Title	**Version Control Final Prototype (D12.1)**
Work Package	12
Authors	Ornulf Risnes, Markus Quandt, Vasuda Peddi, Sirisha Kakarla
Source	WP12 team
Date	
Dissemination Level	PU (Public)

Summary/abstract

The goal of WP12 as described in the CESSDA RI Annex was to develop software prototypes, covering different aspects of the envisioned infrastructure. During the project, it has turned out that real prototyping wasn't meaningful or even possible for some of tasks, and it was the consensus among the relevant work package leaders to work on "paper-prototypes" instead - these are illustrations and mock-ups that were thought to cover substantially more ground than possible by working prototypes.

However, during the preparation of paper prototypes for some tasks, most notably the DDI3-related one (T12.3), it became evident that it might be more productive to create an integrated paper prototype encapsulating the DDI3-complexity in high-level functionality. This was partially due to a lack of detail in outcome-definitions in the CESSDA-PPP, difficulties in the evaluation of DDI3 and contribution of use-cases, and immaturity in the proliferation of DDI3 itself. There are also purely technological reasons. The technical foundations on which we believe a CESSDA RI should be built are very much in flux and relatively immature. This fact has been unravelled by tender reports delivered to WP9 and WP11, research and investigation into available technologies mentioned (explicitly and implicitly) in the CESSDA-PPP Annex as well as contact with other communities, in particular the DDI3 community.

During extended WP-leader meetings there was consensus that developing paper prototypes would be sensible for all prototype deliverables (except D12.3 Authentication Prototype where it was regarded as important to ascertain that the fundamental challenge of creating multinational Shibboleth-based networks for single sign-on (SSO) could be solved). Note that D12.5 - Thesaurus management software - was not intended to be a prototype but a first version production system. Note also that D12.2 (classification database) and to some extent D12.1 (version control prototype) were implemented as working prototypes.

The deliverables will be listed below with links to online or otherwise documented software or prototype instances. The first two sections will however define the background for the implementation of the majority of the deliverables.

DDI3 deployment in CESSDA

Three out of five WP12 deliverables (12.1, 12.4, 12.5) built directly or indirectly on version 3 of the Data Documentation Initiative standard (hereafter labelled "DDI3").

Most CESSDA archives currently use DDI version 1.2.2./2.x (hereafter "DDI2" since the versions are similar for all practical purposes) as the preferred standard for data documentation. The standard is supported by a range of in-house data management and preparation tools, dissemination tools (e.g. Nesstar) as well as the current CESSDA portal.

DDI2 is a comprehensive metadata standard, and its proliferation among CESSDA archives has been fairly rapid and extensive. Through various initiatives, most notably the MADIERA project and the current CESSDA-PPP, the CESSDA community has succeeded in adding value to the standard through creation and wide-spread implementation of controlled vocabularies (CVs) for selected DDI2 elements. Along with the creation of the European Language Social Science Thesaurus (ELSST), this value-added CESSDA-wide standardisation has been of key importance to the development of the current portal.

As this report will show, one important finding in WP12 is that DDI2 has not yet been utilised to its full potential; it is possible to leverage DDI2-documented holdings substantially without moving to DDI3.

To avoid jumping to conclusions, however, a discussion of DDI3-findings is required.

The DDI3.0 specification was released in its final form in April 2008. In October 2009 DDI3.1 was released. DDI3.1 was mainly a bug-fix version compared to DDI3.0 and although DDI3.1 is not backward compatible with DDI3.0, both versions will continue to be discussed under the alias "DDI3".

Even though a final release of DDI3 didn't exist at the beginning of CESSDA-PPP (January 2008), the basic concepts of DDI3 were known among the partners in the CESSDA-PPP. It was clear that DDI3 represented a paradigm shift compared to DDI1 and DDI2, and that the new version would need a great deal of investigation and evaluation during the course of CESSDA-PPP before any large-scale implementation could be planned or recommended.

It is an understatement to say that DDI3-evaluation has been challenging. Version 3 of the DDI-standard is not only extremely complex and detailed; it is also still very immature in the sense that it has not yet been put to any significant use, even outside the CESSDA community.

Therefore, it has been difficult to gather material to guide us on these matters. There have been some detailed reviews of selected parts of DDI3 in the context of the PPP (e.g. in packages WP5, WP8, WP9, WP11), but nothing close to an end-to-end-review has been carried out so far. The analyses that exist are therefore mostly of minor parts of the standard, looking at selected features of DDI3 and the impact these have on implementation, both at the archive-level and at the CESSDA-level.

Best-practice documents on different aspects of DDI3-usage have been published (ultimo June 2009) here http://www.ddialliance.org/resources/publications/working/bestpractices . These documents aren't really "best practices" in the sense that they have been proven to work over a period of time. Instead, they are guidelines on how the authors believe the DDI3 should be deployed and used. Although they give valuable insight into the potential of DDI3, it must be stated that it remains to test the suggestions in production. Under WP12's review of the best practices-documents, certain weaknesses have been discovered. Generally speaking, the weaknesses don't appear in the conceptual model itself (which is very sophisticated and well-designed), but in its applicability in computer-assisted production environments.

One example is DDI3's versioning approach. Taking the best practices document as a starting point, there is an ongoing discussion about the enforceability (by automated computer programs) to the suggested best versioning practices. One direct question, members from the DDI technical implementation committee (DDI-TIC) suggested that application-specific, or institution- or community-wide agreements should be made to enforce a specific interpretation of the versioning construction. The fact that perhaps the behaviour of the most fundamental building block of DDI3 is up for interpretation surely is a signal that the elegant and sophisticated conceptual model will have negative consequences for DDI3-driven interoperability, at least across institutions or communities.

It should also be noted that although the substantive parts of the conceptual model *are* sophisticated, elegant and powerful, WP5 and WP8's reviews of the applicability of DDI3 for concrete use-cases (focusing on comparability issues), suggest that there can be challenges and shortcomings (at least ambiguities) also within the model itself. A great deal of work will be needed to produce integrated reviews of bigger parts of DDI3 and the interplay between the standard, archiving and publication workflows and the demands of the research community.

One of the strengths of DDI3 is its support for metadata sharing and re-use of metadata objects. Basically, a single DDI3-instance no longer needs to be self-contained but can instead refer to metadata objects in other locations. This highly elegant and ambitious pattern does however bring a new issue to the table: dependencies between different DDI-instances. To benefit fully from DDI3's relational/networked approach, metadata producers depend on the availability of reusable metadata elements and tools to integrate these into the production environment. While the relational properties of DDI3 make it a good candidate for a metadata model in a CESSDA network of metadata-driven services, challenges following from this approach should not be underestimated.

Below the most important DDI3-related findings are described as a SWOT-analysis (Strengths, Weaknesses, Opportunities, Threats). Many of the points were also presented in an executive report submitted August 2009, but are still valid and therefore included here as well.

Strengths:

- CESSDA archives have invested in metadata capture, storage and dissemination;

- CESSDA has a long tradition and know-how in using DDI1/2;
- CESSDA archives have shown that they can reach agreements on common practices; controlled vocabularies and policies (e.g. CESSDA template);
- Tools and procedures exist to support DDI1/2-usage, during ingest, archiving and dissemination.

Weaknesses:

- Still little CESSDA archive expertise on DDI3;
- DDI3-efforts risk invalidating DDI1/2-tools and practices;
- No production ready tools for DDI3-processing have been developed by archives (or others);
- No end-to-end-evaluation and review of DDI3 has been carried out by archives;
- DDI3's relational, dependency-heavy data-model prevents incremental transition from DDI1/2. This is perhaps the most fundamental problem.

Opportunities:

- CESSDA is active and influential in the DDI-community;
- No real alternatives to life-cycle metadata capture and storage exist today;
- The world is starting to appreciate metadata.

Threats:

- Competing, simpler alternatives could emerge;
- DDI-community could suffer fatigue if DDI3 doesn't gain momentum;
- DDI2-metadata published in e.g. html w/RDFa (indexable and possibly understandable by search engines) could create competition from own archives;
- DDI2-metadata can be utilized to a much broader extent than today, especially in aggregating services (e.g. portals). While this is probably a desirable feature in the transition towards DDI3, it could also reduce incentives to move on from DDI2-based documentation;
- DDI faces competition from SDMX in the community of statistical agencies. With its focus on aggregate statistics only, SDMX cannot really cover the documentation of respondent level data (e.g. surveys), but it is reason to believe that if statistical agencies invest heavily in SDMX for a substantial part of their data production process, less funding/attention will be given to DDI. It is likely that many agencies will stick with their in-house documentation systems for the data documentation and aggregation processes, focusing on delivering standardised and interoperable SDMX-documented statistics to customers and users.

