represent the same universe, but:

a) we also need to be able to compare methodologies because samples may be drawn differently, etc;
b) we cannot simply match the files since samples are independent - we can only create higher level aggregates at some break level, i.e. at sample level.

T1					**T2**				
File 1	Sample 1	Universe 1	V1	V2	File 4	Sample 4	Universe 1	V1	V2
					File 5	Sample 5	Universe 2	V1	V2
File 2	Sample 2	Universe 3	V1	V2	File 6	Sample 6	Universe 3	V1	V2
File 3	Sample 3	Universe 4	V1	V2	File 7	Sample 7	Universe 4	V1	V2

We may analyse differences between universes, at timepoints (within columns). We may analyse change, development, differences within and between universes (relationships over time). We may, of course, have more than two timepoints.

Relationships over time should preferably be established as a potential, i.e. since it will be somewhere between an analytic result and a data point, and since for this type of data (ordinary

cross-sectional samples) it is not at individual level but at some variety of aggregate level, it is difficult to do this as pre-processing. If we have a variable measured three times, we may define it as a trend. At sample level we could establish the development in satisfaction with life in Germany and Austria and study the differences between Germans and Austrians, but certainly there will be questions about other relevant comparisons, like differences between women and men, young and old, etc, which are other aggregations over elementary units. The intelligence of the file has to be the procedure or map that states that V1/T1 is repeated as V1/T2 and again as V1/T3, and we have to store some convenient lowest level file, at analysis unit level. There will always be a question of the convenience of a procedure versus pre-processed data, and a procedure needs data.

DDI3 uses the GROUP module or the GROUP module in connection with the COMPARATIVE module to document such intelligence. With a hierarchical file construct we do not have to develop specific definitions. We need meaningful grouping substance and grouping technology. Much of this ordinarily belongs to the analytical work which is not necessarily part of an integrated portal solution but rather belongs to the archival cleaning and preparation process. However, the portal should allow/facilitate/support user needs. This indicates that establishing much of this information has to follow traditional archival work, and we are looking for tools that allow reasonable computer actionability. The need to work with weights all the time in most of such comparative analytic work underpins potential supply.

Country / Year	**'85**	**'90**	**'96**	**'06**
Australia	p	p	p	p
Austria	p			
Bulgaria			p	
Canada			p	p
Croatia				p
Germany	p	p	p	p
Great Britain	p	p	p	p
Hungary		p	p	p
Japan			p	p

“Real life” is illustrated as the table above, the intervals are 5, 6 and 10 years.

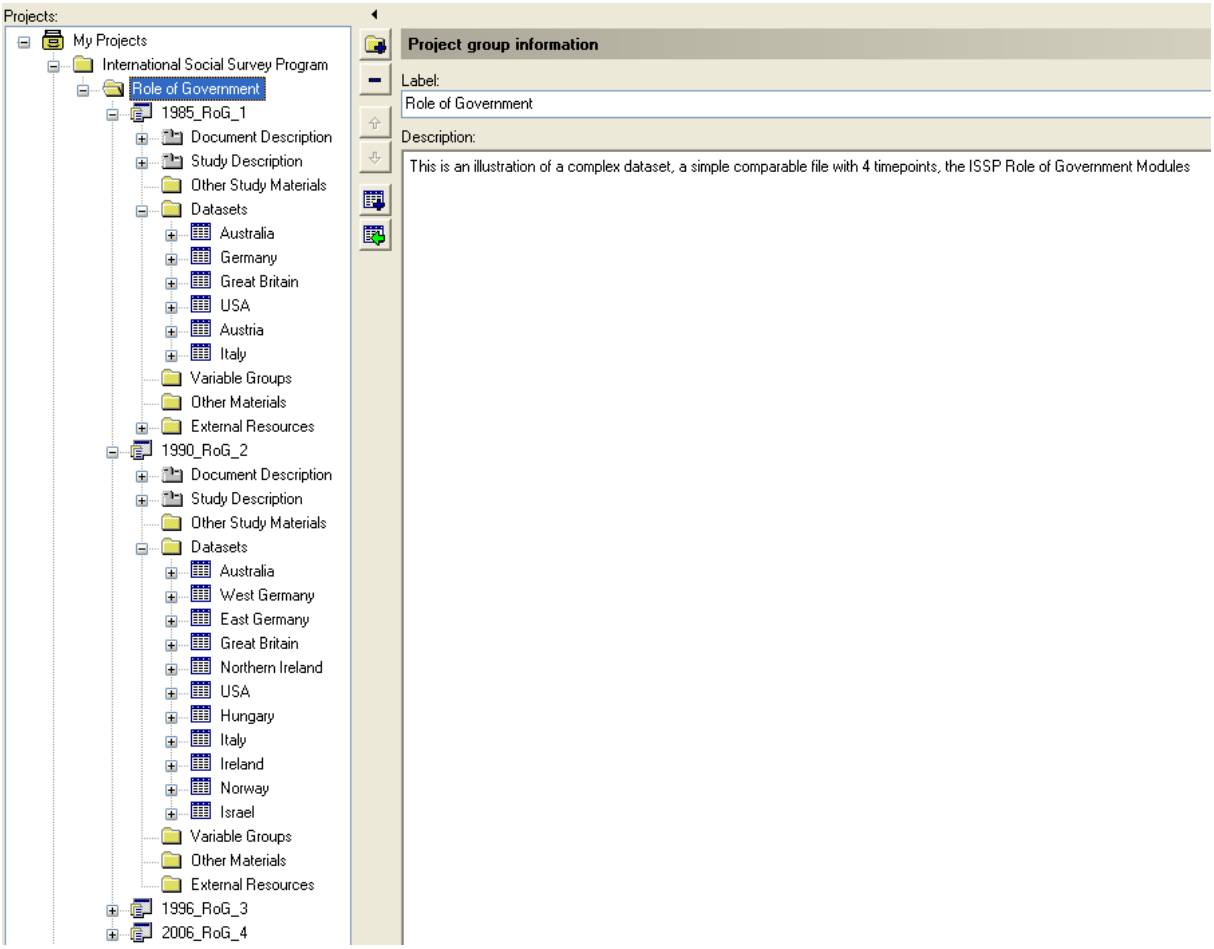
In research based on time variables, the time-interval, and the distance between point measures in time-units, could have some significance. However, this is rarely the case based on sampled or unit record data.

