Title	**A CESSDA Enhanced Data Portal (D5.7)**
Work Package	WP5
Authors	Atle Alvheim (NSD)
Source	Summary of the Bergen Workshop May 2008, Cologne Workshop October 2008, Essex workshop August 2009 and report from Metadata Technology: Technical Specification for a European Question Data Bank
Date	August 2009
Dissemination Level	PU (Public)

Summary/abstract

The motivation for developing an enhanced CESSDA data portal is:

a) Compared to the present CESSDA portal there is a need to handle more complex collections of data;
b) It is necessary to integrate two additional services, a Question Database (QDB) and a Concepts, Conversions and Classifications Database (3CDB) in the total infrastructure that is also incorporating the portal services;
c) The Portal should incorporate possibilities to update and version data.

These requirements can be met in an infrastructure-/portal solution that is based on implementation of version 3 of the Data Documentation Initiative (DDI) metadata standard

Following closely from a decision to base the work on DDI 3, a Service-oriented architectural setup (SOA) for the portal is recommended.

Social science data may potentially be quite complex in its data model. Further, a user oriented data portal has to deliver data for analytical purposes. This indicates that complex data has to be simplified to meet the analytic methodology and technology. In addition, technologies for comparison of data and the comparative research perspective as such will be important to social scientific data use in Europe.

The data provision layer of the portal is built on data resources stored in a decentralised system of (national) data repositories.

Several free-standing additional resources are incorporated into the scheme as part of the documentation / ingest activity and on the discovery / presentation side. Controlled vocabularies and classifications are important as extensions to the metadata standard, the most important such resource is a multilingual thesaurus that allows structuring of information and eases communication across the many languages of the European Research Area (ERA).

A Data Portal: Some general background

The CESSDA portal is primarily intended to provide access to research data. The typical simple social science data file consists of a data (content) matrix and a more or less sophisticated metadata part, sometimes integrated with the data matrix, sometimes organized separately. Metadata plays many roles, from recording and communicating the substantive content of data to allowing technical developments and facilitates linking with diversified applications like information systems on data availability and use, i.e. expressing the data model. When data grow in complexity they by default generate a need for more sophisticated compatible metadata. In particular the Internet has created new uses for metadata that are transforming the management of information, metadata function as the glue of information systems and for the social sciences the ambitions for the development of metadata systems runs higher than in many other scientific fields because of the linkage with more complex data analytic needs. For general multi-purpose IT-systems the metadata-component becomes the most important component, it is through the metadata we enter and access data and understand both structure and content. In a managed collection of resources intended for scientific research purposes there will be a need for a minimum level and a standardized setup for metadata following the actual data to allow information exchange and integration. However, the complexity of social scientific data should be described in a way that allows constructive analytic use of that complexity.

Social science data has traditionally been stored in a simple rectangular data format, units by variables, caused by and for the purpose of maintaining a short bridge over to the dominant statistical analysis technology. Data organised this way are easy to document and fits well with the codebook view of social science data. The majority of data collection efforts in the social sciences still result in single cross-sectional standalone files. We see many varieties and deviations from this main tendency, but the need to move data over from archive / storage to analytic use generally make us think square. Aggregated data represent one simple step further in this picture, in tabular format it normally represents a data model with an elementary analytic element introduced. We are well capable of handling these two dominant types of files, although we have some problems of user-friendly transport of data back and forth between them. Data archives are evaluated by how well they serve their users and the users' perspective is dominated by analytic needs, not practical storage and documentation needs.

However, we have experienced a gradual sophistication of data models. As soon as we move from single files towards incorporating third dimensions like time and/or space into our data model, this organisation and the usage dominated perspective of data becomes harder to handle. With third dimensions of scientific interest our analytic constructs very often become relationships or relative variables more than absolute measurements, they are often generated as part of the investigation process and require more sophisticated tools and often such measures are difficult to generate in a world of separate square files. The square file is the most limited case, where the only variation is across units. For a complex organized collection of datasets, we often need at least a two-level hierarchy to disentangle common attributes from specific measures recorded more than once over 3 dimensions. As empirical social science has grown in sophistication and the need for actual and high quality data has progressed, we now experience a need for greater flexibility in both use and in expression of analytic potential of the data we collect.

Standardisation of metadata is said to be the key to automation of data lookup and exploration processes. As justification and explanation for the importance of metadata this could be extended much further, but is generally part of a set of justifications why the international data producing and archiving community for a long time have worked to develop metadata standards. There are several potential standards available, distinguished by differences in specific purpose and differences in data models and data types covered. Many of these standards have difficulties with data complexity, data dynamics and open relationships; the multitude of available standards also testify or create the same kind of limited flexibility as mentioned above and automatically ask for crosswalks or more comprehensive common denominators. We have many standards because each is only solving part of the problem and with a complex problem we have to face complex explanatory standards.

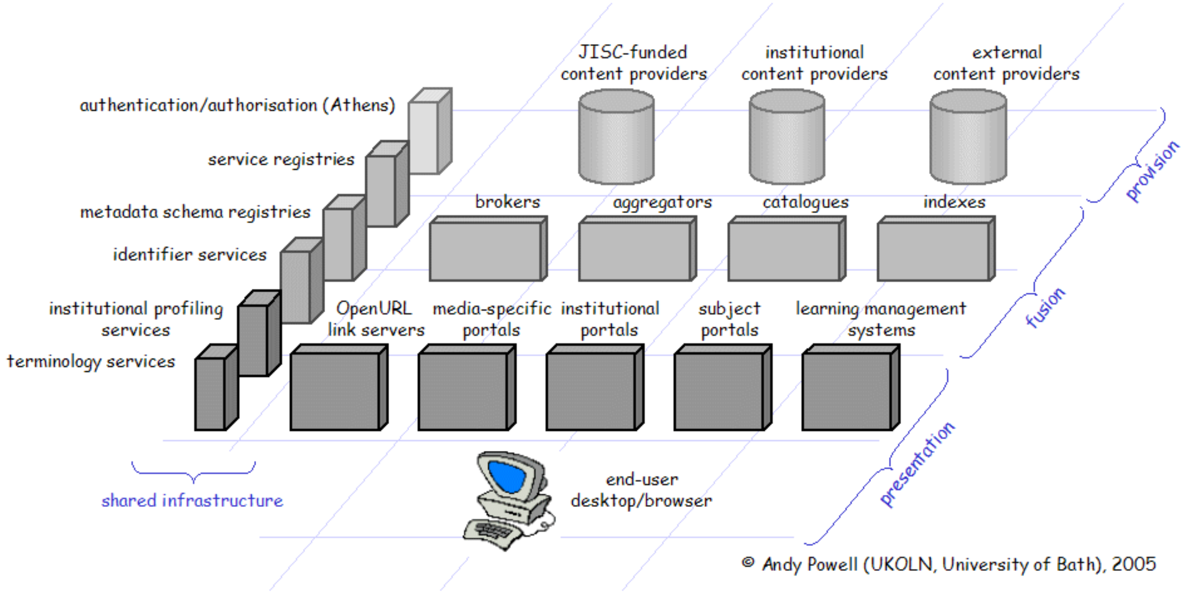
The present CESSDA data portal

In the description of the present CESSDA portal, the JISC IE technical architecture is to some degree used as a background framework to illustrate the relationship between layers and services. The portal work will be focused on three types of work:

- Preparation of the data and characteristics of the local data repositories;
- Organising content of repositories to facilitate development of data location tools;
- Services for data location and data exploration.

