

Title	A CESSDA Common Data Portal; metadata harvesting, indexing and search technology (D5.3)
Work Package	WP5
Authors	Atle Alvheim
Date	August 2009
Dissemination Level	PU (Public)

Summary/abstract

This report addresses the problem of bringing together the information content in a decentralised set of data repositories across Europe as the basis for a common data location portal. More specifically, the task here is narrowed down to the question of how to create a common catalogue as the basis for portal functionality

CESSDA already has a common data portal of substantial value. One important aim of the CESSDA-PPP project is to make this existing portal a better tool for European research, through more and varied data content and more timely and scalable, i.e. better technologies. The new CESSDA portal should be able to:

- Handle more complex organised and over-time related data;
- Handle comparative data and support on comparison problems;
- Integrate micro- and macro-related data to facilitate study of multi-level scientific questions;
- Support data collection with better instrument development tools;
- Support data exploration and retrieval processes with good standardisation, harmonisation and comparison technology.

Development of the new CESSDA portal is expected to be a large project and will be a gradual process, where different technologies may be supplementing each other for some time. In practice the collection and indexing of data from many repositories for use in one single CESSDA portal tool could follow several different strategies. The recommendation is that the complete system should be based on the implementation of the DDI 3.0 metadata standard to standardise data documentation work and a service-oriented technical architecture, also building on a repository-registry model to support data location and data exploration processes. As some aspects of this may become a long term strategy, some intermediate steps may be evaluated, in particular how to develop the DDI XML needed to describe more complex data relationships and to transport such information around the system. Likewise the OAI-PMH protocol for metadata harvesting and the Lucene open source search and indexing tool may be used to generate test data that will make it possible

to work on portal user interfaces and functionalities in parallel with the build-up of the data management system. Beyond this the system may accommodate a variety of data documentation and data repository technologies. The data repository technology has to support some data loading and exploration technology.

WP5.3 Foundations for a CESSDA common data portal

At the CESSDA-PPP project Cologne workshop in October 2008 the participants discussed characteristics of a CESSDA central data portal and generally agreed to the following points:

1. CESSDA needs a focal point, one starting point that holds large amounts of scientific data. The portal should potentially also harvest metadata and link to data from CESSDA-external sources;
2. Those data could be of various kinds, surveys, aggregate data, textual data. The portal should handle most varieties of complex collections;
3. A relatively high level of standardised metadata is required, it should not be possible to publish data to the portal if there are no related metadata;
4. We should bring users to that focal point in the most efficient way when they are looking for data, users need not know about CESSDA. It is a fact that Google is the dominant internet search engine and the one most used even by researchers looking for data;
5. The portal should make it possible to locate data resources by:
 - a) searching across suppliers, systems and languages,
 - b) browsing across defined controlled vocabularies,
 - c) resource location across languages requires translations of returns, so that it is possible to interpret the content;
6. Treatment of long hit lists is a very difficult problem;
7. Exploration of variables could be based on metadata both with and without standardised frequencies;
8. Repositories / storage should cover different types of data, however a comprehensive object model as is support for standardised input/output;
9. For exploration and comparison purposes a data-viewer should be able to load data from more than one file and more than one repository in a session;
10. The metadata standard should build on DDI, either version 2 or 3. Dublin Core should also be supported if relevant for the data;
11. Accessing metadata should be free of restrictions and does not require authorization;
12. Accessing data across national borders requires an explicit access policy;
13. A data archive may publish data to the portal from more than one local repository via an aggregator service;
14. CESSDA members should be able to store complex data instances and to retrieve on the basis of data relationships;
15. A work-bench concept is not of primary importance, data location and data exploration are the prime keywords;
16. Versioning of data should allow identification of the most recent or the authorised version of data;
17. It should be able to link micro- and macro-data;
18. The CESSDA portal needs a system of Persistent Identifiers, to support referencing. Persistency is related to versioning;
19. The multilingual thesaurus and a concepts and harmonisation database should be available for functionality development.

Some background

For the last 15 years CESSDA has been involved in two major projects in parallel:

- a) The development of a common data catalogue;
- b) The development of a common metadata standard.

The two projects have enriched each other and contributed to significant advances for the data archiving community. The present outcome is a common data Portal for 12 archives, making it possible to explore 5.400 datasets. The CESSDA membership varies a great deal in terms of technical capacity and data resources, but these common projects have created large benefits for all. The portal work is based on the belief that European research activities will benefit from better integration and easier access to the collective data holdings.

The expanded common CESSDA Data Portal brought forward by the PPP project is intended as a common single entry point into data repositories stored at many CESSDA data archives across Europe. These data repositories presently represent considerable variety, in technology and architecture, as well as legal frameworks and languages. The overarching problem of the PPP technical project is how to integrate and bring together in a common top solution what may be quite diversified in its basic foundation.

For such a portal to become the central one-stop-shop for researchers looking for data it has to represent the majority of holdings of the data archives, it has to have *content*. This also implies that it has to cover many varieties of data. In addition to substance and capacity to handle it, it has to be able to communicate with these underlying data in two important ways. First, it must be an efficient information system linking these sources together to give prospective users an efficient overview over holdings, in a way that breaks down language barriers. Second, it must open possibilities to explore and investigate data in ways that are relevant for research-related use, beyond the level that is opened by access to metadata holdings only. Very simplified this requires that the metadata content of data-repositories are indexed or made available for search and browse procedures, and it requires that data located may be accessed for further exploration.

CESSDA work to organise the archiving of scientific data can be summarised with some basic general keywords / requirements characterising the holdings that has been built up, as the starting point for outlining an integrated application:

1. Community based maintenance of data, data and tools intended for academic use, data archived for long term preservation, holdings covering varieties of data, data of proven scientific quality;
2. Reliance on common documentation standards, fostering interoperability and compatibility of technologies;
3. Reliance on standard and tested general purpose technologies, scalability;
4. High availability, durability and reliability, both in technical and scientific terms;
5. Simplicity and user-friendliness important for a variety of users.