The nice illustration on the former page may fool us into thinking that it is easy to integrate different samples from the same universe. It is not that easy. For these types of analysis it is more relevant to view this as follows:

85	File 1	Sample 1	Australia	V1	V2
	File 2	Sample 2	Austria	V1	V2
	File 3	Sample 3	Germany	V1	V2
	File 4	Sample 4	Great Britain	V1	V2

90	File 5	Sample 5	Australia	V1	V2
	File 6	Sample 6	Germany	V1	V2
	File 7	Sample 7	Great Britain	V1	V2
	File 8	Sample 8	Hungary	V1	V2
96	File 9	Sample 9	Australia	V1	V2
	File 10	Sample 10	Bulgaria	V1	V2
	File 11	Sample 11	Canada	V1	V2
06					

Since non-panel data cannot be match-merged on individual units, this visualisation would be more relevant as the use-oriented retrieved data has to be presented this way. Potential depends more on the variable dimension, this is the typical "trend file" where we will have added difficulties if variables are not comparable, only the comparable variables could be selected. This can be visualised in a multi-dimensional DDI2 setup as follows:



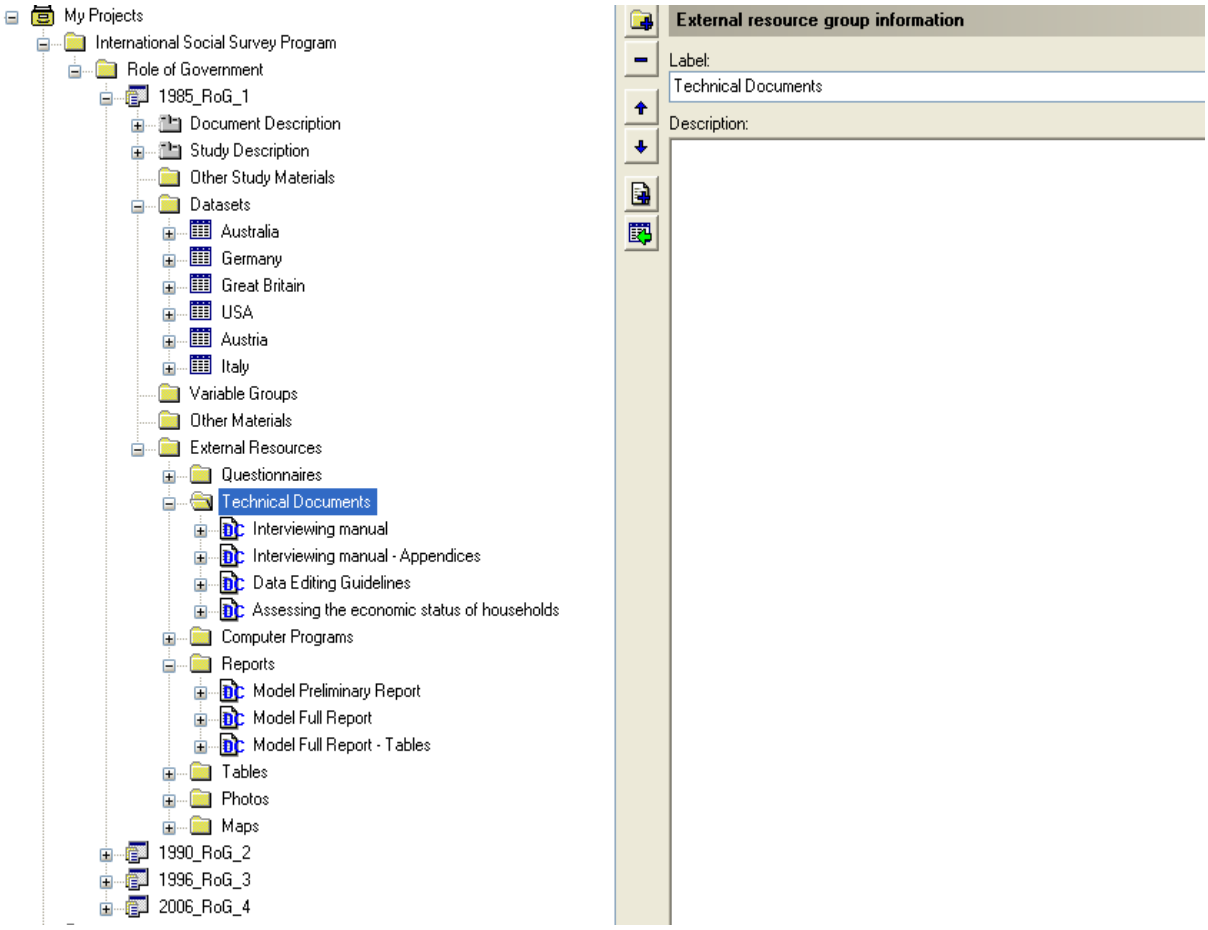
Neither DDI2 nor DDI3 have any particularly simple solution to the problems posed by such data. These datasets/matrices are independent of each other and data can only be compared or put together into trends at some break or aggregation level. The most convenient tool or user functionality for data harmonisation, data exploration and maybe also delivery to analytical procedures would be a tool that allowed the selection of matrices (USA at T1, USA at T2) and the