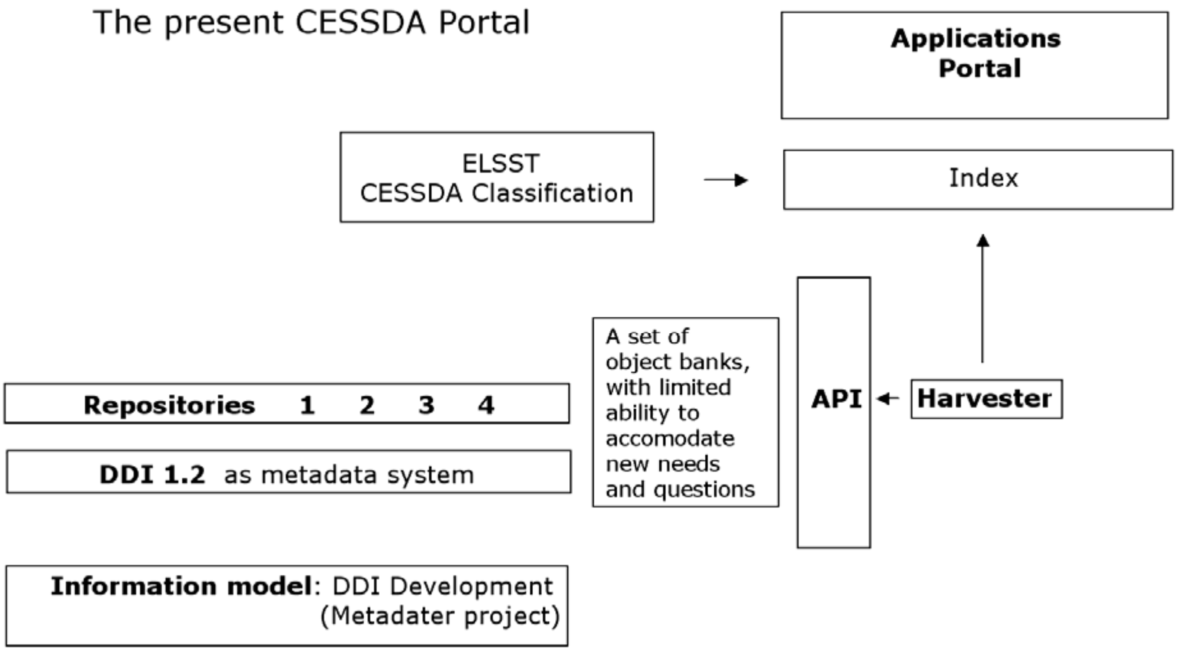
These points may be seen as comparable to the data layers in the illustration.

Fig. 1: the JISC IE portal architecture


JISC-funded content providers
institutional content providers
external content providers
authentication/authorisation (Athens)
service registries
metadata schema registries
identifier services
institutional profiling services
terminology services
brokers
aggregators
catalogues
indexes
OpenURL link servers
media-specific portals
institutional portals
subject portals
learning management systems
provision
fusion
presentation
shared infrastructure
end-user desktop/browser
© Andy Powell (UKOLN, University of Bath), 2005

If we think in terms of this JISC IE architecture diagram above for the present Madiera/CESSDA portal, the CESSDA data *provision* layer is a set of institutional (data archival) content providers, presently without any specific authentication / authorization services attached. These data repositories are filled with data objects/studies (information packages) that are tightly integrated packages of data and metadata. DDI2.x is presently functioning as the de facto metadata standard across, but metadata may also be published to the publicly exposed part of a repository without the actual data, this is one of the means we have to tell about the existence of data without actually exposing the data. The *fusion* layer is represented by a search and browse possibility against a federated common (virtual) catalogue and a parallel common index that is developed from harvested metadata. The *presentation* layer is represented by a common portal solution on top of the metadata indexes, being enriched by some terminology services, the multi-lingual thesaurus ELSST and other varieties of classifications. No clear-cut delineation between fusion and presentation is intended here, and is probably not 100% correctly positioned relative to the JISC specification.

Fig. 2: A tentative visualisation of the present CESSDA portal, the original visualisation of the Madiera portal setup.


The present CESSDA Portal
Applications Portal
ELSST
CESSDA Classification
Index
A set of object banks, with limited ability to accomodate new needs and questions
API
Harvester
Repositories 1 2 3 4
DDI 1.2 as metadata system
Information model: DDI Development (Metadater project)

The present data archive data repository is typically a Nesstar server. These servers are based on an information model following DDI version 1 or 2. DDI 2 as metadata standard is focused on documentation of rectangular files and geared towards great information detail. DDI was a major step forward for documentation of social science data, but the earliest version has become victim to the problems mentioned above; it is restricted to a limited or partial solution, basically only moving the traditional square statistical file to an internet context. It is only Nesstar servers that are supported by the present CESSDA portal; this of course simplifies the metadata harvesting problem. Data, as a rule, is organised with several content-levels of interest, by topic, files, modules/parts of files and questions/variables. Presently data is published to such a Nesstar server in one of two ways, either via the dedicated publishing tool in the Nesstar suite (the Nesstar Publisher), or via home-made database tools that deliver XML-coded information packages of data and metadata that are copied to the Nesstar server. A publishing process via the Nesstar Publisher tool may be supported and standardised through employing a common template accommodating common controlled vocabularies of various kinds, and for this purpose a common CESSDA template has been developed. Home-made database solutions for this task obviously have some greater standardization problems across data publishers. The pros and cons of these two strategies then could be listed as:

The publisher or editor solution: Presently easier to standardise across users and usage through an explicit common template, a standardised tool and a simpler solution to maintain in such a decentralised architecture.

The database solution: A more comprehensive tool that normally is influenced by solutions to several additional problems of a data archive, in particular interfacing with information services and covering curation-related problems.

The present *portal* component is a freestanding software component that is able to harvest metadata of a specific definition from sets of Nesstar servers/data repositories and from that build indexes for search and browse functionality (Lucene search engine do not supply robots or crawlers itself). From the hit lists returned from the data discovery technology (search or browse), the Nesstar Client is employed to load and explore single data files, one at the time. From the Nesstar Client these files may be downloaded to other statistical file formats on users local computer equipment.

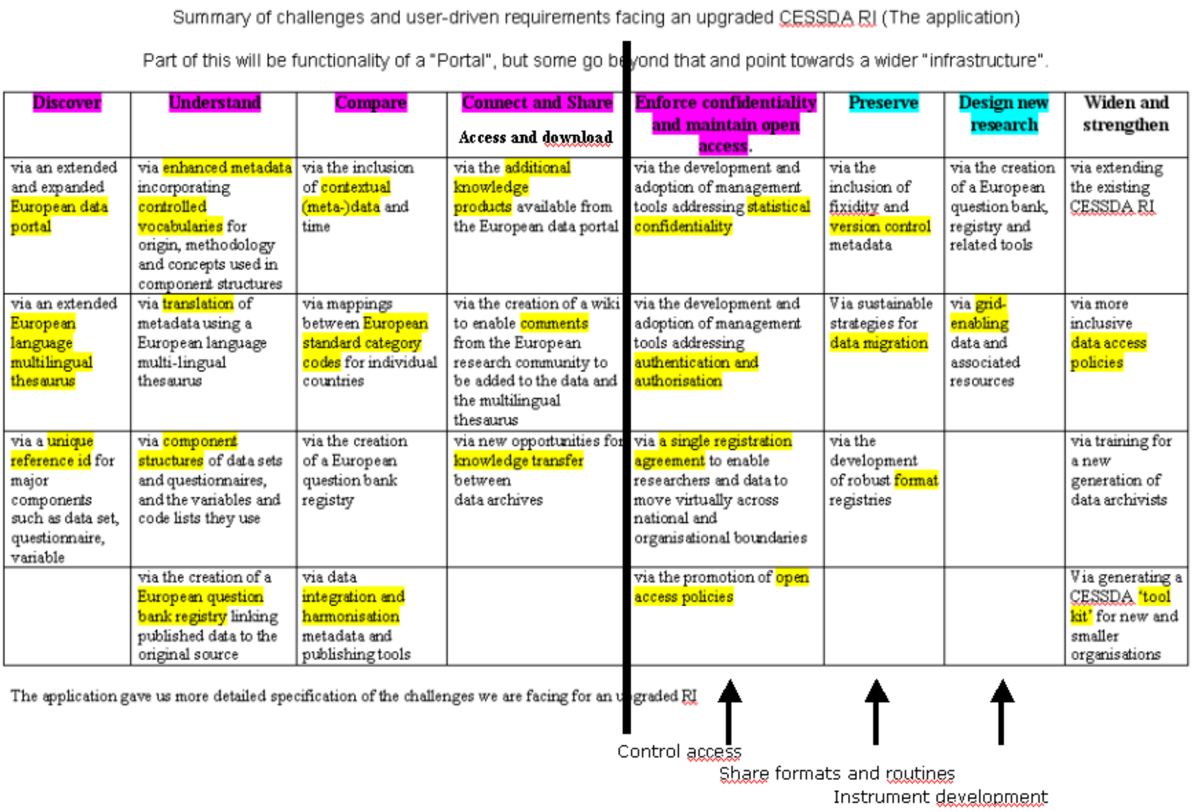
As the present purpose is to illustrate information functionality needs and architectural problems, an actual overlap in terminology with JISC IE is not stressed.