These findings indicate, along with the challenges in DDI3-evaluation discussed earlier, that it is not a given that DDI3 is the preferred choice for CESSDA. In fact, recent developments in the CESSDA/DDI-community suggest that the success of DDI3 depends on it’s adoption by commercial sector and more importantly major and influential software vendors. It is unclear what to expect in this area at this point (i.e. at the end of CESSDA-PPP).

The alternative to DDI3 implementation is - for now - to continue utilization of DDI2. This can, as we shall see, improve CESSDA services substantially with reasonably modest efforts and funding. It is however necessary to note that DDI3's conceptual model is set out to solve a range of problems identified within CESSDA, and that without a joint implementation of DDI3 in our community, many problems will either remain unsolved or will have to be solved by other means. Most likely "other means" in this context will involve extensions of DDI2 and combination with other technologies and standards to create support for creation and management of relations between DDI2 instances and version control and identification of DDI2 instances. Without unique identification and sophisticated (and unambiguous) version control systems, powerful services like the proposed "Constructs, Classifications, Conversions Database" (3CDB) cannot be developed.

It must also be added that some of the findings and suggestions outlined in this document are based upon the existence of a proposed CESSDA-wide Service Oriented Architecture (SOA) which in turn depends on DDI3 being implemented by archives participating in the SOA-network.

There follows a description and analysis of the proposed SOA-model.

A new service network platform in CESSDA

CESSDA does currently have its portal (http://www.cessda.org/accessing/catalogue/).
The portal harvests DDI metadata from decentralised Nesstar-servers located at different archives, and indexes these metadata in a sophisticated manner, drawing heavily on the multilingual ELSST thesaurus for classification and translation of metadata (keywords).

Although the current portal has suffered from both loss of key staff and weaknesses in the development-model (more on this below) after the end of the MADIERA-project, it seemed to be consensus among archives that building new services on top of the functionalities, tools and resources shown in the existing portal would be a good idea.

Originally, CESSDA-PPP set out to extend and improve the portal incrementally, aiming at creating a more feature-rich and user-/research-friendly common interface to the data holdings.
There is, however, reason to believe that the harvesting pattern exercised by the portal could jeopardise other requirements, most notably those of persistent identification and robust versioning. It should be added that without these two requirements, the current harvesting pattern is still very much valid and functioning. The SOA-based solution should therefore not necessarily be regarded as a prerequisite for continued portal developments but instead as a possible tool for expanding the scope of the portal (and indeed the greater CESSDA network of services).

In the tendered report produced for WP9 re: "Question Data Bank", a very different model is suggested. In this model, which is a platform for metadata re-use and interoperability, a central “Registry” would support many (although far from all) of the functions in the current portal (including envisioned extensions), in addition to a range of other services and a system for persistent identification.

The registry would, as the portal used to, serve as a central index for parts of CESSDA held metadata. But the registry does not harvest metadata. Instead, metadata objects must be explicitly registered/submitted/pushed into the registry to become available in the indexes and discoverable through the central interface. Registered objects come with a guarantee that they will be forever present, unchanged and accessible through CESSDA. Changes can of course occur, but a change mandates a new version to the changed object. All registered previous versions must be kept online as well.

For further details on the proposed model, read the tendered report on WP9: http://www.cessda.org/project/doc/CESSDA_PPP_QDB_May09.pdf
The SWOT-analysis below tries to address development and deployment of a new service-architecture for CESSDA RI.

Strengths:

- CESSDA has proven the ability to create a central value-added index (the portal);
- Metadata and resources (ELSST and shared templates/controlled vocabularies in particular) are on a high level (although many are DDI1/2-based);
- Able developing teams present in the organisation.

Weaknesses:

- Lack of tradition in coordinated software development across CESSDA;
- Little/no common software tools are deployed across the community (apart from Nesstar);
- Existing tools do not support DDI3 (see DDI3 SWOT above);
- Archive-staff/organisations not geared up towards the publishing and archiving patterns (e.g. for ensuring versioning/persistence) necessary in the new network;
- Outcome/desired functionality not clearly defined;
- No common support and maintenance teams exist. Such entities need to be established to answer both technical and substantive support queries from archives, end-users and other stakeholders;
- Large degree of autonomy in archives creates heterogeneity in software and practices. A successful SOA-environment depends on homogeneous and standardised interacting service nodes;
- The suggested model is currently not sufficiently reviewed. It shows great potential on paper but it needs to be evaluated for a great variety of scenarios in order to prove its usefulness in CESSDA;
- Uptime and scalability requirements will increase dramatically (see also PID SWOT below) for IT and development teams across CESSDA. Due to the relational and networked structure of the network (it is easy to envision millions of relations between metadata instances), it is crucial that participating nodes have very high availability. If one node breaks, all relations to the node will also break. Fault-tolerance can (and indeed should) be built into this system (in the form of cache-solutions, failovers, etc), but adding this will also come at a substantial cost.

Opportunities:

- Software with potential to ease deployment components/patterns/solutions continue to emerge (grid, cloud, SOA);
- No real competition ready to take CESSDA's place on this scale.

Threats:

- Suggested approach relies on DDI3. See DDI3-threats in DDI3 SWOT above;
- Generic search-engines (Google, Yahoo) are starting to support semantic web (e.g. microformats, RDFa), see: http://www.alistapart.com/articles/introduction-to-rdfa Metadata published in html + RDFa can be "understood" by search-engines, rendering the "discovery"-aspect of CESSDA somewhat superfluous.

See also tendered report WP11, and executive summary WP11 (http://www.cessda.org/ppp/wp11/WP11_Executive_Summary_vFinal.doc) for SWOT on choice of underlying technology (i.e. Grid/Cloud/SOA).

Deliverable D12.4: Portal prototype

The prototype is available at the following URL:
http://extweb.nsd.uib.no/cessda-ppp/portal_prototype/

Note that this is a "paper prototype" in the sense that it just displays certain features of a portal; it does not for example, contain a real data index.

For information on how the prototype should be understood, links to a few demonstration screen-casts have been added along described functionality.

Before the portal prototype is further discussed, a background section now follows, placing the portal within the proposed SOA-based service network platform (discussed above).

Registry/portal relationship

The most important difference between the current portal-architecture and the newly proposed one, is that the portal no longer uses direct harvesting as the means of collecting metadata from archives.

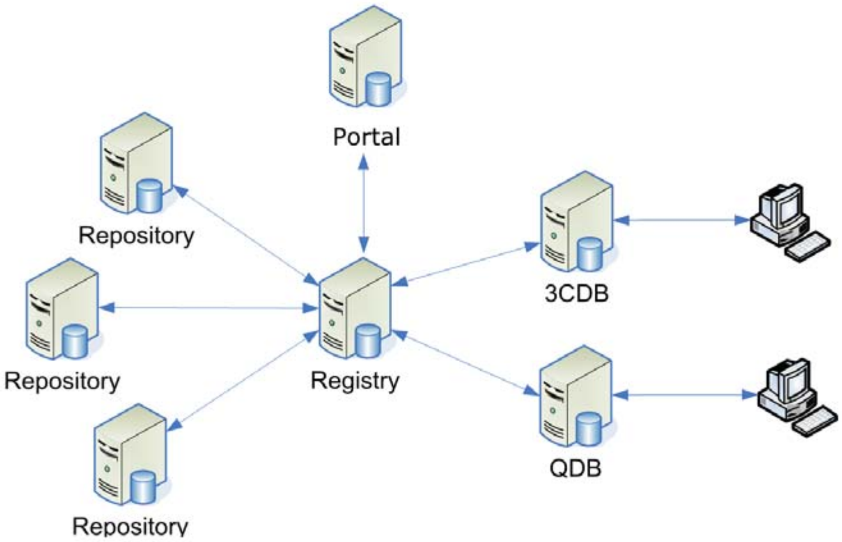


Figure 1 - Portal's relationship to the SOA network

Instead, the portal uses the registry to look up and monitor additions and changes in the published and registered material. Once notified by the registry (via the registry's proposed "notification mechanism") about a change in a metadata record, the portal can fetch/harvest the metadata record in question from the source, guided by references in the registry. This three step process is shown in the figure below.