CESSDA already has a common data portal of substantial value. One central aim of the CESSDA-PPP project is to make this existing portal a better tool for European research, through more and varied data content and better technologies. Compared to what already exists, the new CESSDA portal should be able to:

- Handle more complex organised and over-time related data;
- Handle comparative data and support on comparison problems;
- Integrate micro- and macro-related data;
- Support data collection with better instrument development tools;
- Support data exploration and retrieval processes with good standardisation, harmonisation and comparison technology.

To be able to handle more complex data is a formidable problem, but also an absolute necessity of the portal. A substantial part of the data available for social science research is not adequately represented in the present portal.

The portal as a common catalogue

The CESSDA Portal represents an **implicit or virtual common catalogue** for the collective data holdings of the various CESSDA member archives. For this implicit catalogue the portal is supposed to present **one common data finding aid**. In OAIS¹ terminology CESSDA is a federation, adding together the collected data holdings of its members. In addition, the portal may organise simplified administration of important **shared resources**, both data and application oriented resources.

The general information model spans several complicating elements, both methodological and content-related:

1. Different types of data are of relevance and may be covered by the portal
 - a) Unit record data based on samples and questionnaires as data collection instruments, traditional survey data
 - b) Register- or universe based aggregates, often collected through administrative procedures, stored in databases or as cubes. This may be referred as aggregate, regional or ecological data, these data are partly of a different methodology, partly collected for different types of units and partly of a relational character functioning to describe the contexts of individuals.
 - c) Text- or qualitative data are of increasing value and abundance in a world where data are the direct outcome of administrative or other processes;
2. The supporting metadata may be available at different information levels:
 - a) Data collections;
 - b) Study / samples;
 - c) Topic / sections;
 - d) Concept / questions;
 - e) Variables.
3. A mixture of languages. Europe presently may represent at least 50 states, 50 legal systems and 30 active languages.

¹ <http://public.ccsds.org/publications/archive/650x0b1.pdf>

To organise the information content efficiently for the portal finding aid, a quite sophisticated registry building or indexing procedure has to be implemented. This is only deemed possible if data in the underlying data repositories are documented according to common standards. It is generally accepted that the most efficient solution to diversity in such systems is standardisation, in processes and products. This is one of the points where the portal has to be seen as imbedded in a larger infrastructure, where data preparation and data documentation processes has to follow generalised rules. The Data Documentation Initiative²(DDI) is run by the data archives in common. The DDI project is focused on development of a documentation standard that meet the very specific needs of the data archives in a modern communication oriented Internet setting. Version 2 of the standard (DDI 2) or version 3 (DDI 3) are recommended as common metadata standards also for the CESSDA subsystem of this larger world-encompassing envisioned infrastructure. In addition there have to be clear common rules steering the documentation processes and tools of specific data. One common integrating tool for implementing such rules may be a specified common CESSDA DDI profile or template, with a classification of metadata elements as being mandatory, recommended or optional for description of social science data. In addition such a DDI profile should be incorporating the CESSDA PPP developed sets of controlled vocabularies. Further clear rules for data quality control should be implemented.

The portal as a tool for exploration and preparation

Science is about relationships, while data often represent a specific timepoint/-interval or a spatial reference. This generates a need to **compare and link**, to put data together for generation of relative measurements, contrasts, trends, to create possibilities to measure differences and developments.

For some data this aspect is incorporated already in the data collection or data preparation process, we collect and store complex organised data, data measured also across scientifically important third dimensions because appropriate analysis requires that some measures are generated already in the collection process. CESSDA data archival work needs to describe and store this data complexity, since they represent important research questions and analytic necessities.

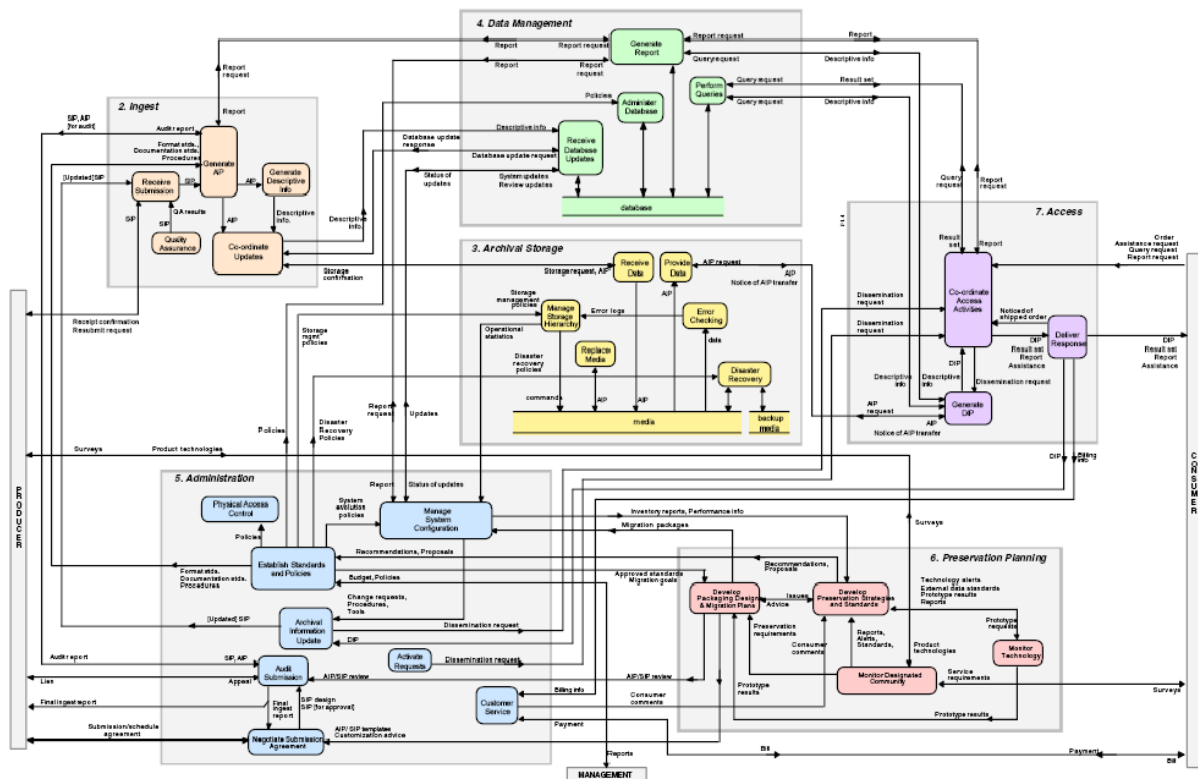
An important enhancement of the CESSDA Portal will be the incorporation of a Questions Database and a Concepts, Conversions and Classifications Database. These expanded databases will facilitate enhanced support for data collection processes and data exploration processes and link the data archives into the user communities in new and relevant ways. In technical terms they underline the need to build integrated systems and share resources.