A high-level generic model

The three basic functionalities of a social science data delivery service are to make available means to *find*, *explore* and *deliver* data to analytic processes. The analytic process as such is regarded as being outside the problem area discussed here, but it is difficult to draw a sharp distinction between elementary analysis as part of exploration processes and analysis as the final stage (as *data use*). To *find* requires access to the metadata component, because the data discovery process is basically based on the metadata *description* of the substantive content. *Exploring* requires in addition access to the data component, since the actual distributions on single variables will be an important element of data exploration. To deliver data for further use will also be dependent on ability to load data / carry data along. Consequently, the move from the data discovery functionality to the data exploration functionality is quite *fundamental*, since it triggers a need for controlled access to data, i.e. through an authentication and authorization service. An authentication / authorization / access (AAA) procedure will be activated by crossing the line between metadata and data.

In the application for the CESSDA-PPP the portal was sketched as a somewhat richer collection of functionalities than what is usually associated with a data portal, it is envisioned as imbedded in an integrated infrastructure or web service that announces CESSDA services. This central point is a combination of a web-site announcing CESSDA services and best practises, a data catalogue entry point and as part of that, a related tool-kit for linking up with additional services for linking data repositories horizontally and potential pre-processing of data before delivering the output product to dedicated (analytic) services. Thinking in terms of a tool-kit makes it a bit more complicated to outline in terms of clear-cut layers in the architecture, because it indicates a more dynamic linking of many of the components in the scheme. The strategy in this document is to link the portal architecture somewhat more with the OAIS Reference model. The portal is seen as a data discovery tool in interaction with a provision layer, and as a flexible tool-kit in the fusion layer, linking in a variety of services for the treatment of the resources returned from the data discovery activity. In particular the interaction with a harmonisation database will be of this middleware character. But the need for the tool-kit view also stems from the fact that data returned from the discovery service may be of complex organisation, the comparative nature also indicates that documentation may need to cross language boundaries on its way back to the user.

Fig. 3: The CESSDA PPP application summary of challenges



Summary of challenges and user-driven requirements facing an upgraded CESSDA RI (The application)

Part of this will be functionality of a "Portal", but some go beyond that and point towards a wider "infrastructure".

Discover	Understand	Compare	Connect and Share Access and download	Enforce confidentiality and maintain open access.	Preserve	Design new research	Widen and strengthen
via an extended and expanded European data portal	via enhanced metadata incorporating controlled vocabularies for origin, methodology and concepts used in component structures	via the inclusion of contextual (meta-)data and time	via the additional knowledge products available from the European data portal	via the development and adoption of management tools addressing statistical confidentiality	via the inclusion of fixidity and version control metadata	via the creation of a European question bank, registry and related tools	via extending the existing CESSDA RI
via an extended European language multilingual thesaurus	via translation of metadata using a European language multi-lingual thesaurus	via mappings between European standard category codes for individual countries	via the creation of a wiki to enable comments from the European research community to be added to the data and the multilingual thesaurus	via the development and adoption of management tools addressing authentication and authorisation	Via sustainable strategies for data migration	via grid-enabling data and associated resources	via more inclusive data access policies
via a unique reference id for major components such as data set, questionnaire, variable	via component structures of data sets and questionnaires, and the variables and code lists they use	via the creation of a European question bank registry	via new opportunities for knowledge transfer between data archives	via a single registration agreement to enable researchers and data to move virtually across national and organisational boundaries	via the development of robust format registries		via training for a new generation of data archivists
	via the creation of a European question bank registry linking published data to the original source	via data integration and harmonisation metadata and publishing tools		via the promotion of open access policies			Via generating a CESSDA 'tool kit' for new and smaller organisations

The application gave us more detailed specification of the challenges we are facing for an upgraded RI

In the application there is a table that summarises the functionalities aimed for, the pure portal functionalities of *discover*, *explore* (understand and compare) and connect and share (access and *download*). Of these, already the division between understand and compare may be quite fundamental, a division between going vertically or horizontally within or between data repositories or data instances within repositories. In addition the portal tool kit should have the ability to authenticate and authorize potential users according to a systematic data access policy, i.e. control access and log use, set up formats and routines for data preservation and allow description of data collection instruments so that the overall infrastructure may be instrumental to promote new research through new data collection, ease data harmonization problems across items and allow more comparative research. A further extension to this could be easy interfacing with modern data collecting tools over the web and delivery of such data in generic and well documented formats into statistical analysis packages. In addition to this it is possible to envision a closer relationship between the pure *data* exploration process and potential knowledge products (*e-prints*) based on or otherwise related to the data. In the CESSDA view e-prints are sub-ordinate, not parallel to data resources, the ordinary library is turned upside down, with data as the primary resource that give rise to knowledge products.

As already pointed out, the CESSDA portal is not preoccupied with the actual *use* of the data, at least not if such use has no further consequences for the data and through that other prospective users. However it will always be a question of how much functionality of *preparation for use* needs to be within the portal tools and how much should be pushed out to the user's local environment. And it might well be that a successful CESSDA portal manages to establish data collection, data preparation and data enrichment as a meriting activity to a level where it creates a stronger need for functionality to publish outcomes of data use back into the data object itself.

The application contrasted the stages of the research process against a set of potentially problem-solving tools and resources, as components of a total research infrastructure:

Fig 4. A process view

Process	Resources			Tool
Conceptualisation	Theory	Former research		Brain
Data collection		Question-bank	←	Instrument
Documentation	Thesaurus	Controlled vocab	Template	Publisher
Archiving	Archival system and information system			
Storage				Server
Location	Thesaurus	Harvester	Index	Portal
Understand	High quality metadata		Translation	
Authentication	SSO →	Shibboleth	Access policy	
Access			TBAA	Client
Exploration	Harmonization			Client
Compare	Time Space Levels		Harmonization	Translation
Analysis, 1	Harmonization	Standards,	Conversion keys	Client
Transfer				
Download				Client
Logging				
Preservation				
Analysis, 2				
Re-purposing				

Within the CESSDA scheme there are two basic recommended principles that also need to be kept in mind:

1. The aim is to build up national repositories for scientific data. They should be under national jurisdiction and national financing. This will in the long run create the largest supply of data;
2. There are international comparative collections of data that exist in many national copies. There should preferably be one common authoritative responsible archive maintaining one authoritative version of such collections.

The variety following from these principles may to a large degree be described and otherwise taken care of through the general principle of standardization. However, the special language problem which is so important in a Europe of 25-35 different languages also needs a special solution.

In the process view it is necessary to explicitly outline which functionalities the portal is supposed to cover. Not everything listed so far is dependent upon or related to what metadata standard is employed.

If we look at the variety of tasks outlined in the project application, they may be boiled down to 5 main points:

1. Develop a more powerful data interface for the data portal:
    - More sophisticated search / browse or data discovery possibilities, more focused, also across languages;
    - Better possibilities to handle results from data discovery functionality.

2. The total system should be able to handle more complex data structures than at present:

    - Complex data models, in particular:
        i. data over time,
        ii. across space,
        iii. across languages,
        iv. with linkage of micro-macro type data

3. The system needs to develop and implement a system for persistent identities, to facilitate the idea of a common catalogue of data resources across different data repositories, and to connect knowledge products (e-prints) to data resources, potentially going both ways, embedding of analytic results in e-prints vs. referencing and lookup of data from e-prints.