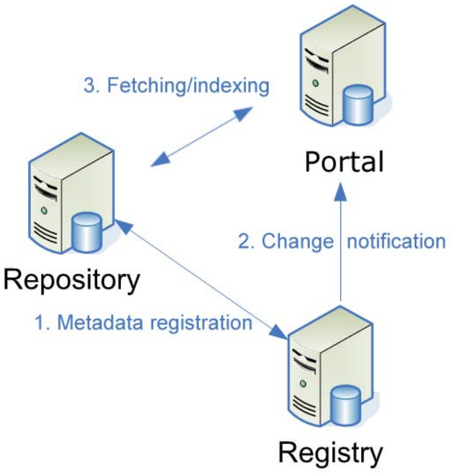


Figure 2 - Registration and subsequent notification and indexing

The registry is at heart of the architecture, holding references to every identifiable, registered object in CESSDA. The portal uses the registry to maintain an updated value-added collection of the CESSDA holdings. As opposed to the registry which indexes only a small subset of the registered metadata, the portal will build a complete index from the available metadata. This complete index is, as we shall see, of key importance to building user-friendly and feature-rich functionalities on the portal level.

The registry may be regarded as the key source of machine-readable metadata in the CESSDA community; the portal will be the main source of human-readable metadata.

Attract users, increase data usage, stimulate research

Perhaps the most fundamental shortcoming of the current portal is that its content is not exposed to generic search engines like e.g. Google and Yahoo. In effect, this means that a researcher, despite the large amount of textual documentation added and made available by CESSDA archives, will not be able to find CESSDA held data and metadata the way they find everything else; via free-text searches in the major search engines.

The single most important feature of the new portal will therefore have to be exposure of all available metadata for indexing by the major search engines. The method for doing so is well understood and straightforward:

- Create dynamic, well-designed web-pages for each "metadata object" (e.g. every Study, every Question, every Variable, etc);
- Make sure each of these metadata object pages can be reached via hyperlinks (which is the way search engine crawlers navigate).

Design choices in the current portal architecture make search engine exposure difficult. The single most important recommendation for a new portal is that it is designed from the ground up with search engine exposure built in. Fortunately modern search engine practices dictate that metadata rich pages with good usability and readability and otherwise sensible designs also achieve better ranking than "difficult" pages. Consequently (and this wasn't always the case), what's good design for humans tends to be good design for search engines and vice versa.

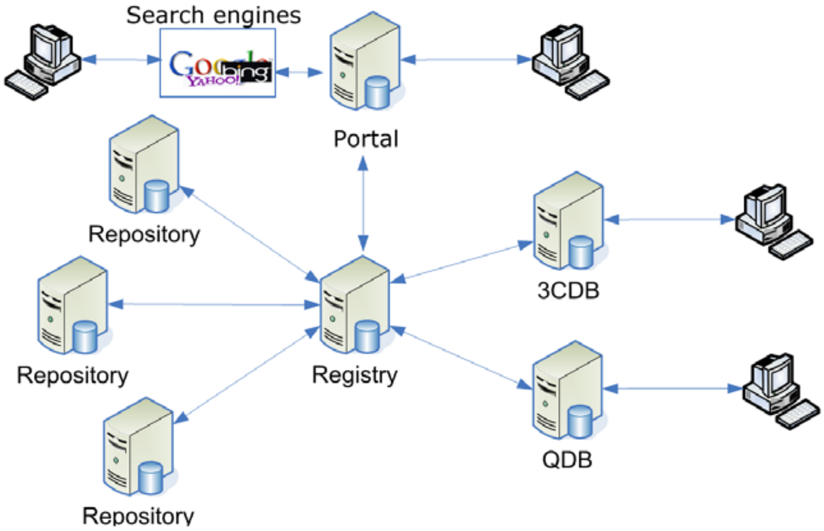


Figure 3 - The portal holding exposed to search engines

Adding value to the research process

Once researchers have found their way to the portal (or, more specifically to metadata objects displayed by the portal) it is important that they are offered:

1. Efficient tools for navigating the vast metadata collection held by the portal;
2. Efficient ways of accessing or downloading data;
3. Tools for comparative exploration.

Enabling efficient access to data is already supported to a degree by the current portal. Every portal metadata object contains a link to the object's "home page" in the harvested Nesstar server located at an archive. Depending on the archive's local access policies, the researcher may or may not access the data. In many cases it is necessary for data users register with each archive holding data of interest. Solutions for single sign-on (SSO) is being evaluated in order to improve and simplify the registration, authentication and authorisation mechanisms for CESSDA users. SSO-questions will be addressed in a section dedicated to the authentication prototype later this paper.

The portal prototype explained

In order to support the above points, a number of suggestions have been added on top of the current portal interface to improve shortcomings in the portal.

Suggestive search

Suggestive search demonstration:
http://extweb.nsd.uib.no/cessda-ppp/portal_prototype/demos/search_with_suggestions.htm
To enrich the search process, search terms will be suggested as the user types. The suggestions are based upon (at least) two sources:

1. ELSST (the multilingual thesaurus);
2. Known concepts.

The thesaurus is added as a resource to the portal already, so it is straightforward to power a suggestive search from it. Also (as of today), all search queries will (by using the thesaurus) be automatically expanded into all 9 languages and synonyms. Narrower/broader terms may be selected subsequently.

The figure below shows suggestions based upon the mentioned sources and the two characters "re" typed by the user.

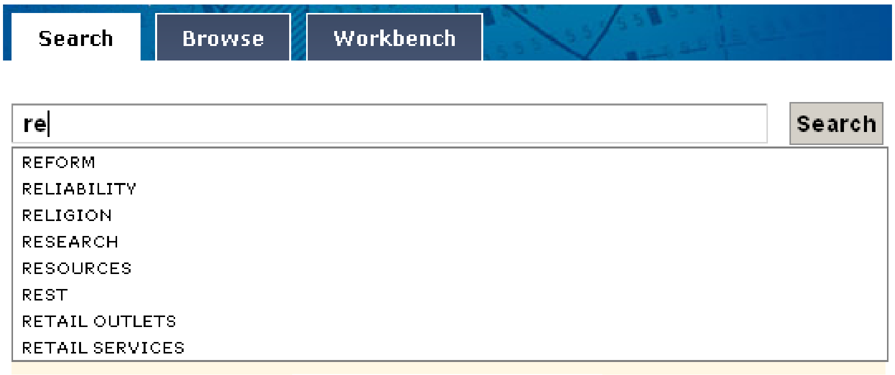


Figure 4 - Suggestive search

Faceted search filtering

Faceted filtering demonstration:
http://extweb.nsd.uib.no/cessda-ppp/portal_prototype/demos/faceted_search.htm

A problem with the current portal is the vast amount of hits for virtually any search. It is difficult to navigate hit lists with thousands of results.

Since the development of the current portal, great advances have been made in the field of searching well-structured (but vast) information spaces. The development is mainly driven by major online sales companies (Amazon.com, eBay, etc), but also by research projects like

Flamenco (http://flamenco.berkeley.edu/). One of the most important improvements in this area is "faceted filtering" that allows users to combine free-text search with clickable refine categories (or facets). Facets are orthogonal (and often hierarchical) metadata classifications that allow users to filter/narrow searches along several dimensions. From online retailers we know this as product categories and sub-categories, vendors, language (for example, for books), etc. The Flamenco-project runs an informative demonstration of this concept: find Nobel Prize Winners by faceted filtering (http://orange.sims.berkeley.edu/cgi-bin/flamenco.cgi/nobel/Flamenco).

The same approach is attempted for the portal prototype. Given that the DDI2-documented archive holdings are metadata rich and well-structured, the potential for faceted filtering on the CESSDA portal is present.