development of summaries, aggregates, tabulations or cubes that would allow hinging of data at some kind of aggregated level. In practice this would require the generation of tabulations/cubes with data from more than one sample as "hinging" of cubes.

To illustrate, in the setup above, it would be possible to take Australia at T1 and T2, explore and select comparable variables (e.g. "Gender" and "Satisfaction with life") and generate a trend for men and women respectively from T1 to T2 as a simple cube. At the Role-of-Government-level it would be possible to operate with trends or other versions of variable relationships as pre-defined higher level variables.

In the setup above a "study" is defined as a collection of material, where datasets are one type of material, other supplementary types may be text-documents documented in Dublin Core, pictures, etc. A trend would be a "super" variable at group level, a variable defined with reference to variables in separate datasets.

It is not very difficult to fill this up with supplementary material. It is a bit unclear what is supplementary *study* material, what is supplementary *dataset* material and what are *external resources*. However, if we are able to publish such a collection of material, it would probably do for most users, at least if we could wrap it up in technology that makes it easy to view, print, etc.



What about discovery and retrieval? What strategic consequences does it have? Related to the four main types of metadata required by the WP8 analysis, we do reasonably well on context, we

need more information on instrument, and we have the same potential on harmonization and probably also on discovery related substance.

However, with such files stored in our repository, what potential do we have to process metadata and build an index and what kind of portal functionality is it possible to support? So far general conclusions seems to indicate that a service-oriented solution may be more elegant and offer better integration of applications on the producer side, but it is more of an open question regarding potential rewards in terms of user functionality for access and exploration. For such data it will be extremely rewarding when we are able to develop good applications using the full potential of the COMPARATIVE module.

To develop 'super'-variables would cover the same ground as the GROUP module in DDI3, but would probably be a bit more dependent on manual user action to realise functionalities. This potential should be explored a bit further and has most likely been considered by the TIC under development of DDI3 - the chopping up of variable relationships into bi-polar pairs is probably a well-investigated decision.

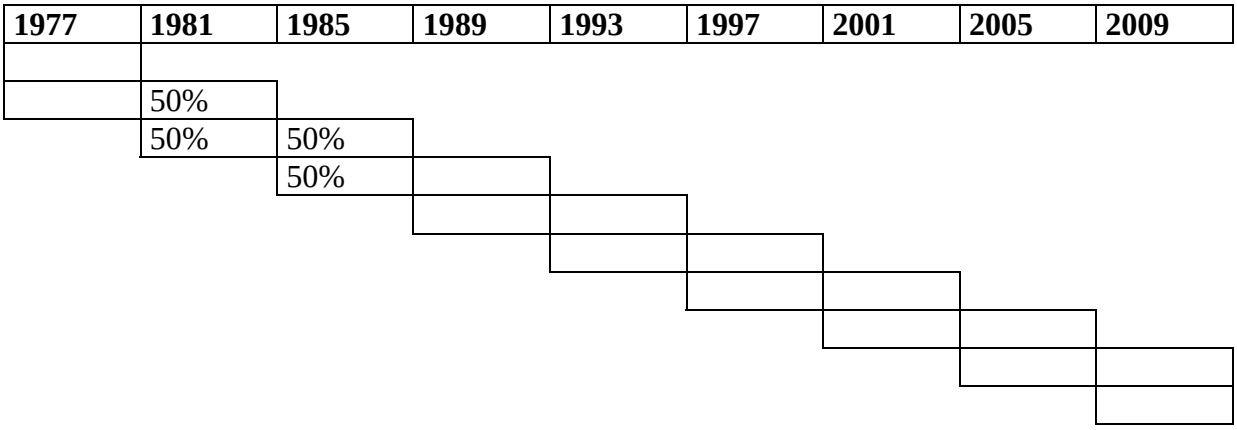
To employ a preliminary solution based on single micro-files linked into a hierarchical system like the one described, gives potential for the generation of DDI3 compatible XML as transport files or for storage in a repository. It would also facilitate the use of alternative DDI3-based end-user-tools.