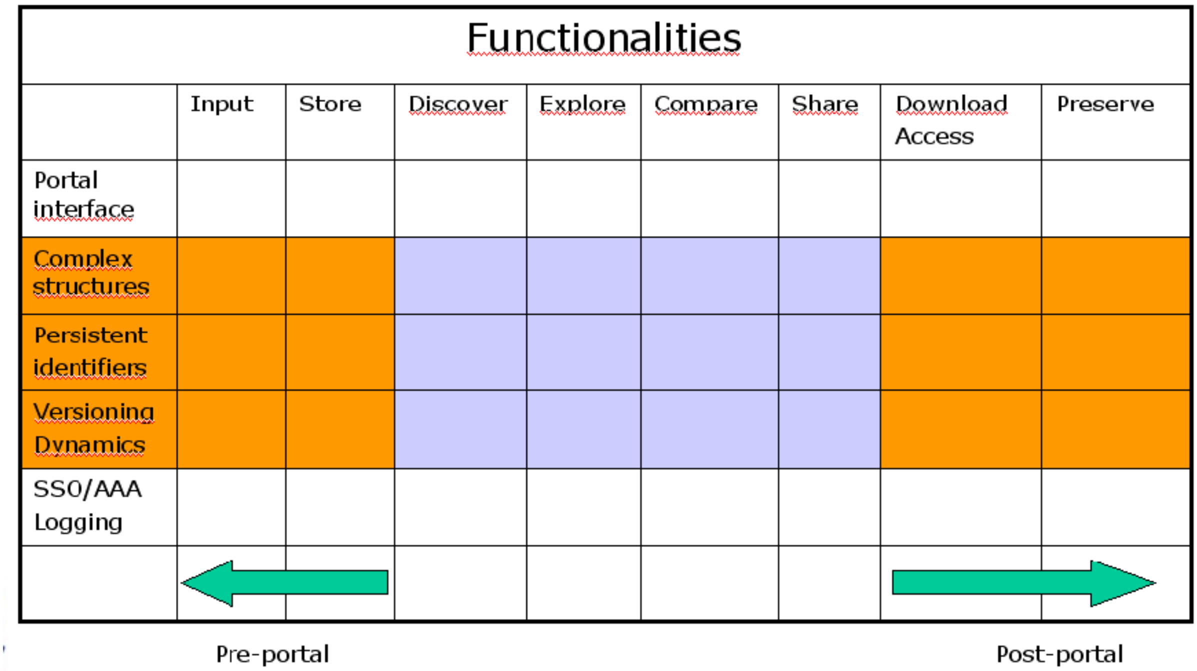
4. The infrastructure should handle problems of versioning, updating and republishing. These are situations that may generate double-/multiple storage situations. In addition the possibility to add comments, links and references to data should be investigated.

5. Data access requires development of a system for federated single sign on. Such a system needs to store/pass on information about the user for use whenever a new data repository (server) is accessed for data exploration. In addition such information needs to be logged, for reporting and control purposes.

The flexibility problems of the present CESSDA data portal reflect limitations of file structures and metadata standard. DDI2 is not an optimal metadata standard given the present ambition level and these concerns have already led to the development of a newer version 3, explicitly aimed at solving most of these identified problems. Some of these problems are related to the time / life-cycle perspective that the data archives are putting on data and relate to the concept of dynamic (meta-)data, (in particular the ability to collect and include knowledge and experiences gained from use and reuse and feed them back into the metadata), which is an ambitious general expansion of the user perspective. Other new / complicating factors are more directly related to file complexity problems, comparative data, time-series data, etc., and the need to develop descriptions and functionalities for relative or relational variables of many kinds.

The main aim of WP5 is to outline a general architecture for a “one-stop-shop” for **data lookup** and **exploration** where we allow for a reasonable amount of a considerably more complicated data picture, and to evaluate to what degree DDI3 as metadata standard manages to describe and potentially solve the problems that have become apparent and, not least, to evaluate the practical implementation problems. To develop such an evaluation we need to specify

a) What are the aims of the portal, so we have to know what the portal is meant to be;
b) What are the potential and problems of DDI3;
c) And what then comes out as the match/mismatch of these two sides;
d) This then has to be tested against the potential for technical implementation.



Functionalities

	Input	Store	Discover	Explore	Compare	Share	Download Access	Preserve
Portal interface								
Complex structures	🟧	🟧	🟪	🟪	🟪	🟪	🟧	🟧
Persistent identifiers	🟧	🟧	🟪	🟪	🟪	🟪	🟧	🟧
Versioning Dynamics	🟧	🟧	🟪	🟪	🟪	🟪	🟧	🟧
SSO/AAA Logging								
	←	←					→	→

Pre-portal
Post-portal

If we summarise this as a table, we see that functionalities discussed can be phased as pre-portal, portal and post-portal. And of our 5 main tasks the first on data discovery and the last on user sign-on and authentication, authorisation and access are not much if at all related or dependent upon actual metadata standard. But the table above, more than anything tells us that the central functionalities will be dependent upon metadata organisation.

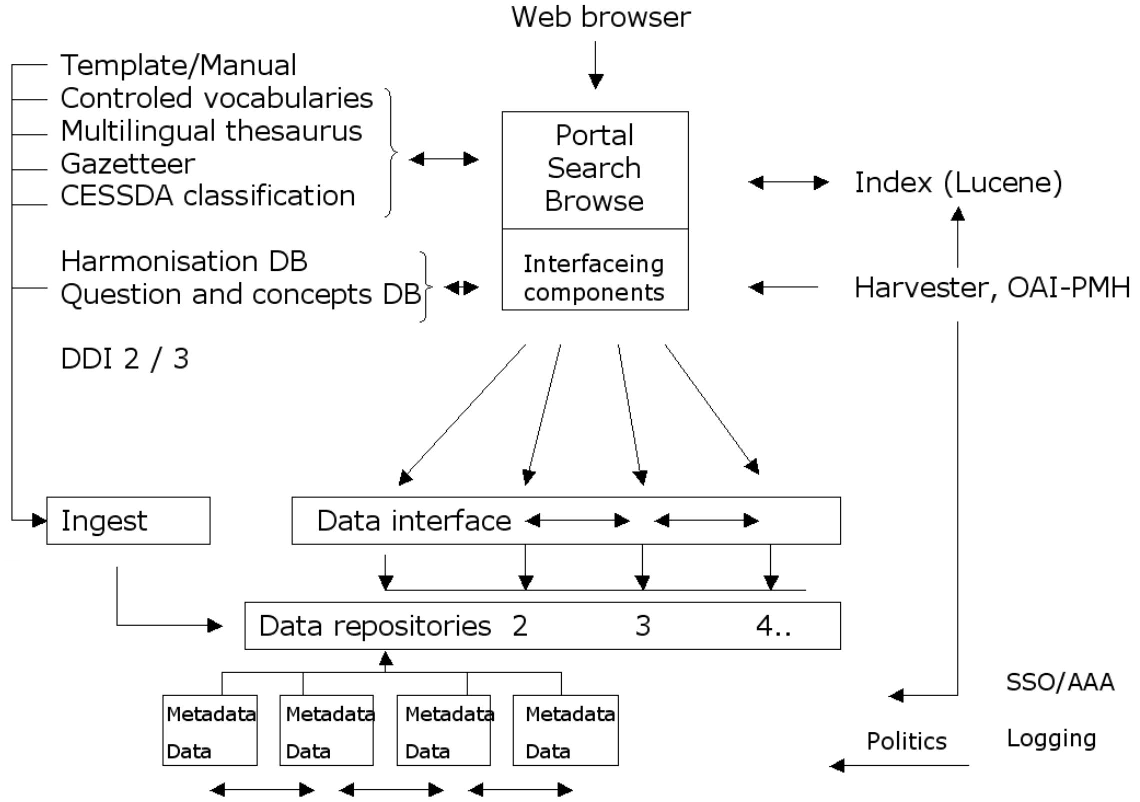
In addition to the above, it could be an aim or technically feasible for the portal to:

a) Include data beyond the CESSDA organisation, as long as the data repositories linked up adhere in a reasonable degree to the rules set for data description and technical solutions, going further in the “horizontal” direction;
b) In the “vertical” direction, any institutional data repository in itself could constitute a portal or an aggregation of primary repositories, so that the harvesting of metadata and access mechanisms proceed in a two-step way, one archival portal may deliver its metadata index readily processed up one step;

c) Or, as already is the situation, CESSDA members could be represented by more than one server, e.g. one for survey data, one for aggregate data and one for qualitative / textual data.

Slightly related to this, but more of a general political/strategic question for a cessda-ERIC, could be the option to develop one or several *CESSDA secure sites*. Such sites could be included under the discovery functionality, maybe also under some explore functionality, but follow its own rules for data access.

The ambitions for an expanded CESSDA data portal may be sketched as follows below. The challenge is to make tools and resources play together in an integrated network of social science tools and resources. The interfacing components of the "portal" element in the drawing represent the *services* description within the fusion layer, i.e. the functionalities available, while the "data interface" element represent the *collections* description, i.e. information collected on the actual resources. Both are normally elements of a "thick portal", i.e. a portal that is aggregating functionalities in one separate fusion layer or component between the data provision layer and the portal presentation layer. The more we think beyond Nesstar repositories, the more we need to include explicit "aggregator" components and collections descriptions, (i.e. a sophisticated metadata system) between the portal top layer and the content providers.