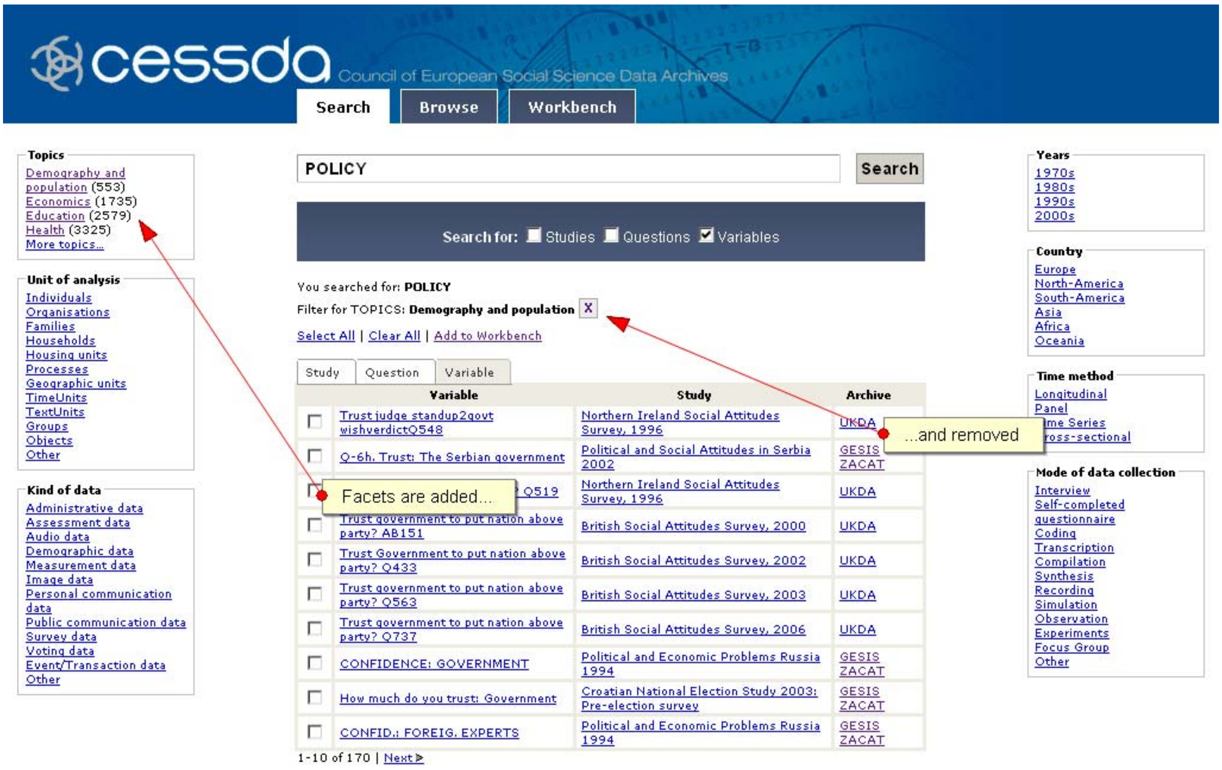
In the prototype, the following facets are chosen:

- Topics (compiled from DDI topical classification);
- Unit of analysis (compiled from DDI and controlled vocabularies developed in WP4);
- Kind of data (compiled from DDI and controlled vocabularies developed in WP4);
- Years (compiled from the index directly (the Time Periods element in DDI), and grouped by decades for space considerations);
- Country (compiled from the index directly (the Countries element in DDI), and grouped by continent for space considerations);
- Time method (compiled from DDI and controlled vocabularies developed in WP4);
- Mode of data collection (compiled from DDI and controlled vocabularies developed in WP4).

There is much to be said about details in implementing faceted search. Some facets are too large to fit on screen; some are mutually exclusive, others should allow for more than one selection per facet.

As a rule of thumb, each facet must be customised individually; in the prototype, all facets have the same simplified behaviour.

The figure below shows how facets are added and removed to enable/disable a specific filter from the search query.



POLICY

Search for: □ Studies □ Questions ☑ Variables

You searched for: **POLICY**

Filter for TOPICS: **Demography and population** ☒

Select All | Clear All | Add to Workbench

	Variable	Study	Archive
□	Trust judge standup2govt wishverdictQ548	Northern Ireland Social Attitudes Survey, 1996	UKDA
□	Q-6h. Trust: The Serbian government	Political and Social Attitudes in Serbia 2002	GESIS ZACAT
□	? Q519	Northern Ireland Social Attitudes Survey, 1996	UKDA
□	Trust government to put nation above party? AB151	British Social Attitudes Survey, 2000	UKDA
□	Trust Government to put nation above party? Q433	British Social Attitudes Survey, 2002	UKDA
□	Trust government to put nation above party? Q563	British Social Attitudes Survey, 2003	UKDA
□	Trust government to put nation above party? Q737	British Social Attitudes Survey, 2006	UKDA
□	CONFIDENCE: GOVERNMENT	Political and Economic Problems Russia 1994	GESIS ZACAT
□	How much do you trust: Government	Croatian National Election Study 2003: Pre-election survey	GESIS ZACAT
□	CONFID.: FOREIG. EXPERTS	Political and Economic Problems Russia 1994	GESIS ZACAT

1-10 of 170 | Next ▶

Figure 5 - Faceted filtering

Workbench

Workbench demonstration:
http://extweb.nsd.uib.no/cessda-ppp/portal_prototype/demos/add_to_workbench.htm

A concept missing from the current portal is a "workbench" (similar to a "shopping cart" on an online retail site) where users can add/collect metadata objects of interest as they encounter them. The workbench may be reviewed and edited at any given point. The contents of the workbench may also be downloaded or emailed to the user upon request. Inspiration for the workbench was gathered not only from online retail sites; the ODESI portal in Canada has a similar concept (albeit dubbed "My List" and not "workbench").

The prototype workbench shows the concept in its simplest form; as a placeholder for interesting information (variables, studies, questions). However, the workbench could be extended to a tool for "exploration" (WP5) and possibly a control panel for the fusion, combination, and harmonisation of complex data. Depending on implementation, the workbench could exist for the user session or more permanently (requiring login at the portal level). If in fact the portal is connected to the registry, and thereby also to all its connected services, it should be technically possible to extend the workbench functionality quite far in this direction.

Search engine exposure

One of the intended consequences of the prototype design is that all metadata instances are reachable through hyperlinks (through the "browse" tab which resembles the current portal interface, or via faceted filtering). This means that generic search engines will be able to crawl and index the entire site. An added effect to the indexing will be the metadata rich interface, where facet terms and headings, combined with the substantive content of the DDI-based metadata, will make up web-pages consisting of additional terms and keywords relevant to the metadata object in question. In other words; the context in which a given metadata object appears in the portal will improve its recovery via generic search engines.

Deliverable D12.1: Persistent identifier and versioning prototype

Background

There are several rationales for implementing solutions for persistent metadata identifiers (PID) in CESSDA. The most important one is perhaps that it is an inevitable step on the road to offering "citable data"; i.e. to enable data collections to be referenced in electronic and printed publications. Also, persistent identification, and persistent access to historical versions is a prerequisite for harmonisation infrastructure (e.g. 3CDB - see WP9), because harmonisation rules will depend on the existence and presence of historical versions of metadata objects.

There are both technical and organisational aspects to PID. On the technical side, a number of solutions and formats. URN, Handle, DOI, ARK and PURL are all examples. See the CLARIN-project's review and comparison of the different solutions here: http://www.clarin.eu/files/wg2-2-pid-doc-v4.pdf. As the SWOT-analysis below will reveal, choice of format comes with a range of associated risks.

The organisational aspects of offering PID-infrastructure should not to be underestimated. A survey carried out among CESSDA member archives during spring 2009 showed that archives have very heterogeneous policies and methods for dealing with versions/editions, and more importantly older/historical versions. Such policies need to be harmonised between local archives and then integrated into the larger PID-supporting network. The responsibility for offering sustainable PID-solutions lie with CESSDA - but the responsibility for implementing technology and procedures will lie with member archives.

The following SWOT-analysis will discuss both technical and organisational aspects:

Strengths:

- The proposed metadata model (WP9 tendered report on SOA architecture/Question Data Bank) ensures PID by design;
- Both the Dutch archive (DANS) and the closely related Inter-university Consortium for Political and Social Research (ICPSR) in the US have knowledge and experience in dealing with PID solutions and policies.

Weaknesses:

- The goal of "citable data" is not yet sufficiently strategically reviewed. Business model for PID/citable data not established;
- DANS and ICPSR use different formats and models (URN/local/granular VS DOI/outsourced/high-level);
- Organisational impact of PID is yet to be investigated; however likely to be costly;
- Technical impact of PID on local archives is yet to be investigated;
- Locally developed (and therefore heterogeneous) archive systems are generally not PID-compatible;
- DDI3-identifiers (URN-based) are not printer-friendly, possibly not acceptable to journal publishers;
- CESSDA-hosted PID-services needs 24/7 uptime, support and very high scalability. These are demands currently not met by most archive-services.

Opportunities:

- With citable data, data use (and re-use) is likely to increase, enforcing archive's position in research community;
- Not yet much competition regarding PIDs for social science data;
- The work carried out in the different initiatives (Handle, DOI, etc) is substantial and systems are constantly improving;
- A few digital archiving tools exist with partial or fully integrated support for assigning PIDs to stored objects/material, including DataVerse (http://thedata.org) and Fedora Commons (http://fedora-commons.org/), both open source tools.