The general framework for the CESSDA portal

The OAIS Reference model³ is accepted as a general framework for the CESSDA data infrastructure. It is stated as an overall requirement that CESSDA data archives should comply with the OAIS model. Below is a presentation of the model, as what is termed a composite functional view. This is not meant to be a specific recommended implementation design, but rather a loose functional framework for discussing concepts and comparing systems.

² <http://www.icpsr.umich.edu/DDI/>

³ <http://public.ccsds.org/publications/archive/650x0b1.pdf>



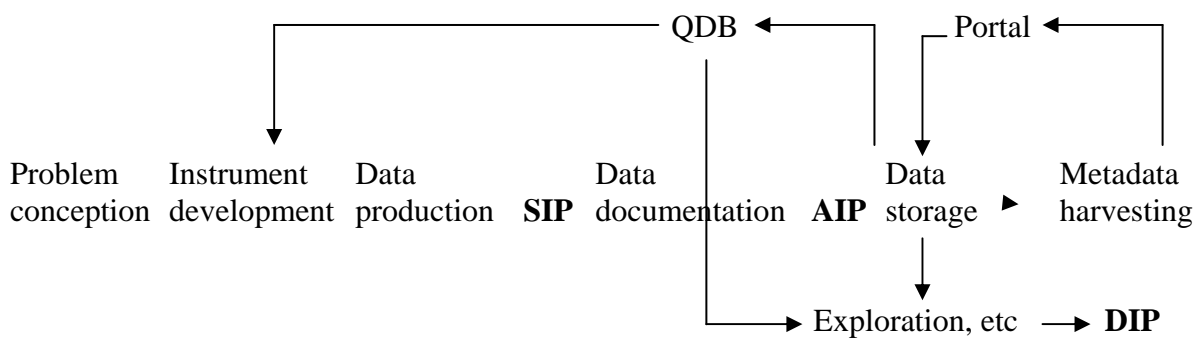
The model spans the space between a data *producer* and a data *consumer*. For a data archive the model is presented as 6 functional tasks:

- a) Ingest, the process of loading data into an archive;
- b) Storage, to store the data in a suitable way;
- c) Management, aspects of data work;
- d) Administration, to administer the workings of the archive;
- e) Preservation planning, organise long-term secure archiving;
- f) Access, organise processes whereby users may access or order data.

If these reflect the functionalities of a total CESSDA infrastructure, the *portal* could be defined as an extra layer over the access function. The portal is supposed to supply efficient overview over data holdings and through data exploration give insight into possibilities for constructive data use before data are downloaded for further processing by the consumer.

Tools for bringing order to chaos, shared models more than same technology

The viewpoints presented above indicate that it is not obvious what constitutes the “portal”. Is it the whole infrastructure, with the aim to generate a common interface to a list of applications, or is it just the common cataloguing of data resources, the extra layer over the access function in the OAIS model? Looking at the whole infrastructure or process from researchers initial problem conception to final data analysis and reporting as an integrated system could reduce development of a data portal to the question of developing an integrated catalogue of holdings across many data suppliers, a problem that technologically is not very different from other applications that CESSDA aim to develop across the many data-archives of Europe, e.g. the common questions databank (QDB). Very simplified the comparability of the processes could be depicted as:



Generally, the OAIS model outlines a process where researchers specify a problem, develop instruments to collect data to carry out empirical analyses related to the problem and as the end product of that phase, develop a Submission Information Package delivered to a data archive. The data archive further puts together necessary material and documents the data according to its internal or professional standards and develops an Archival Information Package, that is stored / archived for long term preservation and secondary use. At about this point in the process, the report on a European Question Data Bank indicates that a QDB can be developed as a registry-based derivative from the archival repositories.

It is the starting point for this present exploration of portal development that index- or registry development for a portal as a general process is comparable to the QDB-building process, but that the functionality and the amount of metadata needed to process may be quite different between the two applications. However, the similarities and the integration potential becomes a very convincing argument for building on the same kind of solution. One problem of a different nature could be that the QDB-report more or less takes DDI 3 as a given, in terms of data model, architecture and technology. The implementation and tools development problems related to DDI 3 however are formidable and will make this a longer term aim and probably force us to explore how we can work towards that aim in intermediate steps.

While the exploration processes of the portal require loading of data, a QDB rarely goes beyond metadata.

To be able to develop central common collections like a QDB or a Portal for general use across a variety of distributed data archives, it is obvious that these starting points, the archival

stores or repositories have to comply with some general rules. Some general recommendations for the practical work are summarised below.