A panel focused more explicitly on time-dimension:

T1					**T2**				
File 1	Sample 1	Universe 1	V1	V2	File 4	Sample 1	Universe 1	V1	V2
					File 5	Sample 4	Universe 2	V1	V2
File 2	Sample 2	Universe 3	V1	V2	File 6	Sample 2	Universe 3	V1	V2
File 3	Sample 3	Universe 4	V1	V2	File 7	Sample 3	Universe 4	V1	V2

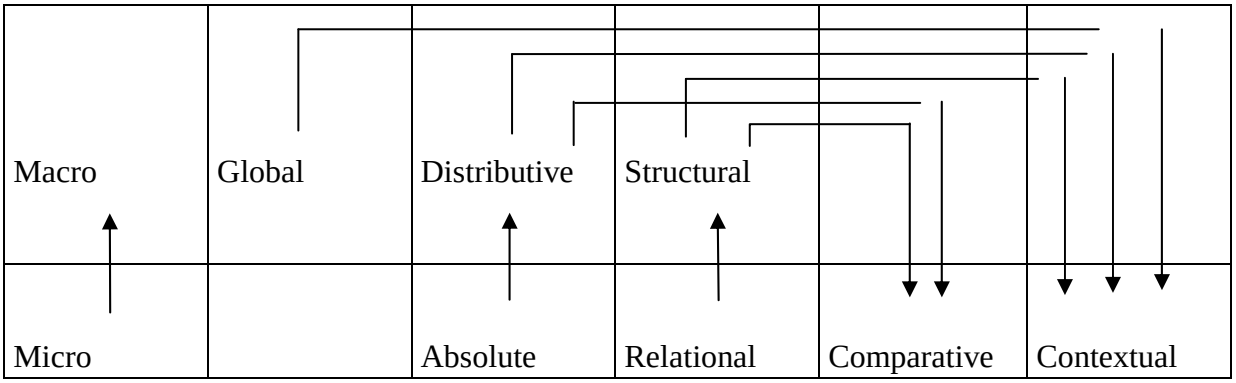
This panel means that we get our time-based variables, trends, changes, etc. brought down to an individual level, giving less representation problems when analysing the time/trend dimension for sub-groupings. We may actually match-merge files at an individual level. However, in practice this is too simple a picture. Panel files often represent varieties on renewal of sample.

The Norwegian Election Studies are quite typical with 50% renewal between each wave.



In European Labour Force Surveys we generally have panels surveyed 6-8 times. Selective non-response or non-availability gives slight initial representation problems, and could represent a significant weighting problem due to treatment of non-response over time. However, most of the practical methodological problems stem from mixing the longitudinal and cross-sectional uses of such data collections, not holding the weighting problems apart. Such data make possible the analysis of cross-sections and individuals, over time. In a DDI2-based setup, introduction of a trend-concept represents the same amount of work for a user documenting data and the same potential for functionality development as the DDI3 GROUP concept.

With a micro-macro-axis ranging from individuals to countries, data at higher levels may be generated in different ways. We sum variables to distributions or calculate averages on the basis of sums, for example.



A microdata file is usually stored in a rectangular matrix, units by variables. This is a format that is very much influenced by dominant statistical analysis technology, where data are intended for further use by standard statistical analysis programs. We may aggregate or calculate values for aggregate levels, e.g. geographic areas and add these as contextual data to original micro units in such a file. It is of substantive interest to be able to study social gravitation forces, etc.

In contrast to the micro file, aggregate data are quite often presented as tables, variable by variable, as a final product. In a computer this could be efficiently treated as a multidimensional cube, but it is neither straightforward to put such a table into a statistical analysis program, nor to link it with other files.

A trend based on micro data is a cube, sample x time x measure, appropriately weighted. Most data at context level are usually summaries or results from mathematical/analytical manipulations, strongly dependent upon which kind of variable scale we have at a micro level.

A data portal would benefit from technology for the exploration of datafiles to establish possibilities for merging or integration of files into multi-level structures. Basically, this means that when users are exploring data they are given possibilities to:

- control what identifiers are available on files;
- "unfold" tabular data to rectangular matrices if there is any identifiable "unit of analysis" in the table.

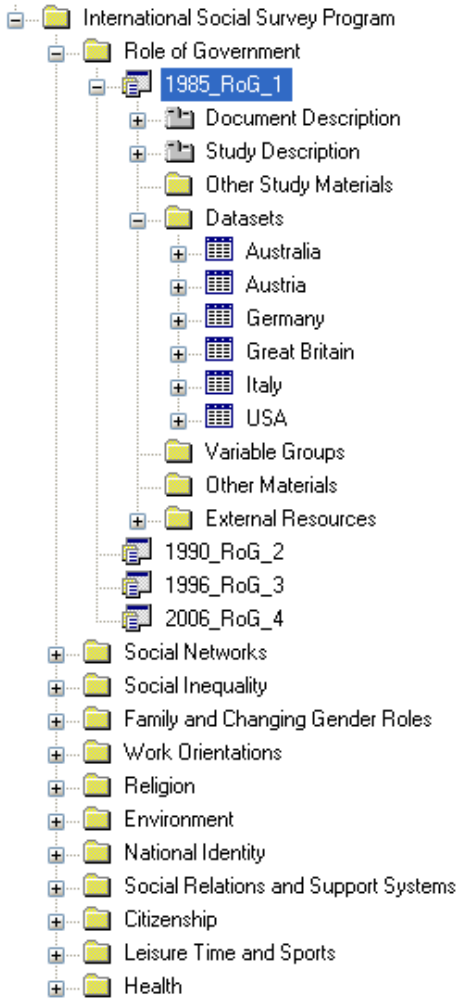
If there are two different types of data-files, for example a survey file resting in one server and a context type file resting in another server, it would be very convenient to be able to match-merge files on the fly into a new multilevel structure:

- if identifiers allow, to match-merge files into multi-level structures;
- before downloading data to preferred format.