Threats:

- Publisher-associations (e.g. STM - http://www.stm-assoc.org/) might not accept all types of PID-formats. A clear "winner" format has yet to emerge. It will be crucial to bet on the right horse. Choosing a format NOT accepted by STM should not be seen as an alternative;
- Cross-ref (a commercial DOI-registrar) is member of STM. (Others might be too.)
- Choosing a commercial solution (e.g. DOI) creates economic risk as payment models may change (similar problems as with Cloud-computing: see WP11);
- Commercial solutions become very expensive with high granularity/volume (e.g. variable-level PIDs);
- Other ESFRI-projects (e.g. CLARIN) are leading in this process (although these things are always hard to evaluate from outside of a project);
- National guidelines can require archives to implement other formats than CESSDA agrees upon;
- Research institutes could start hosting own “DataVerse”s, and start to offer an infrastructure for "citable data" outside archives, decreasing archive relevance in the research and publication community. However, as we shall see below, this is both a threat but also a potential opportunity.

The persistent identifier prototype: combining DataVerse and Nesstar

Following from the SWOT above, DataVerse adoption in archives can be perceived as both a threat and an opportunity. DataVerse is an open source data archiving tool designed by the Institute for Quantitative Social Science (IQSS), Harvard University.

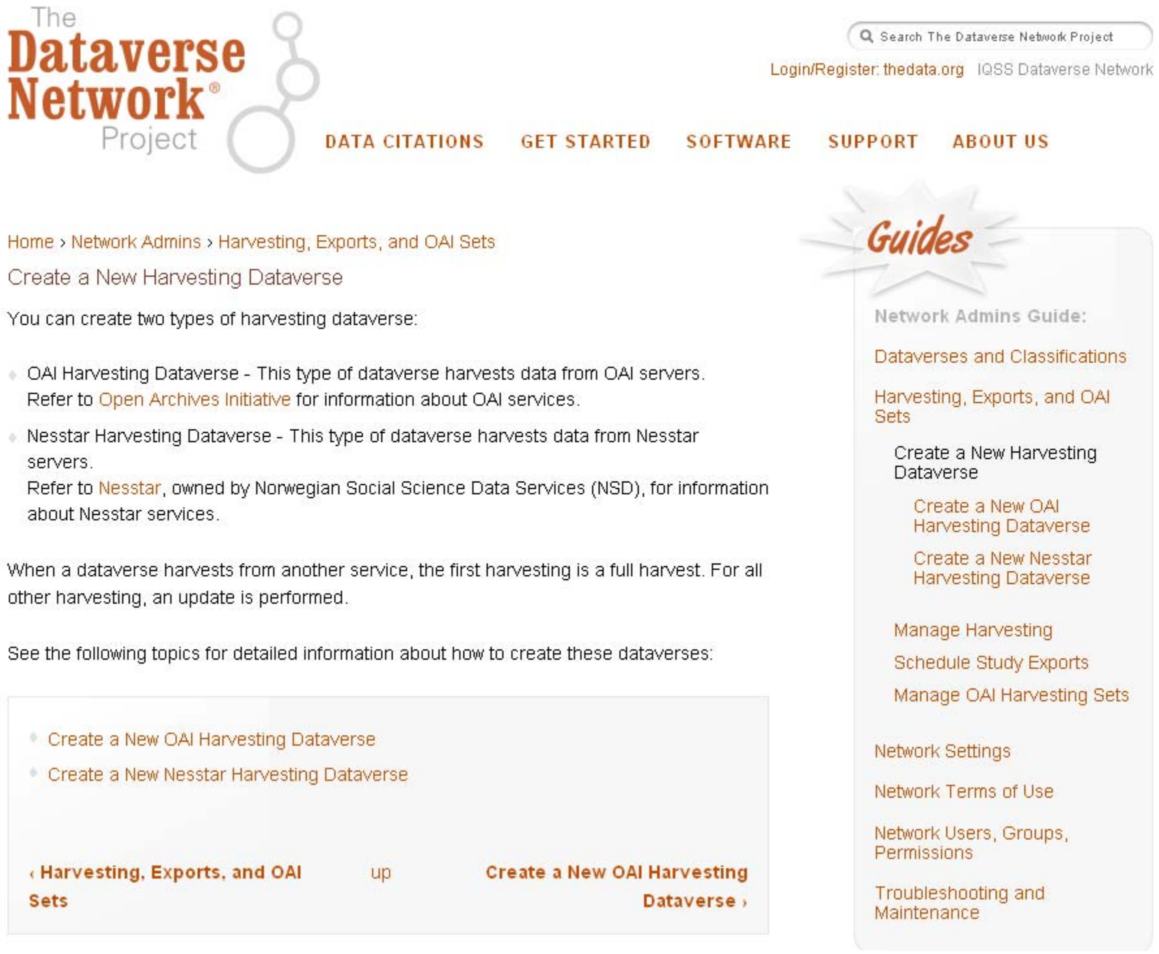
DataVerse is a tool for long-term archiving of data sets. One of the main features of DataVerse is that it supports citable data through a combination of persistent identifiers (by integration with the Handle system) and a method for creating unique "fingerprints" for data sets, DataVerse-held data may be quoted with confidence from printed and electronic publications. Depending on access policies researchers may download data through the DataVerse directly. There are also limited analysis capabilities built into the system.

For detailed descriptions of DataVerse, we refer to the project's home-page http://thedata.org/.
An interesting feature in DataVerse is its support for harvesting metadata and data from external sources through the Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH - http://www.openarchives.org/). This enables institutions (or individual researchers) to configure their DataVerse instance to harvest metadata and data from other sources for long-term preservation (and citability) in DataVerse.

Now - as most CESSDA Archives run and publish their data to Nesstar servers, it would be interesting to either add OAI-PMH-support on top of Nesstar, or to enable DataVerse to harvest Nesstar servers via other means (i.e. a custom Nesstar harvester). In March/April 2008, opportunities in this direction were discussed between the DataVerse team and the undersigned.

Even though it would be a good idea in itself to add OAI-PMH-support to Nesstar (as mentioned in the portal section above), it was decided that an acceptable intermediate solution was for NSD to provide the DataVerse team with a Nesstar-harvesting library that could be integrated in DataVerse, enabling support for harvesting Nesstar servers in addition to OAI-PMH harvesting. The library was shipped April 2008, and was subsequently included in DataVerse v1.3.

As a side-effect, DataVerse is now promoting Nesstar-harvesting on equal terms as OAI-PMH-harvesting (see facsimile from the DataVerse home page below).



The Dataverse Network® Project
Search The Dataverse Network Project
Login/Register: thedata.org
IQSS Dataverse Network
DATA CITATIONS
GET STARTED
SOFTWARE
SUPPORT
ABOUT US

Home › Network Admins › Harvesting, Exports, and OAI Sets

Create a New Harvesting Dataverse

You can create two types of harvesting dataverse:

- OAI Harvesting Dataverse - This type of dataverse harvests data from OAI servers. Refer to Open Archives Initiative for information about OAI services.
- Nesstar Harvesting Dataverse - This type of dataverse harvests data from Nesstar servers. Refer to Nesstar, owned by Norwegian Social Science Data Services (NSD), for information about Nesstar services.

When a dataverse harvests from another service, the first harvesting is a full harvest. For all other harvesting, an update is performed.