The data packages or data sets made available through the CESSDA local data repositories need to comply with a reasonably **standardised information model**. Given the recent history of this work within the data archival movement, it is recommended that data are documented according to **DDI 2 or DDI 3**. There are major differences between these two versions of the metadata standard, however as framework for data work both will facilitate delivery of a data package coded in XML as an AIP into the data repository. We then have the potential of a fairly integrated metadata model. Such a comprehensive information model is a must for tools development, in particular tools should be developed for the data preparation and documentation process. Presently we see these tasks solved in two main ways, either through the Nesstar Publisher or via database solutions. In order to develop such tools, the detailed information model is the basic requirement.

Development of such a CESSDA / DDI metadata or information model can build on the DDI development work and work carried out within the Metadater project⁴. Without such a model it will not be possible to develop a suitable common data documentation or conversion tool for the DDI3 standard.

To move a data package from SIP status to AIP status generally requires some data management and some data documentation work.

Within the CESSDA federation, there should be common guidelines for such work, for quality control and minimum level of documentation.

Because of the differences in languages, data should as a minimum be documented in two languages, the national and a common language. English is the only common language of practical relevance.

The present existing CESSDA portal employs DDI 2.0 as metadata standard. To standardise documentation, CESSDA developed a common documentation template⁵ for the Nesstar Publisher tool. This was based on a thorough working through of the complete standard and a classification of elements in 3 groups, mandatory, recommended and optional elements.

First: This template should be re-evaluated, for national and common language respectively. The archives have now accumulated extensive experience working with these tools and such experiences should be incorporated into the work. In addition the development of the new version 3 of the standard brings up new types of elements.

Second: This template should be mapped⁶ against the DDI 3.0 set of elements, and a comparable template / DDI profile should be developed, in 2 versions: for national language and for common (English) language. It is generally expected that the common language version moves some of the recommended elements (i.e. question text) from the recommended to the optional category.

⁴ <http://www.metadater.org/index.htm>

⁵ <http://www.ddialliance.org/related/cessda-rec.pdf>

⁶ Such a mapping may be found at <http://www.ddialliance.org/DDI/ddi3/variable-fields.txt>

An important part of a common template / a DDI profile is to incorporate the controlled vocabularies developed within WP4. If the present CESSDA Common template is kept unaltered for archives still mainly documenting data following the DDI 2 standard, it should be evaluated to what degree even that version of the template should be updated with relevant controlled vocabularies.

The most important controlled vocabulary in this connection is the multilingual thesaurus ELSST and the two-level CESSDA study classification. These should if possible be incorporated in the process and used to insert concepts and keywords where appropriate. A separate task could be devoted to study the problems of having keywords inserted at question level. This would be relevant for development of a Question Database, and registration of questions in such a database should be located in the process at the end of the AIP development process.

Whatever instrument is used to develop the AIP, the AIP should be coded in DDI compatible XML.

The AIP is ingested into an archival data repository.

Archival data repositories may represent other varieties, some of them quite practical. (Meta)data repositories may be the starting point for many other archival applications not discussed in our present scheme, maintenance-, service- and information systems. This may create a certain danger for conflicting developments resulting in integrity problems. A central archival data repository may be mirrored or viewed on a gradual scale from the full-fledged archival storage to a simple metadata storage for publication or local application purposes only, i.e. it may mirror a main archival storage completely or be completely separate, e.g. metadata without data, maybe for data protection reasons. For a more complete discussion of repositories as a set of layers (representing both preservation and services), see the section below on OAI-PMH.

However, to develop a common portal over a set of archival repositories along the lines that is described here requires a relatively stable starting point. Whatever other services a CESSDA data archive runs, they cannot be allowed to interfere with the stability and status of the data repository. The performance and scale of this system is not expected to become the most important problem to solve, rather it will be the maintenance and consistency of the system that will be the major problem to solve.

There have to be a clear set of guidelines or recommendations on what kind of (data) repository is presented through the common CESSDA applications, the data Portal, 3CDB or QDB. And the archival repositories set up for the CESSDA applications have to comply with a common metadata model. To practically obtain this the simple solution at present is to state that a data repository should support DDI in either version and comply with the OAI reference model. Beyond that, repositories may still vary in architecture and technology, but support DDI as common basic information model.

The AIP as a package is usually depicted as consisting of metadata + data, although it is possible to think of this as metadata without data if the local repository is only for visualisation on the internet and not for archival storage. This may be problematic in terms of rational consistent activity and data preservation policies, all this should, if possible, follow the general principle of storing the data only once and use it by reference.

Metadata may be expressed in one, two or many languages. We need to identify which languages.

Whatever the representation in the underlying repository, metadata must be extractable for indexing/registry-building purposes.

The portal is interacting with the decentralised set of data repositories in two ways:

Harvesting metadata to build an index/registry for location purposes;
Loading of data-files or collections for exploration purposes.

According to the OAIS reference model, in response to requests, the archive (the OAIS) provides all or part of an AIP, or collections of AIPs to a Consumer in the form of a DIP.

Following this, the portal works against AIPs and some of the functionality may contribute to create the DIPs.

Data instances may be more complex than at present under DDI 2.0, data-objects may (at least) be, in practical language:

Instance (project), module/topic, time/wave, levels, datasets, sections, questions, variables, options

In this list, there is no absolute hierarchical relationship between objects, but objects have to be identified in some general scheme for later application/presentation purposes. This list does not intend to reflect the whole metadata model.

A data package in a data repository may consist of a (collection of) metadata + data combinations. For portal purposes metadata are usually under few restrictions and mainly used to locate resources for further use, data are in comparison usually under restrictions and have to be loaded into a suitable computer program for exploration and processing. Crossing over from reading metadata to loading data triggers access control procedures.

However, for data location and exploration purposes access to metadata may be limiting but remains of great value. It is not an absolute requirement that data should be available for online download.

Access control is usually located at repository level, not at individual data package level, i.e. the main part of access policies are formulated for the whole repository or institution, not for single datasets. However, it is possible to diversify access rules at lower levels as part of the metadata following a data package. That means that access conditions should be defined as part of the metadata.