Some specific expansions originally envisioned for the enhanced CESSDA Portal:

1. The portal harvester could harvest metadata from all data repositories that support the Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH), be it as today from Nesstar servers or potentially from other systems. This could be standardised since it is possible to define the record syntax beyond the default Dublin Core of the OAI protocol.
2. Such a strategy could employ the Open Source search and indexing tool Lucene to build a common index across the different data repositories registered with the portal. This index would be the main tool used by the portal to discover data, but the discovery service will be linked up with a variety of controlled vocabularies, classifications and services for translation.
3. The portal is supposed to give controlled access to data. However, data may be found in a variety of data repositories/servers; thus the situation may be that:

    - Within one repository, data may be more or less complex organised files or collections of files, even text-based data documented in Dublin Core or similar is possible;
    - In practice at data object level, formally more often across data repositories, data may be under different legal rules/data regimes or access policies. Access to data is filtered through one common or a number of specific access policies;
    - Data may be stored in different technical systems, sometimes with overlapping versions of data;
    - Data may be documented and organised under different metadata standards;
    - And data may be (documented) in different languages.

4. Thus, it will be a major problem to access and compare data and metadata across both single files and different repositories unless there is a substantial degree of standardisation. Even within the same repository or file collection there will be problems of contrasting data.
5. It will be obligatory in such a system that common guidelines for documentation of data should be established, to guide the data deposit/ingest process. Thus a common template or equivalent DDI profile for minimum levels of documentation is necessary. An ingest strategy should never be considered complete, and the strategy dictates procedures and mechanisms. A clear ingest strategy aims at creating persistence.
6. Further, the very nature of a data archive will underscore the need for data curation / long-term data storage. This question is probably outside this system, but influences metadata requirements considerably.

As indicated, this picture was originally analysed as divided into a fundamental provision layer, an application oriented fusion layer and a presentation layer, it indicated several ambitious extensions or potential extensions to the present CESSDA portal complex. However, the portal interface problem and the SSO problem came out to have minor consequences for the system architecture. Since this is intended as an infrastructure for research data, and since research in its very nature is studying relationships more than mere descriptions, the major problems will be related to the handling of the data, i.e. storage and use of complex data, comparisons of datasets (horizontally in the scheme), integrating this comparison problem with a harmonisation database, and similar. Other derived problems in this same area are about persistent identifiers and versioning of data.

This then indicates that the ability to handle data, in particular the complex collections of data resulting from large comparative and over time data collection projects, becomes extremely important for this project. And it further indicates that the metadata component and its ability to support complex functionality development in the fusion layer, its ability to bring data up from the provision layer and available for efficient presentation also will be of the utmost importance.

In the data provision layer we could list problems / ambitions:

1. Data storage and data complexity and its relation to metadata standard.
    - Data have to be stored in a way that makes them available for loading, for Display in an exploration process;
    - It is an aim to load more than one data object, for comparisons;
    - It is an aim to allow data complexity that covers the most common complex cases of present experience, comparison for exploration could mean comparisons of units or sets of units within one (collection of) files;
    - It is an aim to allow different technical solutions for data repositories;
    - For metadata standards, the requirement should be that they should be able to document data instances at a defined minimum level and be supported by a technical solution.
2. Metadata requirements and organisation
    - It should be possible to harvest a minimum level of metadata;
    - It is an aim to make the data repositories "crawlable" to enhance visibility in ordinary web search engines;
    - There should be different entry points, study, section, question or variable;
    - A specific problem of using the thesaurus synonym/related terms idea with external crawlers. (Maybe a low priority problem).
3. Necessary to be able to harvest metadata in such a generalized picture by own means
    - Crawlers are not by default available for Lucene.
4. SSO solution and relationship to metadata standard/setup
    - Incorporate necessary data for access policies as metadata at data instance level.
5. Practical ingest and storage solutions for many languages. Facilitate ease of translation and insertion of keywords and concepts.

At this stage of analysis, the prime conclusion has become that the data-handling problems of the portal are the most important problems to solve, to be able to develop the portal further. User interface, search and explore, and likewise user authentication and authorization are important problems, but in this sequence the metadata and the data handling problems have to be solved first. It is a basic problem with the present DDI2-version that it is so tightly linked up with rectangular files or tables/aggregate data. At the moment, we have the possibility to document single files and to link them together in hierarchical systems, but we have no retrieval-related or other functionality making use of relationships. Shifting from DDI2 towards DDI3 as main metadata standard represents a possibility to incorporate more complex data models in a more flexible way. DDI3 has specified a general mechanism for the grouping of files. However, to base further work on DDI3 requires some legitimising analysis of what kind of needs there actually are for the ability to handle complex data and to what degree DDI3 actually solve such problems in an acceptable way.

Illustrations of problem types to solve
If we take as examples some of the most central multinational data collections, they represent:

Eurobarometers:
National samples = some grouping of data units into national groups;
Relatively standard cross country questionnaires expressed in national languages;
Some basic variables are distinct national, political parties, regional system, etc;
Substantive content grouped as themes or modules, recurring at irregular intervals;
Some important questions asked repeatedly as trends;
Many other questions repeated more hap-hazardly;
Presently almost 150 “studies”, 5 – 32 countries a round;
Approximately 20 languages;
16000 questions;
60000 variables of different kind (lots of grid- and multi-response variables).

ISSP/ESS:
As a data model, not very different from the Eurobarometers Sets of units in national samples. Main themes repeated by intervals. One of the major problems is that the true comparative nature generates a huge amount of extra documentation.

British Household Panel Study (BHPS):
This represents more than one type of analytic unit, individuals, households, families, i.e. a hierarchy in the file structure that necessarily should allow multi-level files. It is also panels, data collected over time for the same unit, an additional “variable”-level construct where relative or relational variables are brought down to analytic unit level. However, not common to regard this as time-series, more like a trend concept.

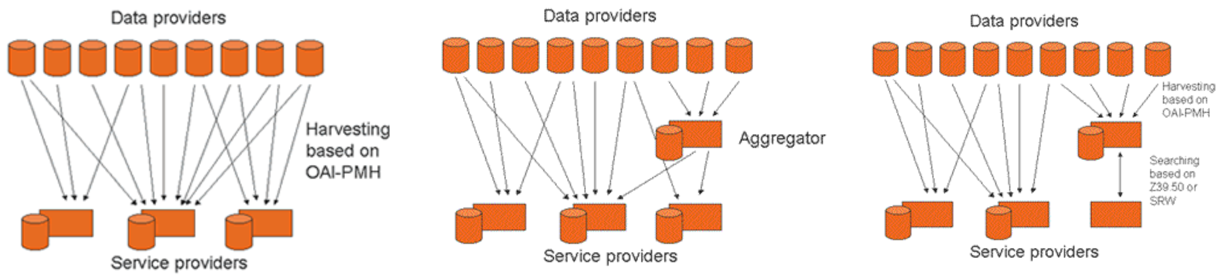
Problems for a generalised data documentation project:
When we work with complex organised data collections, we could single out some topics:

How we communicate structure and content efficiently;
How we organise complex data for optimal extraction;
Complex collections soon become cases with extant reuse of metadata in data collection instruments, in data documentations processes, etc. These are prime cases to benefit from service oriented architectures. For more in depth discussion of complex, in particular comparative collections; see WP5.2

The social sciences (and similar) need very detailed documentation. It is difficult to see other metadata solutions than DDI3 that solves both the analytic user’s problems and the data archival storage, cataloguing and curation problems. In fact DDI 3 has been developed for these problems.

The provision layer / The data repository

The OAIS Reference model [1] spans the space between a data producer and a data consumer, specifying the functions filled by the data archive. The task of documenting data according to the needs generated by data access is not clearly located. The term “ingest” is used about 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 an Archival Information Package into the archival repository. Data are stored in a distributed set of archival repositories. However, a data repository may be seen as a layered storage between a basic storage and a service level, and the repository model may be configured in a variety of ways, as illustrated below:

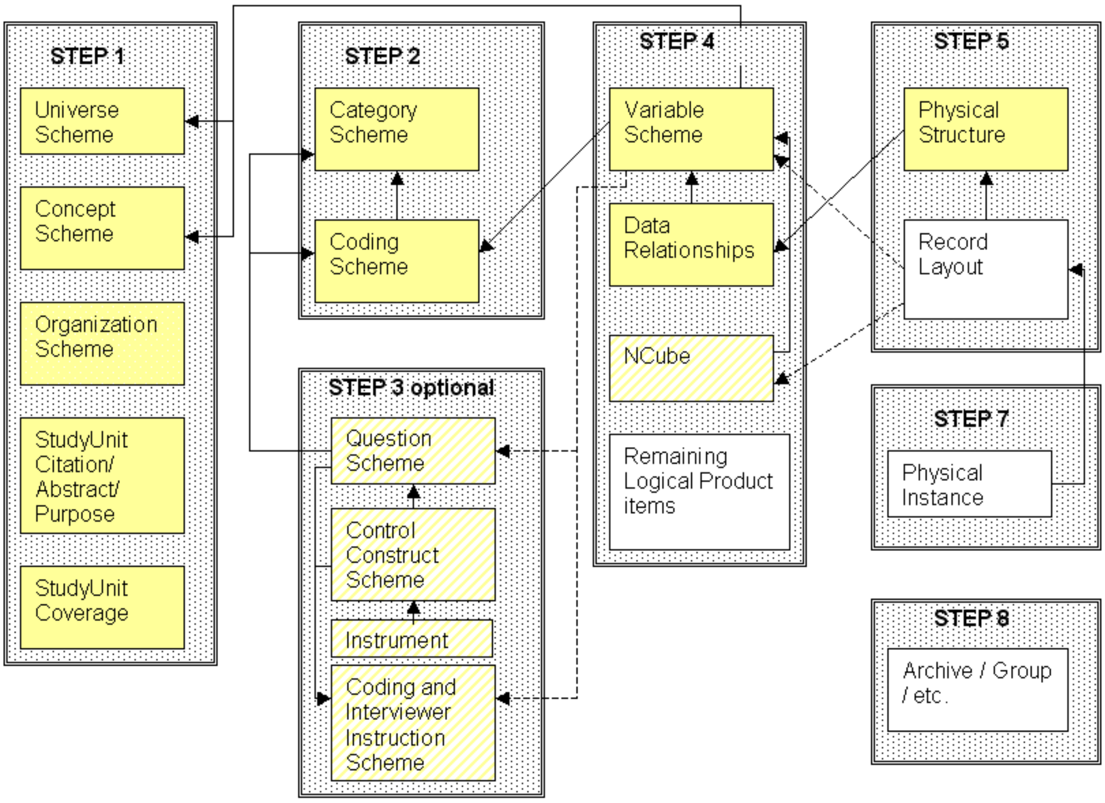


Held against the OAIS Reference model this presupposes that the distributed data provision layer is seen as equivalent to OAIS storage. Such archives cover two basic aims, preservation of data and active scientific use of data. CESSDA data repositories are focused both on preservation and use, data are not only stored to be stored but to be actively used. However, such repositories may also be seen to represent several layers, in different configurations. The Open Archives Initiative (OAI) has defined this as a division between a basic data providing layer and a service providing layer, where portal services or other common functionality will be built on top of the service providing layer. This will have some consequences for updating and maintenance that have to cover both layers.

Suggestion for an updated portal architecture

CESSDA cooperation is a decentralised structure that has clear similarities with the configurations illustrated above, and the major problem is to bring metadata together, across systems, technical platforms and languages, to build a common data location and extraction system.

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



Even if there are major coordination tasks involved, because of the decentralised starting point this is a project where it is natural to think in terms of a service-based architecture, and much of the argumentation is linked to the use of a common metadata standardisation, DDI3.

Version 3 of the DDI metadata standard may be seen as building on service oriented principles itself. Documentation of a data matrix is organised and developed as functionally grouped steps, bringing together elementary elements or objects where every element in the documentation is identified and is brought into the actual use situation through reference or with the potential of being referenced.

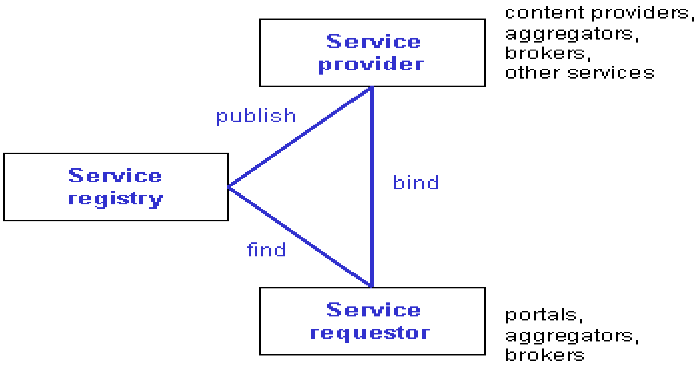
In addition to having the potential of supporting in a very efficient and economical way the work processes of data archival work, the service oriented setup also holds the potential of efficient development and integration of work and applications building on these same principles and based on the same internal communication across many archives.

According to W3C, Web services are a useful solution because:

> *The advent of XML makes it easier for systems in different environments to exchange information. The universality of XML makes it a very attractive way to communicate information between programs. Programmers can use different operating systems, programming languages, etc, and have their software communicate with each other in an interoperable manner. Moreover, XML, XML namespaces and XML schemas serve as useful tools for providing mechanisms to deal with structured extensibility in a distributed environment, especially when used in combination.*

DDI3 is implemented in state-of-the-art XML. Further, DDI3 is particularly developed to solve the problem of organising documentation of complex collections of data, which is one of the major practical problems not yet satisfactory solved in the data archival world. The arguments for DDI3 therefore both concerns the positive benefits of DDI3 and web-services as technology at one level of the work process, and put decisive stress on the integrating power of a service-oriented architecture.

IBM's Web services architecture give a simple illustration of the components involved in a service oriented architecture. It builds on a *service requestor*, a *service provider* and *service registry*. The services offered by the service provider are described using the Web Service Description Language (WSDL), with descriptions made available through the service registry.



The report delivered by Metadata Technology on a technical specification for a European Question Database discusses a comparable setup for a CESSDA context. CESSDA represents a distributed set of data repositories that may be brought together by one centralised registry. A shared metadata standard and a shared metadata model are a big advantage in this respect, both in terms of development and maintenance of tools and integration of data repositories. A common metadata model underlying the whole structure makes it much more realistic to base the portal development on a Service Oriented Architecture (SOA).

As mentioned, DDI is implemented in XML. Already from the documentation process we have the setup to build standardised packaging and exchange of data. The (XML-coded) input packages from a documentation process are basically used for transport into the data repositories and not yet as basis for functionality development. The actual metadata is broken up again in the data repositories, to make it more useful for functionality development. This makes it possible to develop many varieties of data storage or data repositories locally; they only have to a reasonable degree to follow a common information model. In the QDB report these repositories are identified as 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). Many varieties of relational or XML databases may be used to store the data, and such databases may also still function as basis for other or comparable services.

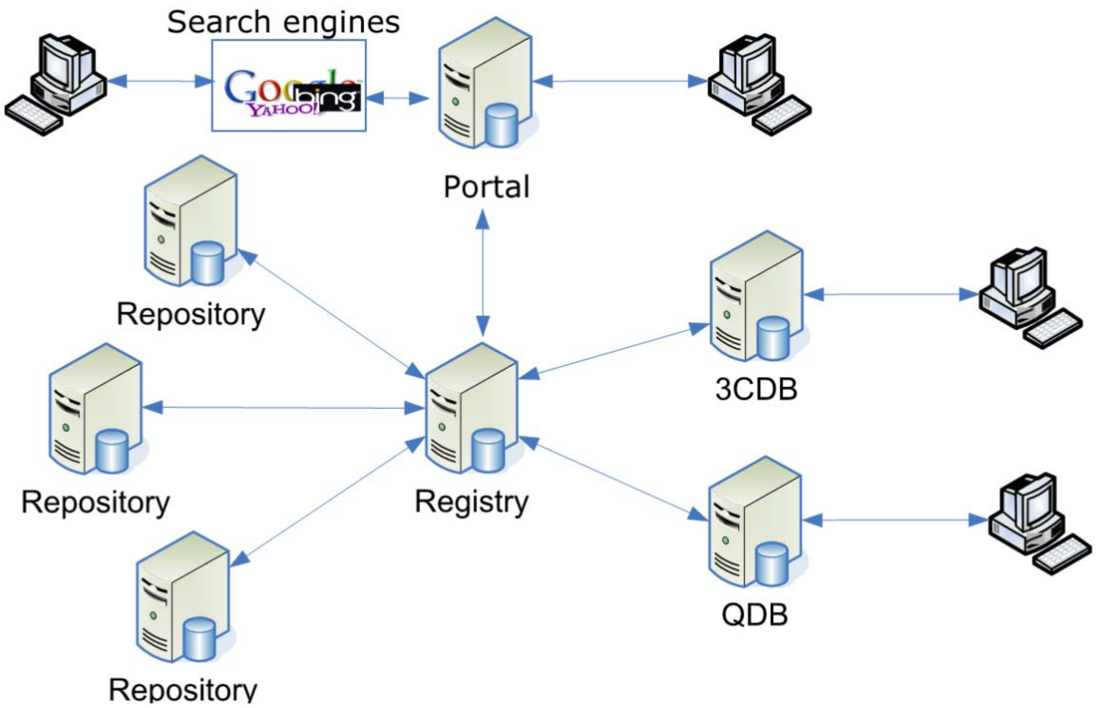
A service-oriented architecture is efficiently held together by a registry that may be indexed as basis for data location and extraction services. The resources are still mainly stored in the decentralised network of servers where the centralised service is fetching elements when required. The requirement is the ability to communicate, messages and content.