See the following topics for detailed information about how to create these dataverses:

- Create a New OAI Harvesting Dataverse
- Create a New Nesstar Harvesting Dataverse

‹ Harvesting, Exports, and OAI Sets   up   Create a New OAI Harvesting Dataverse ›

Guides
Network Admins Guide:
Dataverses and Classifications
Harvesting, Exports, and OAI Sets
Create a New Harvesting Dataverse
Create a New OAI Harvesting Dataverse
Create a New Nesstar Harvesting Dataverse
Manage Harvesting
Schedule Study Exports
Manage OAI Harvesting Sets
Network Settings
Network Terms of Use
Network Users, Groups, Permissions
Troubleshooting and Maintenance

Figure 6 - DataVerse support for Nesstar harvesting

With Nesstar-harvesting support added to DataVerse, it would be appropriate to create a prototype to test if the combination of the tools could, in effect, enable CESSDA archives to add persistency and citability to their Nesstar-held data with relatively little effort. Nesstar would continue to serve its purpose as the day-to-day channel for data publishing and dissemination, whereas DataVerse could periodically harvest snapshots from the Nesstar server and add Handle-based persistent identifiers to the snapshots. The figure below displays the envisioned prototype setup and the proposed roles of Nesstar VS DataVerse:

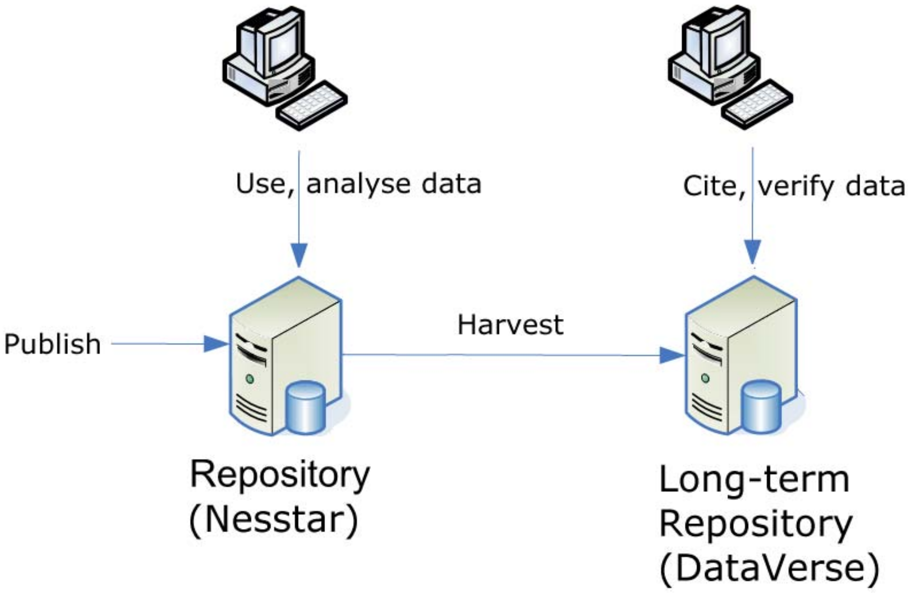


Figure 7 – DataVerse/Nesstar prototype setup

The German CESSDA member archive (GESIS) started working to set up a prototype like this early in 2009 (originally for internal testing and experimentation purposes, and not within the scope of CESSDA-PPP), and were able to successfully harvest data and metadata from a Nesstar into DataVerse.

The Nesstar server used for this prototype (http://demo.koeln.gesis.org/webview/) contains demonstration and test-data (i.e. no quality assured data).

Handle-based persistent identifiers were assigned to the harvested material dynamically. As a consequence data originally solely published into Nesstar, now would become persistent and citable from printed and electronic publications more or less automatically. Below screenshots from the resulting test-DataVerse will be added.

For documentation purposes, the link to the test DataVerse is included here: http://dvn-qa1.hmdc.harvard.edu/dvn/dv/gesis/.

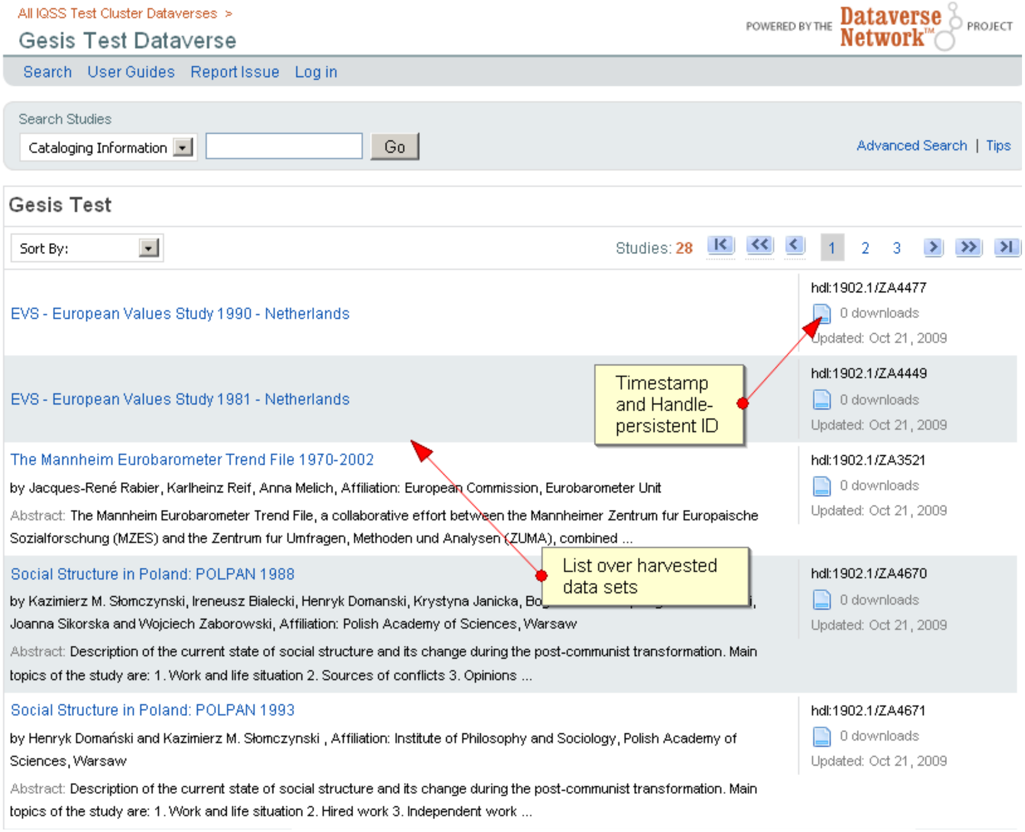
Important note/disclaimer regarding the GESIS DataVerse/Nesstar test:

The GESIS team stresses that this DataVerse is hosted at the DataVerse-"farm" at Harvard, and is volatile by definition. It may therefore go offline at any given time. Also, as the data harvested from the test-Nesstar-server are **not** public or quality assured, no citations should be made to this DataVerse.

Moreover, the GESIS team experienced a number of issues in the test process that should be included here for future reference:

1. The above links should be treated for what they are; mere test instances that may go offline at any time;
2. Some metadata are lost/changed in the harvesting process (this is likely due to minor bugs in the harvesting libraries);
3. Handles (persistent identifiers) are generated too liberally without necessary "organisational support" behind them. This means that you risk creating an abundance of persistent identifiers to data you have no intention of persisting. (For this reason, we recommend that a Nesstar-harveting DataVerse only is set up to harvest public and quality assured Nesstar servers as a minimal requirement. Nesstar servers with data of temporary character should not be harvested to DataVerse, and at least not be assigned a persistent ID);
4. Version control is not enforced by this setup. Studies altered on the Nesstar Server may (and indeed should) be assigned a new persistent identifier when harvested by DataVerse. However, the relationship to the previous version of the said study will be unclear; to DataVerse they will probably appear to be two independent studies altogether. Seemingly identical versions of one study are likely to cause confusion to researchers trying to navigate the DataVerse.

Explanatory screenshots from the test DataVerse:



Gesis Test Dataverse

Gesis Test

EVS - European Values Study 1990 - Netherlands

hdl:1902.1/ZA4477
0 downloads
Updated: Oct 21, 2009

EVS - European Values Study 1981 - Netherlands

hdl:1902.1/ZA4449
0 downloads
Updated: Oct 21, 2009

The Mannheim Eurobarometer Trend File 1970-2002

by Jacques-René Rabier, Karlheinz Reif, Anna Melich, Affiliation: European Commission, Eurobarometer Unit

Abstract: The Mannheim Eurobarometer Trend File, a collaborative effort between the Mannheimer Zentrum fur Europaische Sozialforschung (MZES) and the Zentrum fur Umfragen, Methoden und Analysen (ZUMA), combined ...

hdl:1902.1/ZA3521
0 downloads
Updated: Oct 21, 2009

Social Structure in Poland: POLPAN 1988

by Kazimierz M. Słomczynski, Ireneusz Bialecki, Henryk Domanski, Krystyna Janicka, Bo , Joanna Sikorska and Wojciech Zaborowski, Affiliation: Polish Academy of Sciences, Warsaw

Abstract: Description of the current state of social structure and its change during the post-communist transformation. Main topics of the study are: 1. Work and life situation 2. Sources of conflicts 3. Opinions ...

hdl:1902.1/ZA4670
0 downloads
Updated: Oct 21, 2009

Social Structure in Poland: POLPAN 1993

by Henryk Domański and Kazimierz M. Słomczynski , Affiliation: Institute of Philosophy and Sociology, Polish Academy of Sciences, Warsaw