Only metadata are read by harvesting processes to build indexes / registries. For the CESSDA harvesting problems employed to develop a portal index, it could be specified as a requirement to be able to harvest metadata from servers supporting DDI and the Open Archives Initiative Protocol for Metadata Harvesting. OAI-PMH has the ability to support various metadata standards and identifies both data providers as single servers and

aggregators over diversified sets of servers. The protocol also allows for persistent treatment of repository deleted material.

The Lucene open source search tool is an efficient tool to build a common index across a set of data repositories / servers.

The OAIS model distinguishes between consumer inputs as queries, orders and report requests. A query is mainly working against an index, a simple search session returning a list of hits that constitutes the starting point for further exploration. An order, in contrast, involves the distributed set of data repositories, similar to what we have termed exploration, where additional data are loaded from the data repositories. It will probably become of some interest to log various types of requests, to generate data on how users are using such services.

The OAIS model also distinguishes between ad hoc and event-based requests. In such a system as this portal, it is difficult to see any particular relevance of the concept of event-based requests. However, it could be of interest to link this up with some kind of notification based on Publishing or harvesting processes.

The CESSDA Portal will employ 2 different technologies for data location: Search and Browse. The procedures work through the ELSST thesaurus for systematic access to hierarchies of concepts and concept translations and synonyms.

Presently the CESSDA portal only locates studies by substantive criteria. As the basis for the browse options, the CESSDA 2-level classification for studies and the ELSST thesaurus are used as a more fine-grained hierarchical browse list with synonym possibilities.

The future search setup will be slightly expanded, incorporating some recommendations coming out of the FLEXible information Access using METadata in Novel COmbinations (FLAMENCO) project at UC Berkeley School of Information, so-called faceted orthogonalities. Working through the ELSST thesaurus will still represent content / substance as the prime source for structured searches, covering also the principles of hierarchical drill-down and synonyms. This will then be sought orthogonally with measures of time, space and methodology. In practice there will be two main sources for search keywords, either free search with phrases supplied by the user that are matched against the appropriate language version of ELSST or through look-up via ELSST. This may then be narrowed down by specification of time-point or -interval, by specification of geography or aspects of methodology.

The portal interface may also function via ELSST as an interface to a question database or a harmonisation database.

A CESSDA Portal and its data location technology

The CESSDA organisation is in itself a decentralised structure, the CESSDA nodes are represented by national data archives. Social science data collection is to a large degree linked up with national characteristics. Financing, lawgiving, languages, educational systems and researchers training ground are national in character, and there is a general agreement that a structure with national data archives is the one that produces the largest amount of data. The CESSDA data infrastructure is to be built on the principle that the member archives themselves maintain their data repositories, in a decentralised system. To have possibilities to develop integrated services on top of such collections of decentralised holdings, these holdings have to be managed according to common general rules. The most important technical rules would be that repositories should follow the same standards for work and support the same communication protocols.

The Data Documentation Initiative (DDI) is a metadata standard developed by and for the social science data archives and is a natural candidate as a common metadata standard for the CESSDA infrastructure.

DDI3 was developed to break out of the codebook-centric, rather static view of data documentation that the second generation DDI2 represented. In addition to allow for a very flexible setup for versioning of metadata and data elements, the standard is also expanded with considerable new potential to describe complex collections of data. This is highly relevant for the CESSDA archives. For the CESSDA archives the ability to handle data complexity is of great interest.

DDI 3 is built on a service-oriented paradigm and naturally inspires application building on top of the standard to follow the same principles. DDI 3 breaks up the documentation standard into atomic elements, bundles elements into functionally oriented modules and supports a process where elements are collected or put together according to user needs. Using DDI 3 to document data means fetching elements and putting them together into larger integrated constructions or composite metadata objects. This requires that all elements are uniquely identified and may be located and communicated around the system following the web-services principles. However, it is possible to write DDI3 XML without identifiers, but then the service orientation breaks down.

When data are properly documented, they are stored in archival repositories, and over these repositories we build user-applications.

Different applications built over the same foundation should be easier to integrate, different applications are to a large degree just using the same data, i.e. the same or different elements out of the same repositories in different ways. A QDB and a Portal are good examples of applications that could be built over the same repositories and be developed based on the same principles.

To be able to build a common data location tool over a series of data repositories, it is a highly efficient technique to develop one common central index that is used as basis for search and browse procedures. In modern Service Oriented Architectures ⁷(SOA), this uses a

⁷ http://en.wikipedia.org/wiki/Service-oriented_architecture

strategy of setting up a registry, a common catalogue as an intermediate database, this registry is indexed to speed up efficiency of search and browse procedures.

The functionalities envisioned for the CESSDA data portal are to locate and explore data resources, so that we could decide which data resources are the most relevant for us to download to our local computer for further more intensive work.

We could contrast three different ways of developing the indexes necessary for efficient search and browse.

A. We could build on the same technology as is used in the present CESSDA portal.

The present CESSDA portal is built over a collection of Nesstar servers and is based on DDI2 as the metadata standard. Participating archives run their local Nesstar server to make their data collections available over the Internet. However, it has been stated as a general principle of the PPP project that the future portal should not be based on proprietary software or solutions and Nesstar is owned by one of the CESSDA member archives. Therefore the close link between the CESSDA portal and Nesstar products has to be generalised.

In the present portal, there is harvesting technology that builds up an index. The present harvester fetches metadata objects expressed in RDF via the Nesstar API, indexes the metadata via the Lucene indexer and stores the index in a new object database. In the same way the ELSST thesaurus and the CESSDA study classification are also converted to RDF and stored in the same database. The user interface on top of this database employs an object query language to fetch objects, supported by some ELSST-based services. The returned hits are used as links to look up the appropriate files in the underlying repository via the Nesstar Webview tool.