The MT report outlines a solution for a Question Database, but the architecture may be expanded to cover several inter-related applications. In that respect the CESSDA Portal may be seen as an overarching umbrella application sharing most of the underlying data resources and the application tools, in particular the concept of a standard gateway software package and tools for communication with backend systems. This should be fairly simple to organise, the main difference between the portal application and the QDB application is the portal need to work against a somewhat more complete metadata model. In particular the portal functionality has to take into account the extension of the model following from the need to handle complex collections. The extension of the model will create a need to describe the collection as a metadata component.

In the suggested setup there is a solution to the versioning problem if the object identification system of DDI3 is implemented with the possibility to add version identifiers. The suggestion is that data published to the applications here are very explicit published/registered and made non – deleteable. This will make the process and products easier to control and administer, and will safeguard the integrity of the system. If a complete DDI3 solution is implemented for the data archival metadata work, then one of the main justifications are that metadata elements may be reused, included by reference. If this comes into widespread use it is obvious that metadata elements, even if they have become obsolete in the original context cannot be deleted because that may break down integrity of the system, elements may have been referenced by others.

Somewhat premature, the portal discussions led to a vision of a “dynamic” QDB based on a general harvesting process. Because of that the QDB index was seen as stored on some central server along with other central indexes / registries. Now it is more appropriate to build on the QDB as a decentralised set of repositories, more or less integrated with legacy systems of an archive, where such a non-deletable status of published data make possible persistence of the referencing system. This does not differ much from the initial architectural ideas, the change in architectural recommendations mainly concerns how to build central registries and indexes on top.

Below is the visualisation starting of the architecture presented for the QDB and 3CDB applications. The Portal is then added in as an additional application, to some degree overlapping with the QDB, which is just using a subset of the information needed by the portal application.



This architecture in the QDB report is argued to support various types of CESSDA applications, and one important argument for promoting this architecture for the portal concerns the ability to integrate applications. This concerns both ability to access resources and to develop interfaces and tools.

Whatever instrument is used to develop the AIP, the AIP 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 a large degree be remedied as a conversion of DDI2 XML to DDI3 XML. This is the same strategy that has been used in the Dutch DatapluS project[1].

Nesstar Publisher v4 operates with the possibility to build simple files together in complex collections and also to aggregate data and produce cubes. A detailed mapping of DDI 2.1 to DDI 3.0 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 A relatively easy implementable solution to generate DDI 3.0 XML, also for the GROUP module could be to use Nesstar Publisher v4 or comparable products as an interface to a XML-writer, since their file format holds all the necessary information for writing out the GROUP-specific XML. This strategy is ignoring the reuse-idea since there is as yet no clear plan or tools for an identifier system, specific implementation plans for an identifications system have to be a major ingredient in a DDI3 production/conversion tool.

To use and develop functionality based on the COMPARATIVE module is substantially more user application oriented.

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

A totally DDI3-based solution will take a long time to develop. DDI3 is a very ambitious project and requires in a service-oriented architecture an identifier system that few have been working on or tried to implement in practice. The present suggestion is therefore an experiment to find out to what degree we may test out the portal functionality for complex instances based on the same solution for producing the XML as indicated in the MT report.

To illustrate how a repository publication could work, here is an example from a hypothetical survey documented in DDI 2 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 give several hundred single files). Variable level documentation should include universe, question text and interviewer instructions. Concepts have been captured in the study description.

- Aiming at developing CESSDA XML based on the DDI3 model, the metadata is imported in 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 DDI 2 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 DDI 2 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 to store 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 has been validated, various optimization steps can take place:

- 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; etc.

- These metadata import / optimization / curation procedures should be accompanied with 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 is under the umbrella of a StudyUnit so it remain a coherent package (no loose objects)
- Once the optimization 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

Documenting ISSP

In our two most relevant use cases we could list versions of hierarchies: Actually, we could have them in many versions, which soon become an argument for preferring DDI3 to DDI2. The present Nesstar implementation does not have the same reshuffling potential as a full-scale DDI 3 version.

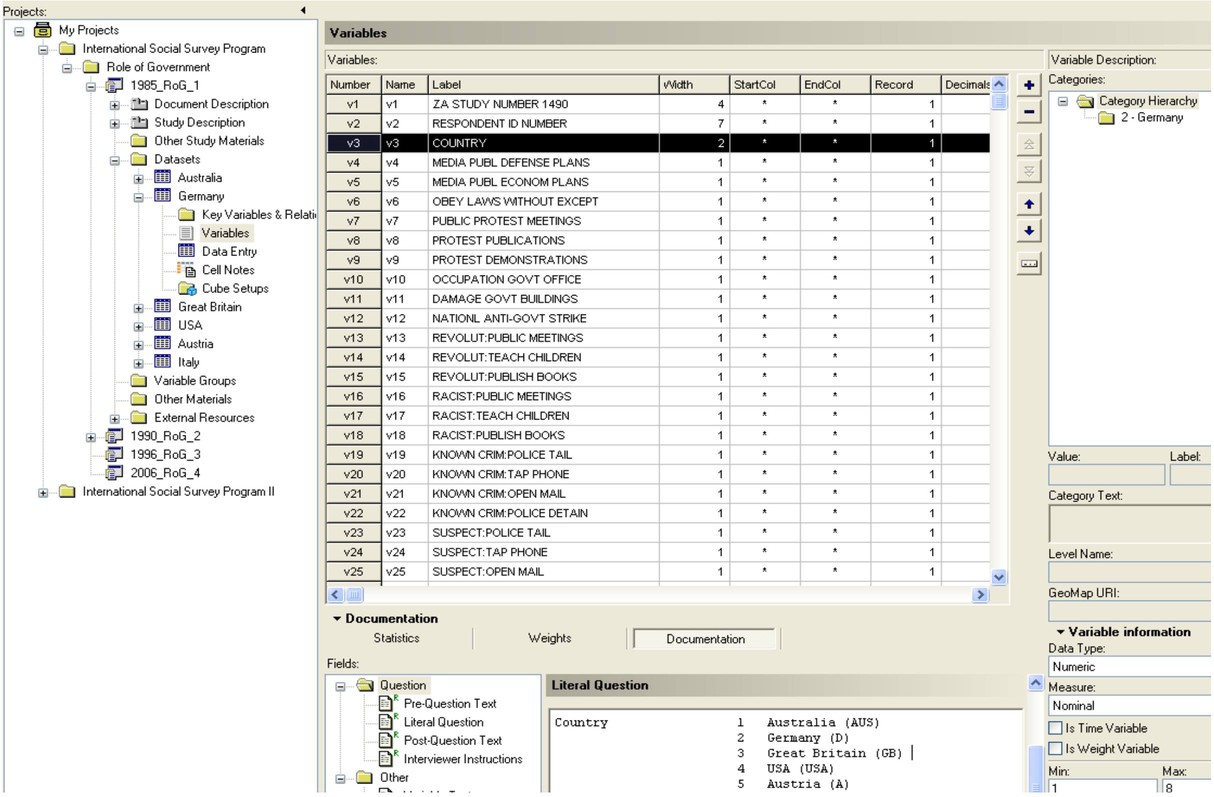
ISSP	ESS
Role of Gov't	Wave (2002, 2004, 2006, 2008)
1985, 1990, 1996, 2006	Countries
Module/Topics	Topic
Question	Question
Variables	Variables

We want to include/combine all relevant descriptive information in a comprehensive package, and we want to develop the functionality a user needs to play around with this to be able to produce the useful end product to investigate analytically. So it is actionable information Vs more dead descriptive info.