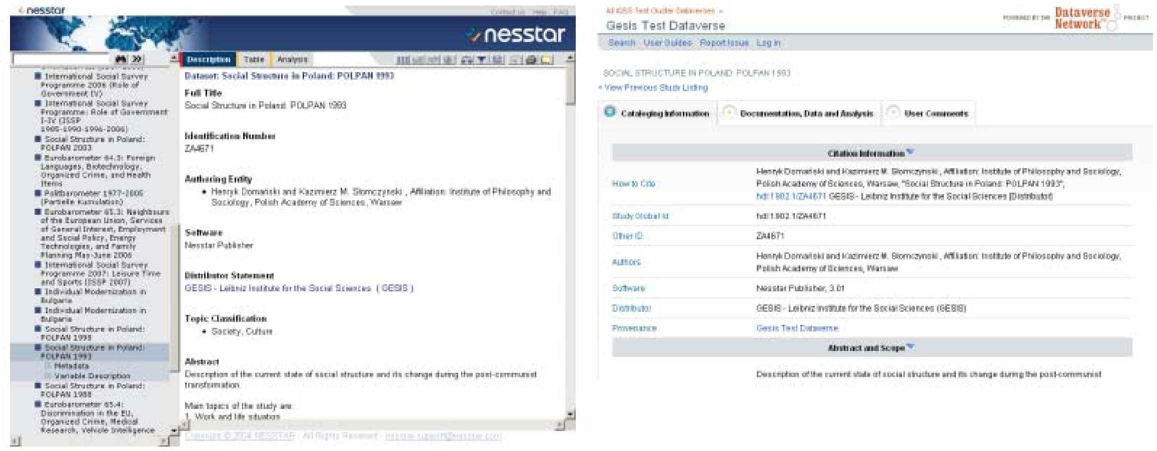
Abstract: Description of the current state of social structure and its change during the post-communist transformation. Main topics of the study are: 1. Work and life situation 2. Hired work 3. Independent work ...

hdl:1902.1/ZA4671
0 downloads
Updated: Oct 21, 2009

Timestamp and Handle-persistent ID

List over harvested data sets

Figure 8 - List of data sets in the DataVerse test instance



- International Social Survey Programme 2006 (Role of Government IV)
- International Social Survey Programme: Role of Government I-IV (ISSP 1985-1990-1996-2006)
- Social Structure in Poland: POLPAN 2003
- Eurobarometer 64.3: Foreign Languages, Biotechnology, Organized Crime, and Health Items
- Politbarometer 1977-2006 (Partielle Kumulation)
- Eurobarometer 65.3: Neighbours of the European Union, Services of General Interest, Employment and Social Policy, Energy Technologies, and Family Planning May-June 2006
- International Social Survey Programme 2007: Leisure Time and Sports (ISSP 2007)
- Individual Modernization in Bulgaria
- Individual Modernization in Bulgaria
- Social Structure in Poland: POLPAN 1998
- Social Structure in Poland: POLPAN 1993
  - Metadata
  - Variable Description
- Social Structure in Poland: POLPAN 1988
- Eurobarometer 65.4: Discrimination in the EU, Organized Crime, Medical Research, Vehicle Intelligence

Dataset: Social Structure in Poland: POLPAN 1993

Full Title
Social Structure in Poland: POLPAN 1993

Identification Number
ZA4671

Authoring Entity
- Henryk Domański and Kazimierz M. Slomczynski , Affiliation: Institute of Philosophy and Sociology, Polish Academy of Sciences, Warsaw

Software
Nesstar Publisher

Distributor Statement
GESIS - Leibniz Institute for the Social Sciences (GESIS)

Topic Classification
- Society, Culture

Abstract
Description of the current state of social structure and its change during the post-communist transformation

Main topics of the study are:
1. Work and life situation

Copyright © 2004 NESSTAR · All Rights Reserved · nesstar-support@nesstar.com

All IQSS Test Outer Dataverses »
Gesis Test Dataverse
POWERED BY THE Dataverse Network PROJECT
Search
User Guides
Report Issue
Log In

SOCIAL STRUCTURE IN POLAND: POLPAN 1993
« View Previous Study Listing

Citation Information	
How to Cite	Henryk Domański and Kazimierz M. Slomczynski , Affiliation: Institute of Philosophy and Sociology, Polish Academy of Sciences, Warsaw, "Social Structure in Poland: POLPAN 1993", hdl:1902.1/ZA4671 GESIS - Leibniz Institute for the Social Sciences [Distributor]
Study Global Id	hdl:1902.1/ZA4671
Other ID	ZA4671
Authors	Henryk Domański and Kazimierz M. Slomczynski , Affiliation: Institute of Philosophy and Sociology, Polish Academy of Sciences, Warsaw
Software	Nesstar Publisher, 3.01
Distributor	GESIS - Leibniz Institute for the Social Sciences (GESIS)
Provenance	Gesis Test Dataverse

Abstract and Scope

Description of the current state of social structure and its change during the post-communist

Nesstar version

Harvested DataVerse version

Figure 9 - Listing of the study instance "Social Structure in Poland: POLPAN 1993" in Nesstar and in DataVerse. The DataVerse version is harvested from the Nesstar one.

Deliverable D12.2: Internal project database for classifications and conversion information

The internal project database and corresponding software application is not available online. Instead, the application (with screenshots) and associated database structure is described in the following reports:

http://www.cessda.org/ppp/wp09/WP9_T2_Aug09.pdf
http://www.cessda.org/ppp/wp09/WP9_Database_Models_Aug09.pdf

An offline version is delivered and can be requested from GESIS.

In the first report, the D12.2 prototype is described at length, including screen-shots from the prototype. The second one describes the database structure of the prototype.

For this deliverable we will therefore just add the conclusion from the first report.

Conclusion quoted from the WP9-report:

> *"As part of the general work being done within WP9, this document has aimed to describe the basic functional and technical specifications of the proposed Constructs, Classifications and Conversions Database and the corresponding software application. Our analysis suggests that the three-layer or working step model of a harmonization project that has been presented is helpful to the researcher in the process of building and documenting a harmonization routine.*
> *CCCDB holds these harmonization routines as well as the target variable data and the data required for the graphical representation of the three layers created in the process of building the actual routine. This would help researchers to better*

> *understand these routines and maybe reuse them by transforming another source variable into a target variable. Thus, it is evident that reusability of this routine is supported by the docu-mentation of the complete harmonization project.*
>
> *Some basic decisions should be taken in collaboration with the other WPs, as well as with the manufacturers of NESSTAR (or any other data repository used in CESSDA). It seems that data dissemination systems that incorporate routines for manipulating data sets and provide a language for building these routines would be a helpful tool for CCCDB to use when creating the actual harmonization routine. Moreover, it would provide the ability to perform the harmonization routine regardless of the format the source data set is in.*
>
> *The general CCCDB architecture could easily stand within the proposal of Gregory et al., 2009 for the general CESSDA architecture. Even if this architecture is not adopted, CCCDB could also stand on its own, since the proposed CCCDB architecture is unaffected from the overall architecture. It would however be necessary to have a decision on the infrastructure architecture before starting to implement CCCDB, since there must be made certain changes to the database and to the web services. This flexible, service-based architecture also makes CCCDB available to other CESSDA projects and software applications.*
>
> *The benefits of CCCDB would be great for CESSDA's user community, since the researchers will have at their disposal an assisting tool to transform and interpret already published studies as well as to easily build new cross-national studies.*
>
> *Finally, we must point out that even though the current purpose of CCCDB is to provide a context for building harmonization projects, the architecture, software tools and the specifications proposed make it possible for further expansion and use of the database. More functionality can therefore be added in the future without having to change anything significant in the technical, functional and security requirements described in this document."*

Deliverable D12.3: Authentication prototype

Background

As researchers increasingly want to move painlessly between data sources or combine data from different sources, it is necessary to streamline and harmonise authentication and authorisation mechanisms across CESSDA archives. The goal is to achieve Single Sign-On (SSO), so researchers remain logged in (i.e. authenticated) when moving between services located in different archives.

Authentication is planned to be mostly handled in cooperation between national Identity Providers (IDPs) and CESSDA service-providers, relying on the SAML/Shibboleth platform, which recently has gained momentum in Europe.

Authenticated users are not however necessarily authorised to view/operate on a given object. Authorisation rules therefore need to be built on top of the authentication systems. It is recommended that CESSDA develops a taxonomy/controlled vocabulary of roles in such a system, in order to achieve harmonised authorisation mechanisms across the organisation.