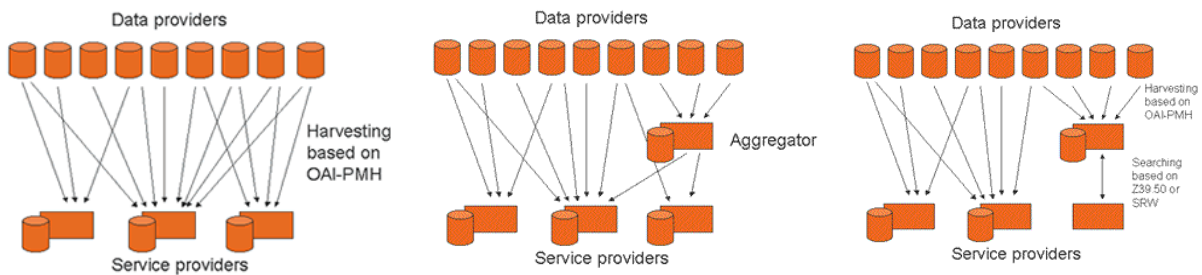
This solution does not meet many of the requirements of a stable, flexible, modern and scalable portal solution, but the architecture has proved its value. It has demonstrated that it is possible to develop a highly efficient search and browse solutions for a common catalogue in the CESSDA context. At present the portal contains well over 5,400 datasets, which is a fairly large amount of data.

B. An alternative is to employ the OAI – PMH⁸

The Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH) provides an application-independent interoperability framework for metadata harvesting from repositories. The OAI-PMH gives a simple technical option for data providers to make their metadata available to services, based on the open standards HTTP and XML. The metadata that are harvested may be in any format that is agreed by a community (or by any discrete set of data and service providers), although unqualified Dublin Core is specified to provide a basic level of interoperability. Thus, metadata from many sources can be gathered together in one database, and services can be provided based on this centrally harvested or "aggregated" data. The protocol arose out of the e-print community, where a growing need for a low-barrier interoperability solution to access across fairly heterogeneous repositories led to the establishment of the Open Archives Initiative (OAI). Originally the protocol was developed to

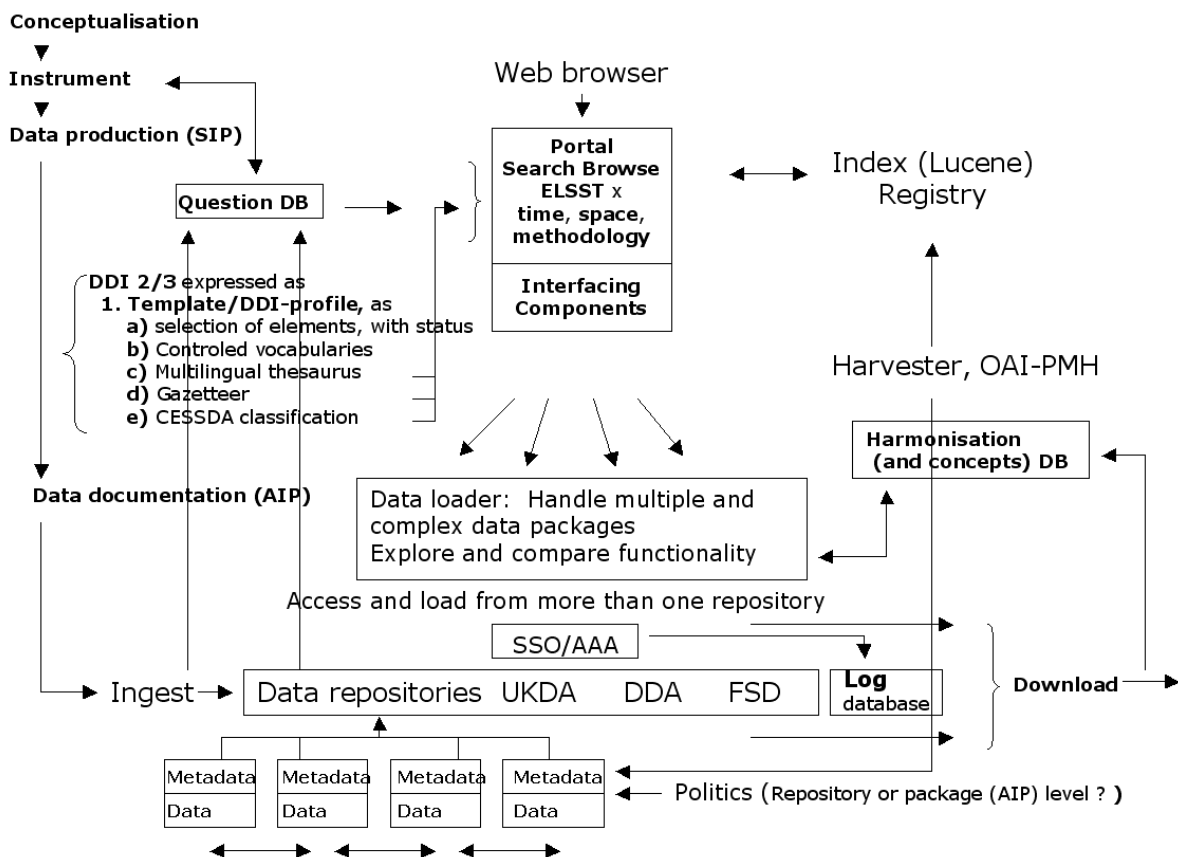
⁸ <http://www.openarchives.org/OAI/openarchivesprotocol.html>

enhance access to e-print archives, but is now taking into account access to other digital materials.



Several configurations are possible for more complex collections of repositories. Multiple Service Providers can harvest from multiple Data Providers, aggregators can sit between Data Providers and Service Providers, or the harvesting approach can be complemented with searching based, for example, on Z39.50 or SRW.