The Nesstar example presently has 4 conceptual levels:

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

A DDI2-based solution with ISSP Role of Gov’t 1986 as use case could look like:



Variables:

Number	Name	Label	Width	StartCol	EndCol	Record	Decimals
v1	v1	ZA STUDY NUMBER 1490	4	*	*	1	
v2	v2	RESPONDENT ID NUMBER	7	*	*	1	
v3	v3	COUNTRY	2	*	*	1	
v4	v4	MEDIA PUBL DEFENSE PLANS	1	*	*	1	
v5	v5	MEDIA PUBL ECONOM PLANS	1	*	*	1	
v6	v6	OBEY LAWS WITHOUT EXCEPT	1	*	*	1	
v7	v7	PUBLIC PROTEST MEETINGS	1	*	*	1	
v8	v8	PROTEST PUBLICATIONS	1	*	*	1	
v9	v9	PROTEST DEMONSTRATIONS	1	*	*	1	
v10	v10	OCCUPATION GOVT OFFICE	1	*	*	1	
v11	v11	DAMAGE GOVT BUILDINGS	1	*	*	1	
v12	v12	NATIONL ANTI-GOVT STRIKE	1	*	*	1	
v13	v13	REVOLUT:PUBLIC MEETINGS	1	*	*	1	
v14	v14	REVOLUT:TEACH CHILDREN	1	*	*	1	
v15	v15	REVOLUT:PUBLISH BOOKS	1	*	*	1	
v16	v16	RACIST:PUBLIC MEETINGS	1	*	*	1	
v17	v17	RACIST:TEACH CHILDREN	1	*	*	1	
v18	v18	RACIST:PUBLISH BOOKS	1	*	*	1	
v19	v19	KNOWN CRIM:POLICE TAIL	1	*	*	1	
v20	v20	KNOWN CRIM:TAP PHONE	1	*	*	1	
v21	v21	KNOWN CRIM:OPEN MAIL	1	*	*	1	
v22	v22	KNOWN CRIM:POLICE DETAIN	1	*	*	1	
v23	v23	SUSPECT:POLICE TAIL	1	*	*	1	
v24	v24	SUSPECT:TAP PHONE	1	*	*	1	
v25	v25	SUSPECT:OPEN MAIL	1	*	*	1	

Documentation

Fields:

Literal Question

```
Country             1   Australia (AUS)
                    2   Germany (D)
                    3   Great Britain (GB)
                    4   USA (USA)
                    5   Austria (A)
```

Variable Description:

Categories:

Value:

Label:

Category Text:

Level Name:

GeoMap URI:

Variable information

Data Type: Numeric

Measure: Nominal

☐ Is Time Variable

☐ Is Weight Variable

Min: 1

Max: 8

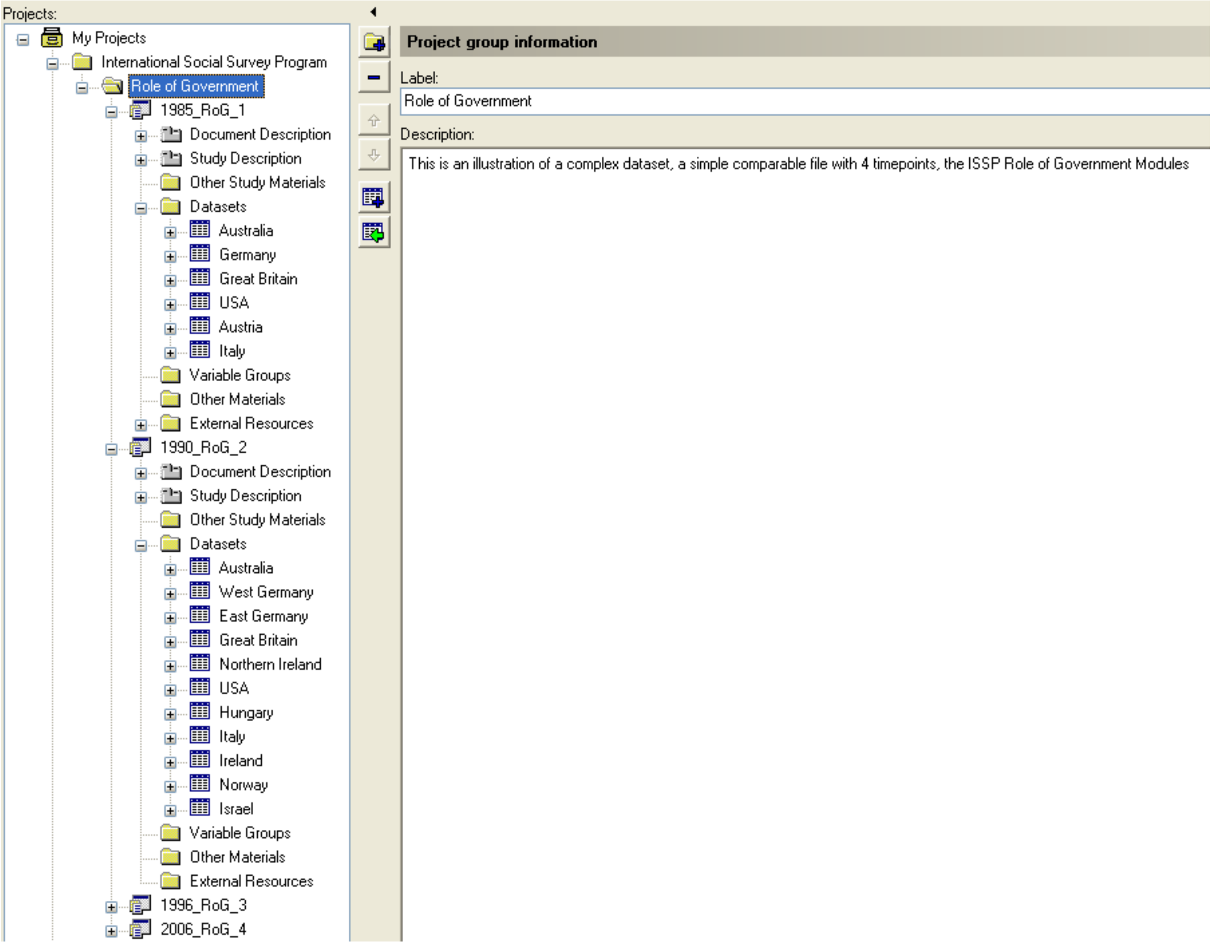
This example illustrates documentation at:

- Project level (The ISSP as a total);
- Module-level (Role of Government is conducted 4 times);
- Wave level (time = 1985, 1990, 1996 and 2006) and finally;
- Dataset (single country) level;
- In addition to the dataset-internals.

Presently Nesstar produces a simple DDI2-based description for every dataset, the study level (the next level in the hierarchy) may be extensively described, and the file and the project levels only have summary abstracts. But the information stored in this file system makes it fairly simple to write out a very rich and flexible GROUPing DDI3 XML. However, it is not making use of the DDI3 identification system. A major project would have to be carried out to develop functionality for development, use and maintenance of the DDI3 identificator system.

In the “dataset-internals” it would be possible to include explicit elements to describe deviations from the common standard. This is the starting point for development of the functionality related to the COMPARATIVE module. It outlines potential for measures in terms of universe or sample, concepts, question, category, codes and variables. It is a formidable task to design software for such functionality and the uses made of it is probably not yet well understood. The conclusion here is that this would become very detailed and potentially very useful information for some collections of data, but it is fairly far over into the analytic part of data work. It is possible to use the structuring power of the 6 generic maps, and it is possible to generate or write out the XML of the Comparative module if DDI3-compatible XML is asked for. However, it is probably a task that is more use-oriented and less of a portal toolkit task.

The time-dimension visualised as a uni-dimensional DDI2 setup:



Projects:

- My Projects
  - International Social Survey Program
    - Role of Government
      - 1985_RoG_1
        - Document Description
        - Study Description
        - Other Study Materials
        - Datasets
          - Australia
          - Germany
          - Great Britain
          - USA
          - Austria
          - Italy
        - Variable Groups
        - Other Materials
        - External Resources
      - 1990_RoG_2
        - Document Description
        - Study Description
        - Other Study Materials
        - Datasets
          - Australia
          - West Germany
          - East Germany
          - Great Britain
          - Northern Ireland
          - USA
          - Hungary
          - Italy
          - Ireland
          - Norway
          - Israel
        - Variable Groups
        - Other Materials
        - External Resources
      - 1996_RoG_3
      - 2006_RoG_4

Project group information

Label: Role of Government

Description: This is an illustration of a complex dataset, a simple comparable file with 4 timepoints, the ISSP Role of Government Modules

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 to standardise 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 it can be automated.

With a hierarchical file system as a potential ingredient the relationship between components becomes pre-defined, we do the job when we read in the file(s). This goes more for group than comparison; groups are just technical where comparisons are substance based along many dimensions.