SWOT-analysis:

Strengths:

- A number of CESSDA are already part of national SAML-based federations. National initiatives outside CESSDA carry most costs and investments;
- Shibboleth has been shown to work in combination with Nesstar 3.5 (UKDA/ESDS);
- CESSDA shows willingness and ability to harmonise authentication and authorisation.

Weaknesses:

- Not all archives are part of adequate national federations (not all countries even have one);
- We cannot expect all CESSDA end-users to have a Shibboleth ID from a trusted ID-provider;
- Harmonisation of authentication and authorisation across CESSDA will require multilateral negotiations on fairly detailed rules;
- Attributes forwarded from ID-providers (IDPs) to Service-providers (SPs) are likely to be very limited (lowest common denominator). Therefore, auxiliary authentication/registration procedures must be added to the CESSDA service environment;
- Authentication is just one part of the solution. Granular authorisation mechanisms supporting access control policies on individual objects will be required. In order for authorisation to work seamlessly along with the SSO-pattern, authorisation rules (and object classifications) need harmonisation across CESSDA members;
- SSO will set the standard for simplified user-interaction with CESSDA holdings. Authorisation barriers (caused by e.g. national law) stopping data collection at "the last step" could become unpopular.

Opportunities:

- Lots of EU-initiatives (e.g. in TERENA) are going on creating SSO-federations and technical solutions;
- Europe seems to consolidate on Shibboleth/SAML;
- Technologies for managing authorisation rules across big networks seem to emerge, although still not on the same level as authentication.

Threats:

- Shibboleth has taken off in Europe, but OpenID is more popular elsewhere. (Note: OpenID-support would come with a range of threats in itself and should be subject to a separate SWOT-analysis);
- Investing in Shibboleth-based infrastructure will cause dependencies on technical decisions taken beyond CESSDA's control. Timing of upgrades, migration to new versions/schemes will likely be forced upon CESSDA.

The authentication prototype

The prototype is available at the following URL:
https://shib-portal-cessda.data-archive.ac.uk/Cessda-Test/

To simplify developments, the "protected resource" used in the prototype is just a simple web-page. Knowing that Shibboleth integration across distributed data services works well for UKDA services in Essex e.g. http://esds.ac.uk/newRegistration/newLogin.asp, it was regarded as sufficient to protect a simpler "object" for the purpose of this prototype.

The question was if it was possible to create a multi-federation Shibboleth protected access to "resources/services" held at the UKDA, to users registered at the UKDA, NSD (Norwegian Social Science Data Services - University of Bergen) and The University of Essex. Note that UKDA and University of Essex belong to the same Shibboleth federation; the UK Access Management Federation for Education and Research (UKAMF). NSD users, on the other hand, belong to the Norwegian counterpart of UKAMF; Felles Elektronisk IDEntitet (Feide) (Eng: Common Electronic Identity). In sum, the prototype spanned two national Shibboleth-based federations.

A team from UKDA was assigned the task to implement the prototype, given their previous experience with Shibboleth and UKAMF-configurations in the UKDA.

The work was mainly carried out between August and November 2009 (with a final breakthrough on 4 November), and required much correspondence between the developers and technicians from Feide.

Implementation details and guidelines worked out during and after implementation are published here for future reference and guidance when the greater federation are to be expanded beyond the two in this prototype:
http://extweb.nsd.uib.no/cessda-ppp/sso-prototype/

Deliverable D12.5: Thesaurus management software

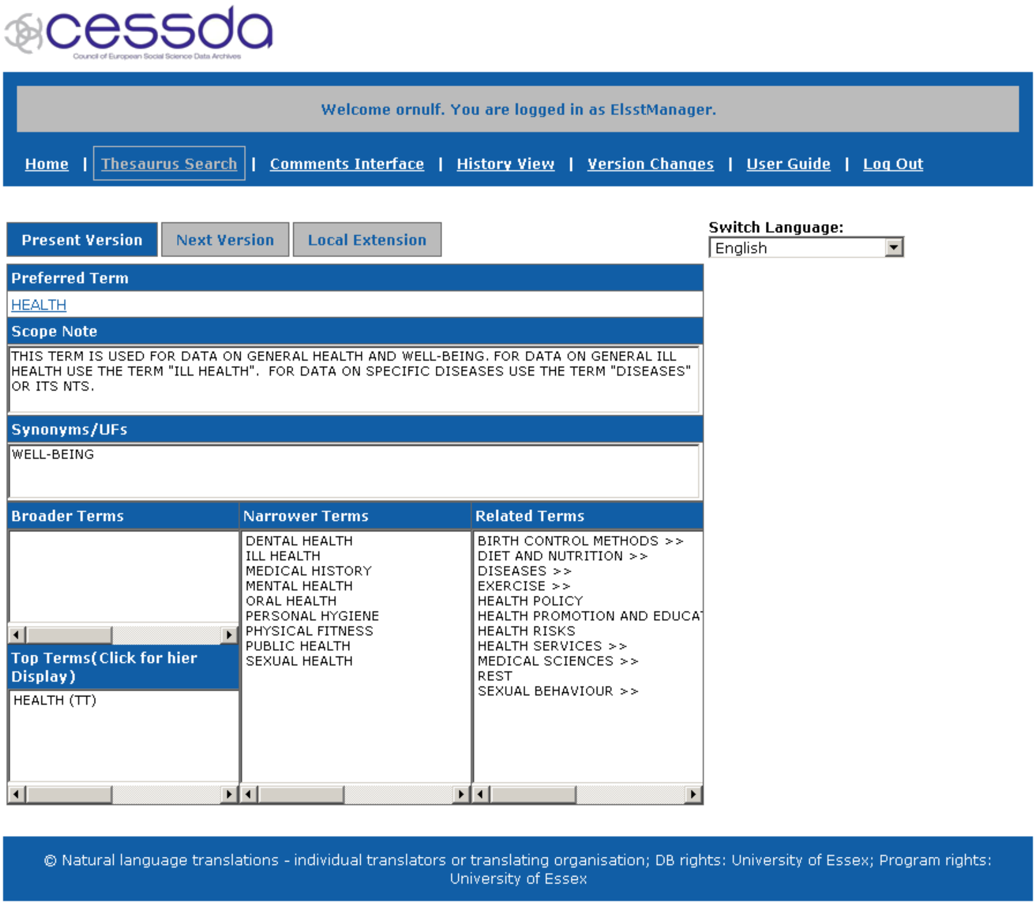
The thesaurus management software is available at the following URL:
http://elsst.esds.ac.uk/

The thesaurus management software development was coordinated under WP4, by Taina Jääskeläinen (FSD) and Ken Miller (UKDA).

The functional specification was published in May 2008:
http://www.cessda.org/ppp/wp04/D4_1_functional_specifications_for_thesaurus_20090603.pdf

The actual implementation gained much momentum during late summer 2009, and from August 2009 onwards, a team from UKDA initiated a series of bi-weekly releases with members of the WP4 team to discuss usability and possible future development.

The final version went live 16 December 2009, and is a workable, first version, online tool for managing content of the ELSST thesaurus.


cessda
Council of European Social Science Data Archives
Home | Thesaurus Search | Comments Interface | History View | Version Changes | User Guide | Log Out
Present Version
Next Version
Local Extension

Welcome ornulf. You are logged in as ElsstManager.

Switch Language: English

Preferred Term

HEALTH

Scope Note

THIS TERM IS USED FOR DATA ON GENERAL HEALTH AND WELL-BEING. FOR DATA ON GENERAL ILL HEALTH USE THE TERM "ILL HEALTH". FOR DATA ON SPECIFIC DISEASES USE THE TERM "DISEASES" OR ITS NTS.

Synonyms/UFs

WELL-BEING

Broader Terms	Narrower Terms	Related Terms
	DENTAL HEALTH	BIRTH CONTROL METHODS >>
	ILL HEALTH	DIET AND NUTRITION >>
	MEDICAL HISTORY	DISEASES >>
	MENTAL HEALTH	EXERCISE >>
	ORAL HEALTH	HEALTH POLICY
	PERSONAL HYGIENE	HEALTH PROMOTION AND EDUCAT
	PHYSICAL FITNESS	HEALTH RISKS
	PUBLIC HEALTH	HEALTH SERVICES >>
	SEXUAL HEALTH	MEDICAL SCIENCES >>
		REST
		SEXUAL BEHAVIOUR >>

Top Terms(Click for hier Display)

HEALTH (TT)

© Natural language translations - individual translators or translating organisation; DB rights: University of Essex; Program rights: University of Essex

Figure 10 - Screenshot from the thesaurus management tool