OAI introduces the highly relevant notion that data repositories often are used for many purposes and may represent many practical complications. For that reason they have introduced the distinction between basic data providers or basic stores and the use-oriented service providers, the actual application repository. This distinction has practical consequences for updating, versioning, etc



In connection with the CESSDA technical infrastructure, the OAI-PMH has been a serious candidate for production of search and browse indexes, and an initial architectural sketch was based on using this technology.

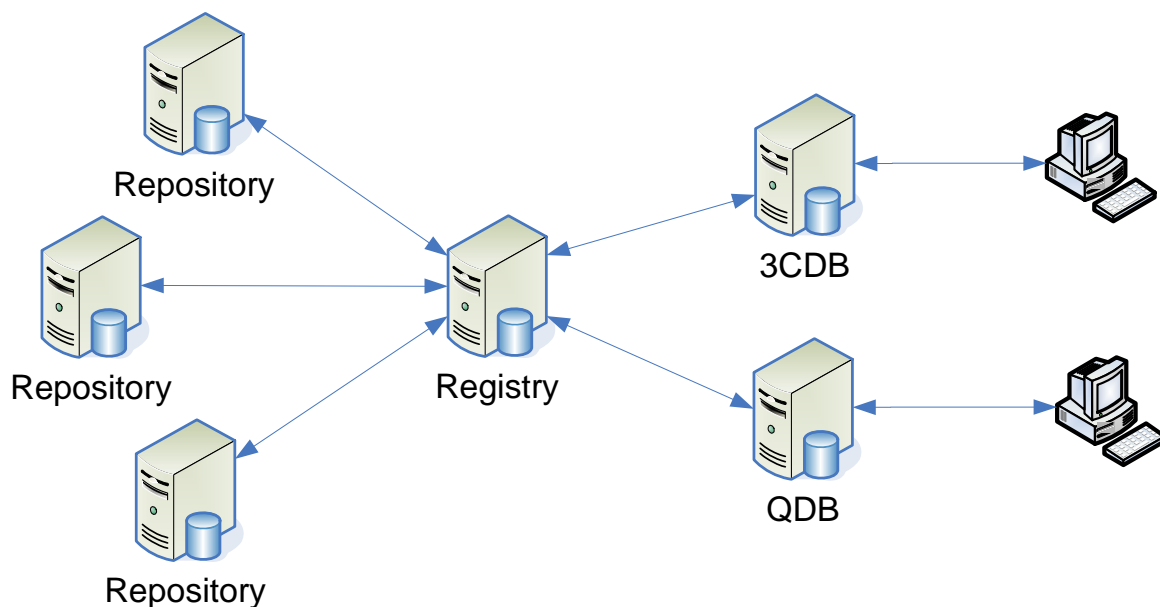
C. The third and recommended alternative for a CESSDA architecture is to follow the same strategy as is outlined for the QDB

The CESSDA cooperation is a decentralised structure, and the major problem is to bring metadata together, across systems, technical platforms, languages.

Version 3 of the DDI metadata standard builds on the same kind of principles. DDI3 is a service-oriented structure. In addition to having the potential for supporting, in a very efficient and economical way, the work processes of data archival work, it also holds the potential for building and integrating very flexible applications. The arguments therefore both concern the positive benefits of DDI 3 and web-services as technology at one level of the work process, and put decisive stress on the integration power of an SOA.

This suggested architecture builds on data repositories that hold data documented according to DDI. On top of the repositories, a registry is positioned, for efficient search and browse services, with potential applications built on top of the registry. All components are able to communicate via XML.

Below is the visualisation of the architecture presented for the QDB and 3CDB applications. The Portal may be regarded as an additional application, to some degree overlapping with the QDB



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

When the general architecture and the web-service principle is established also for the Portal application, several of the further arguments also follow more or less by default:

The starting point for a new CESSDA portal states that:

The portal should be able to handle more complex organised, over-time related data;
The portal should handle comparative data and support on comparison problems;
The portal should make it possible to integrate micro- and macro-related data to facilitate study of multilevel designs;
The portal should support data collection with better instrument development tools;
The portal should support data exploration and retrieval processes with good standardisation, harmonisation and comparison technology.

Repositories

Data archives store the documented data, the AIPs in their data store. More or less overlapping with that concept we may define as the application oriented repositories, but conceptually they are separate. Repositories means the data stores that interact with user applications, be it a QDB or a Portal. Note that a QDB and a Portal may have different data needs, while the exploration processes of the portal requires loading of data, a QDB to a lesser degree goes beyond metadata. Likewise, while a QDB mainly focuses on questions and associated metadata, the Portal has to handle data where the question as a metadata element may be irrelevant. As is depicted for the OAI-PMH example above, different configurations between data providers, their data stores and the application-oriented repositories are possible.

These repositories could be of various kinds. As data are becoming more diversified, such repositories have to be able to handle this diversity. The main requirement is an ability to communicate, to receive and respond to messages and to handle packages of varied information content.

To be able to communicate more efficiently, the QDB report introduces the strategy of splitting the metadata model of the repository into its most important parts and to group metadata into “banks” or groups. The Portal could theoretically run into the need to access all of these banks, although the general functionality of location, exploration and download files generation does not always generate that need. But in addition to what is described in the QDB report, the arguments for recommending this architecture have focused on the ability of DDI3 to allow treatment of GROUPed or COMPARATIVE data. This requires that the data documentation processes have to allow creation of such documentation and the set of registry banks have to be expanded by the GROUP type metadata: Abstract, Concepts, SeriesStatement and SubGroup.

The great integration benefit for a potential CESSDA Portal is that by just a slight expansion

- a) An information model that incorporates the GROUP and the COMPARATIVE modules of DDI3;
- b) A procedure that makes it possible to produce the accompanying XML;
- c) Expansion of the repository banks by an additional concept bank holding the GROUP-type metadata elements.