This kind of problem appears not to have been discussed in DDI3 related documents. Probably it has not been identified as a specific problem; for most practical purposes it can be reduced to the merging of different types of data files.

Exploring hierarchical set-ups of file collections

A comparative file means accumulation on the unit dimension:



This illustrates a strictly comparative file, i.e. an ex.ante example.

Instance (project) ISSP
Module/topic Role of Government
Time/wave 1985

Levels?
Datasets Australia, Austria, …
Sections Are in Variable Groups
Questions
Variables
Options/values

The Comparative module:
Measures comparability on 6 dimensions, with potential for later dynamic construction of relationships (for users)

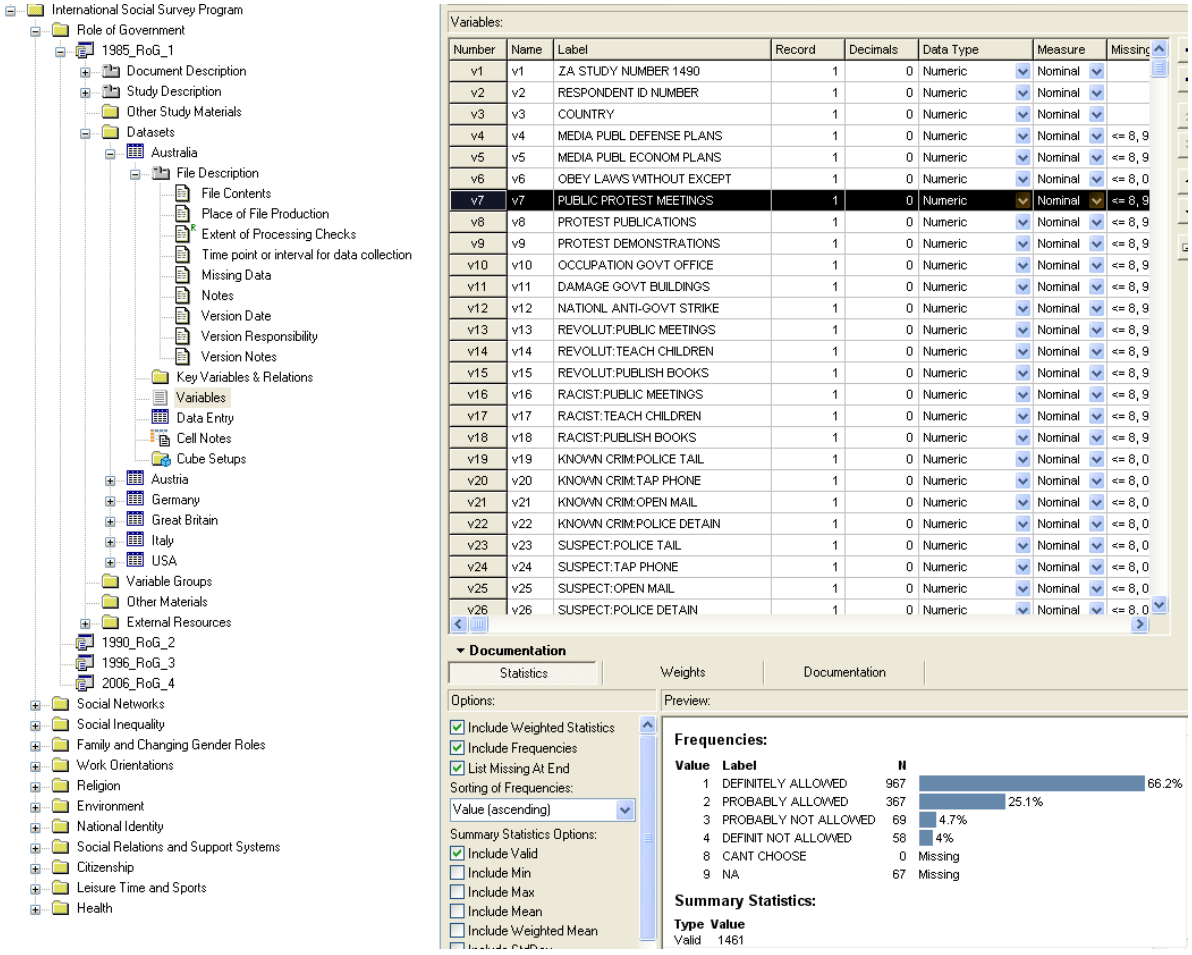
Universe, Concept, Question, Variable, Category, Code — Specifies a relation-ship — descriptions or values

The comparative module in DDI3.0 does not always specify the same type of relationship. For the universe dimension, it is probably most interesting to compare datasets (countries in this case) to each other in a descriptive manner, and to assess to what degree they are comparable. There is one universe per dataset. A description of each universe as a relationship to another universe is methodological information for an analytical unit and has some direct interest for analysis. This is a reduction of its value, in the web-services set-up it is intended to tell software something of machine-actionable value. However this function is redundant here as the software no longer needs this information. Instead it becomes only ordinary documentation and most users would question its value given the work needed to record it. What we need, to be able to undertake comparative analysis, is ordinary methodological documentation, these bi-directional comparisons are requested by many persons. On the other hand, this interface offers one of the few possible ways of generating the description. It will not be easy to automate the process.