This strategy could allow the portal to exist as a separate application using the same system-internal services and bank functionalities on a slightly enhanced registry.

Functionalities of a CESSDA data portal

Search phrases in the CESSDA portal pass through ELSST as a filter. Via ELSST the search phrase may be multiplied into several languages or positioned in a hierarchy of concepts that make it possible to resend searches to the engine with alternative search words.

Every time a search is conducted a unique list of hits is returned.

How the hit-list is generated and set up would depend on what information is available in the index/registry and the need to fetch information directly from the repository banks.

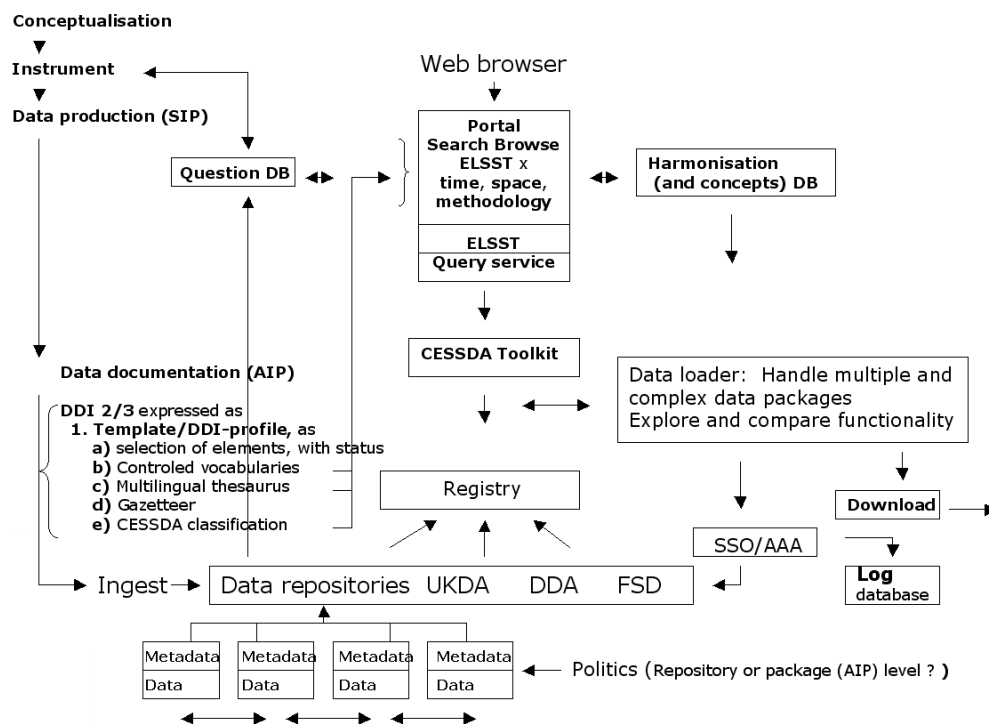
Technologies to explore data located via the CESSDA portal could be divided between

- a) The ability to find and put together simple files;
- b) And the ability to keep and explore the structure of complex files, if such files are available.

Point a) would be very similar to functionalities developed for a QDB, but also basically be to load study units (more than questions) for comparison, i.e. basic display.

Point b) would be to display the structure of complex collections and use that as some kind of menu for loading single study units.

For download, information about requested format would be necessary to generate the files. Since few statistical packages support complex files, it would not be too easy to request complex files.



An intermediate solution for publishing DDI 3 XML for complex data

A major problem in the initial development phase of a CESSDA Portal is how to develop DDI 3 XML. Further, but as a distinct additional problem, since DDI 3 requires internal referencing and communication via web-services all objects need to be uniquely identified using URNs. Managing such identifier systems is a formidable task. But the full web-services capabilities of the system is more of an archival need; to be able to process DDI3 documented data in a portal, we will come a long way by just being able to produce well-formed XML-files.

This following part is copied and somewhat adjusted from the report on a CESSDA Question Bank, to illustrate an intermediate strategy for production of the necessary XML, but also with some ideas on how to optimise this.

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 national files). Variable level documentation should include universe, question text and interviewer instructions. Concepts have been captured in the study description.

This requires an information model that handles a collection level and a wave level above the ordinary rectangular file. Since this is handled by Nesstar Publisher in the hierarchical file format, Nesstar Publisher has the information available that is needed to write out DDI3 XML for the GROUP module.

- Aiming at developing CESSDA XML based on the DDI3 model, the metadata are imported in a CESSDA Toolkit and broken down into several components, A CESSDA Toolkit has to have the Nesstar Publisher ability to handle Nesstar v4 hierarchical files:
 - o 1 or several Study Unit(s) (docDscr + stdyDscr)
 - o Parallel Logical Product(s) (dataDscr)
 - o Variable Schemes (one per file, or since this is a comparative study, one pr. wave) also holding variable groups (fileDscr)
 - o Several Category and Code schemes containing categorical variables' code & labels (one per categorical variable)
 - o Question Schemes and Instruction Scheme (likely one per fileDscr)
 - o Appropriate Concept /and Universe Schemes (depending on how survey and variable level universes and concepts are merged)
 - o 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
 - o Logical Record (in LogicalProduct, one per file)
 - o Physical Data Product (one per file) defining the file characteristics
 - o 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
 - o If cubes are present in the DDI 2, they will generate various NCubePhysical DataProducts

Various other materials can be generated, PDF-files for questionnaires, maps, pictures, etc.

- The 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 have been validated, various optimization steps can take place:
 - o Code and categories used by more than one variable can be merged into a single scheme
 - o Questions and Instructions reused by more than one variable can be aggregated
 - o Concepts and universes can likewise be aggregated (if applicable)
 - o Variables used in multiple files could also be aggregated into a common variable scheme and reused by reference
 - o 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 are under the umbrella of a StudyUnit so they 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.