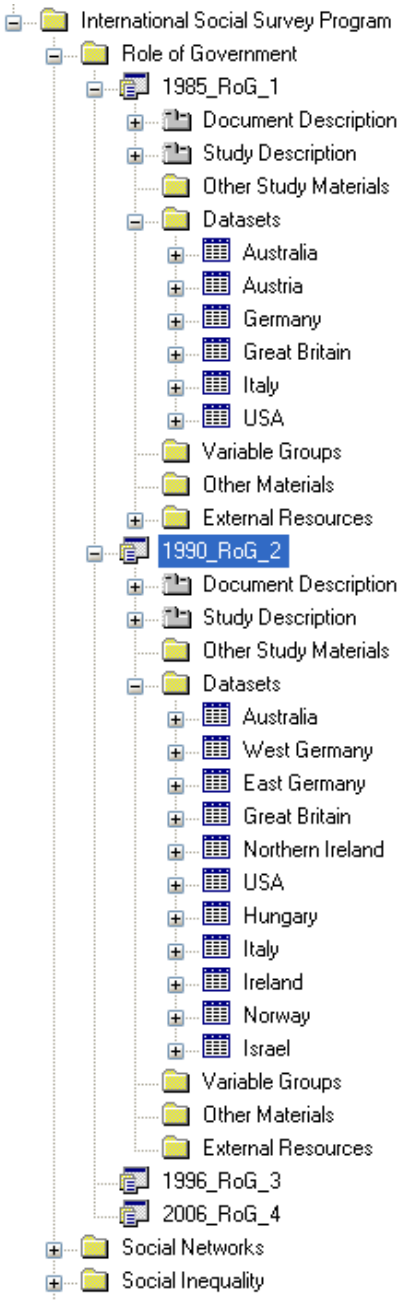
The five others are more or less specific measurement of comparability on substantive dimensions. Some of these could be compared by computers, but would in practice require some layers in-between, like a multilingual thesaurus. For concepts and questions the most interesting relationship

could also be between every single dataset and a common standard questionnaire, it is not necessarily between the single country questionnaires. At least it is a somewhat more complex relationship, because usually there will be a common denominator defined at a higher level, a common questionnaire for the module/wave combination. Here we usually find that the common denominator is translated into a mother tongue, other adjustments are also possible for national situations - the simplest could be to leave out some questions, or to add some national deviations.

A valuable piece of technology for practical work would be the ability to load the six national datasets of the 1985 Role of Government module into six different windows, thus allowing immediate visual inspection and comparison of any metadata element.



If we introduce time in addition to space:



	T1	T2	T3	T4
Australia	x	x	x	x
Austria	x			
Bulgaria			x	
Canada			x	x
Chile				x
Cyprus			x	
Czech Republic			x	x
Finland				x
France			x	
West Germany	x	x	x	x
East Germany		x	x	
Great Britain	x	x	x	x
Northern Ireland		x		
Hungary		x	x	x
Italy	x	x	x	
Ireland		x	x	x
Israel		x	x	x
Japan			x	x
Latvia			x	x
Norway		x	x	x
USA	x	x	x	x

With such a setup we have the ability to analyse changes between countries, differences both in space and time. However we can only study trends at some aggregate level.

The DDIGroup module group study units are separated into time, instruments, panel, geography, datasets or language, all of which are shown above as cells, rows or columns.

At the module level, for example at the Role of Government level, it would be convenient to have a key file/table, wave x variable, i.e. lower level x variable. Such a technique could work very efficiently for long trends, such as for Eurobarometers.

Related to the present Nesstar V4 file system we need three extensions:

1. The ability to open lowest level files in separate windows, this is possible in some GESIS software;
2. At higher levels, to define key tables as described above;
3. Standardisation technology.

Publishing complex datasets in an enhanced CESSDA infrastructure

We have so far discussed three types of problem:

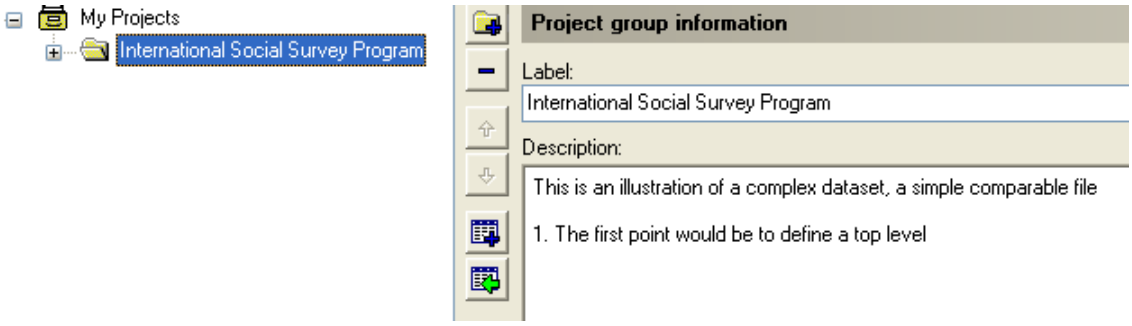
1. Repetitions of data collections over time;
2. Comparative, over space collections;
3. Linking of microdata and macrodata.

These problems are very different; basically it is only the first issue that represents a unique documentation problem. The remaining points represent more of an application based on a thoroughly documented collection or set of data.

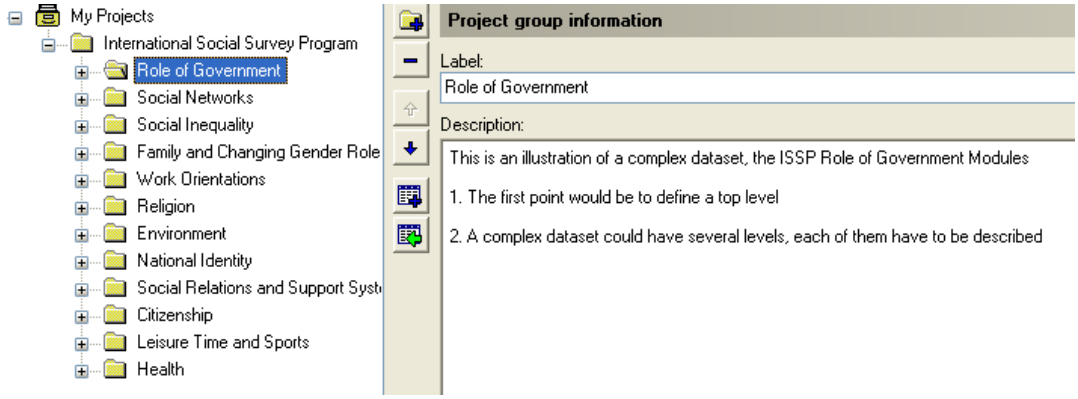
Issues relating to point 1:

1. data being documented as a collection before ingest in a repository, based on a file concept following DDI2, as a hierarchical system of files;
2. data being documented (as a collection) before ingest in a repository, based on a DDI3 conceptualisation, a functional module-based process;
3. a process where data documentation only concerns single files, but where any connection between files, as groups or comparisons, are functionalities in applications built on top of a registry and other services.

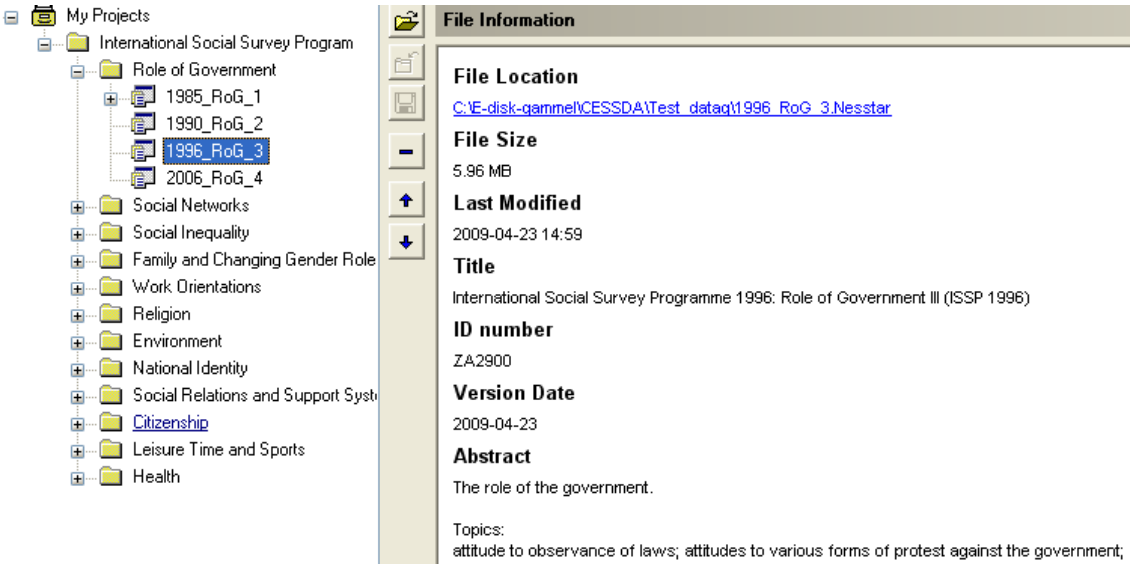
Point 1 is simplest to understand: we have to proceed level by level and define and document each level in the hierarchy with all the appropriate information for each level. For International Social Survey Programme (ISSP) we could start by simply defining that as our collection level:

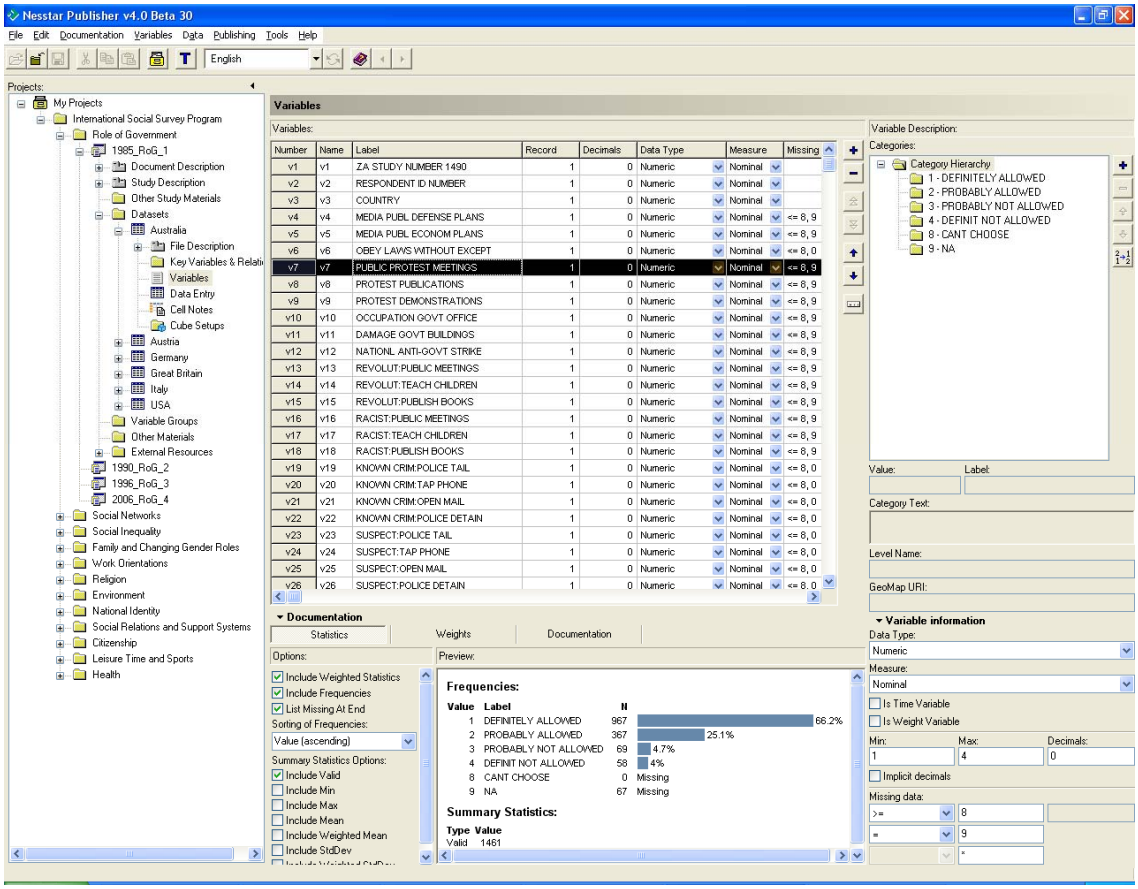


Next, we need to define what can be regarded as the next logical step. In the case of ISSP this could be the module level, since ISSP is conducted as modules repeated at intervals. Subsequently, we would arrive at the question of how general the visualisation possibilities should be, and whether or not we need to contrast module abstracts.



It is not completely obvious what the hierarchical relationship is between the components involved, however we illustrate time as the third level. The important point here is that this becomes a defined hierarchy where it will later be difficult to shuffle the structure around. However, as the individual datasets are the basic units being explored, this is not necessarily a problem. Here it would more likely be of value to be able to contrast metadata directly (for example, abstract) at time-point level.





Usually the single StudyUnit is the lowest level in the hierarchy.

A technical question which has not been raised here is how to update metadata in a collection being explored.

Summary

We presently have tools that can document data as described in the illustrations above. These tools or any other tool need to interact with:

1. DDI-defined controlled vocabularies; Nesstar Publisher does this via the CESSDA template, which is defined for DDI2. DDI3 has a general solution named a DDI profile. To expand the DDI2 template to a DDI3 profile requires an expanded metadata object-model;
2. The ELSST Thesaurus, for multilingual concepts and keywords. This requires a general product interface and *clarification of the IP rights*;
3. In a pure documentation procedure, interfacing with a QDB could mean more rational and standardised work as well as risks associated with consistency and correct wording. However, a QDB-tool is likely to be employed in instrument development at an earlier point in the process;
4. While a QDB-tool points toward the earlier stages of the process, a harmonisation and standards database points toward user scenarios, and a somewhat more dynamic use.

The issues which we have tried to explore here can be summarised as follows:

1. It is easier to develop immediate tools for complex data if we think in terms of complex hierarchical files than in terms of modularised XML-based frameworks;
2. Both solutions can be used to generate DDI3-compatible XML in a storage and transport format, this is not a 'burning bridges' tactic, initially it is more like 'How do we interface with the problems?';
3. However, it may be more complicated to come up with good solutions for the versioning problem. The versioning problem implies persistency in what is stored in a repository, and it requires that the object identification system is implemented. We somehow have to think in terms of versions of the single objects in the object model that have to be developed: this requires a comprehensive identification system that also contains a version number;
4. Whatever software is developed for "intermediate" solutions on the way toward full DDI3 web-service use could hamper development, but certainly runs the risk of being left behind and becoming the victim of development. Such software will be expected to skip the identification solution set up in favour of a full service-oriented solution;
5. It is not easy to run a 'DDI3-only' strategy. So far nobody has done much practical work on the really difficult topics and the issues that are important to European researchers but applications making full use of the GROUP and COMPARATIVE modules would be of great immediate value to researchers exploring collections.

Any version of DDI 3.0+ needs an instrument/interface to produce meaningful code, and to use a general XML-editor is quite difficult and not suitable for standardising work across a European arena. The DDI 3.0 XML-code does not come by itself and some of it is extremely complicated. Sooner or later in the work process we have to define the relationships, the groupings, the mappings, etc. However, with good software some of this can be automated.

With a hierarchical file system as a potential ingredient, the relationship between components becomes pre-defined and much of the job is done when we read the file(s). This applies more to group than comparison and does not cover the ideas about development of higher level variables; groups become technically linked where comparisons are substantive based across many dimensions.