Project N°: 262608

Acronym: Data without Boundaries

Deliverable D12.5

(Roadmap describing varying levels of compliance with metadata model)

Work Package 12

(Implementing Improved Resource Discovery for OS Data)

Reporting Period:	**From: Month 18**	**To: Month 48**
Project Start Date:	**1st May 2011**	**Duration: 48 Months**
Date of Issue of Deliverable:	**1st October 2013**	
Document Prepared by:	**Partner 6, 8, 16, 18, 19, 25**	**NSD, RODA, MT, UEssex, KNAW-DANS, CED**

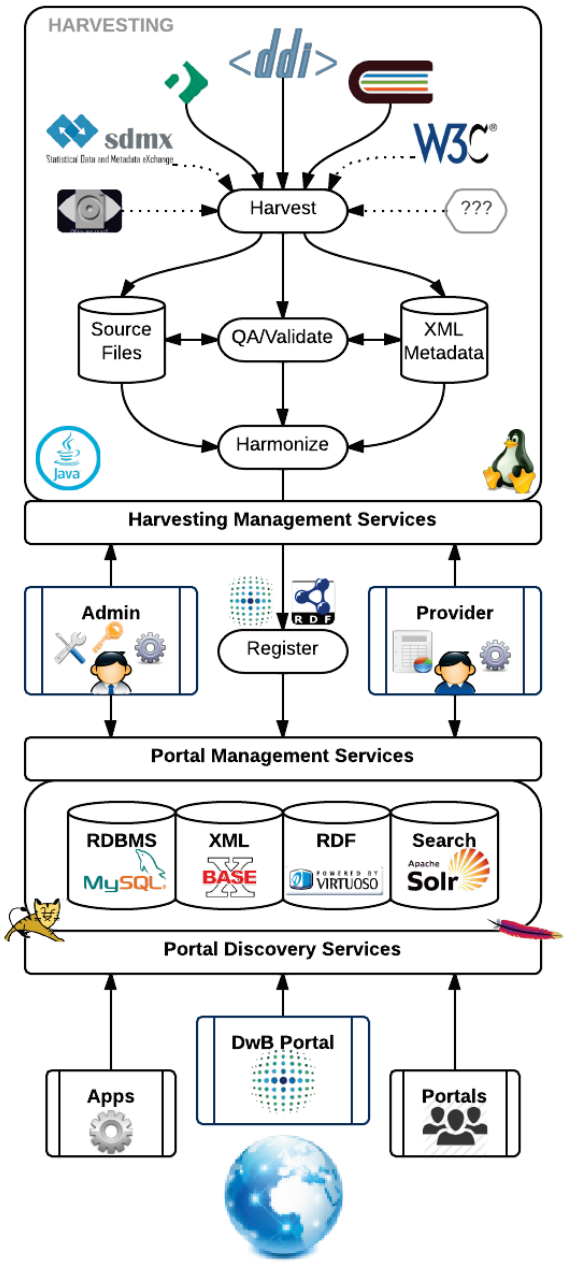
Combination of CP & CSA project funded by the European Community
Under the programme "FP7 - SP4 Capacities"
Priority 1.1.3: European Social Science Data Archives and remote access to Official Statistics

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 262608

Table of Contents

0. Overview

The diagram on the right provides a high level overview of the architecture of the DwB prototype for a CESSDA Resource Discovery Portal being implemented under WP12. It has been designed based on inputs from other work packages (particularly WP8 and WP5), consultations with domain experts, community trends and best practices, and internal research.



It is composed of the following parts:

- **The Harvesting Framework** and related tools to facilitate the retrieval and ingestion of metadata in various formats from a variety of participating providers.
- **An Administrator Dashboard** and underlying management services to configure and monitor the infrastructure, and orchestrate the various tasks
- **A Provider Portal** enabling participating agencies to gain insights on the information present in the system. This includes reports (quality assurance, usage, harvesting) and tools to control their profile, visibility of metadata, or profile information.
- **The Storage Framework** (databases) where the information is hosted and indexed in various shapes and form (Relational, XML, RDF) in order to support the search, discovery and other services built on top of that. The metadata is based on commonly used standards and the DwB Metadata Model[1].
- **The Discovery Services** exposing the platform and information to the outside world, and enabling integration in applications or web sites, along with the Resource Discovery Portal user interface

The implementation leverages several open sources technologies and packages, chosen for their robustness. The DWB-RDP prototype implementation is woven around these in an innovative fashion to deliver an enterprise grade scalable solution.

This document focuses on a Roadmap and the future and is part 5/5 of the project deliverable.

[1] See D8.2 "Metadata Model", available here

1. Introduction[2]

Data is one of the most important components necessary for a science-based understanding of society. However, empirical comparative social research in Europe has been hampered by a fragmentation of the scientific information space, so it is difficult to gain an adequate overview over data resources that are collected, available or relevant for research purposes. The European reality is that data (and its derivatives, information and knowledge) often are scattered in space and divided by languages as well as legal and institutional barriers. This state of affairs has influenced negatively the development of a thoroughly comparative and cumulative research process that would integrate and nurture the entire European Research Area. There are also major differences in how the answers to such challenges have been formulated until now, in terms of centralization and establishment of large-scale European-wide institutions, versus the potential that modern communication technologies offer.

This does not mean that data is a scarce resource in Europe, but information about the existence and availability of data is not as available as it should be. Over 50 well-developed official national statistical systems combined with a variety of both academically and commercially driven data-gathering programs and activities are producing a wealth of data and information about various aspects of the European societies, at an increasing speed. However, these institutions have little focus on systematic data library functions and data dissemination to the general research community. This contrasts with the social science data archives, which are established to secure the longer-term preservation and efficient dissemination of large parts of the available resources for secondary use. But these archival institutions do not to any significant degree collect the data themselves, concerning themselves with data that others have collected. We get an unnecessary severe division between production and dissemination, leading to two separate systems that have had difficulties solving problems of cooperation and integration. Resource discovery across these various types of data repositories is one area that presently is not optimally organized. Whereas the data archives are model institutions for data dissemination, the remit of data collectors may not include dissemination, so their data may not be optimally or adequately documented and often are not saved or intended for future alternative use.

One potential answer is to focus on the power of emerging information technologies, to encourage communication, sharing, integration and collaboration across spatially dispersed but scientifically related communities. However, such a virtual infrastructure will only make sense if it connects existing and well functioning providers of content and services; and it will only survive if it meets the demands of its users. What is required is a Semantic Web extension of the ordinary web, where information is given well-defined meaning and where the documented resources are made available over the internet.

If “sharing” is the most important single keyword characterizing a true grid, the key to realizing the benefits of both grid computing and the Semantic Web is standardization. Standardization facilitates integration of computer applications so that the diverse resources that make up a modern computing environment can be discovered, accessed, allocated, monitored, and in general managed as single virtual systems – even when provided by different vendors or operated by different organizations. The requirement is standardization of metadata, at the semantic, structural and syntactic level, to facilitate interoperability, and technical systems built according to the best practices for modularity, openness and exchange.

[2] See DwB Application, Part B 1.1 Concepts and objectives

The scenario developed in this report concerns the production of science-based knowledge as the main aim, and access to relevant high-quality data as the most important means towards that aim, and it points at a long list of stakeholders in such a project:

Data users need ability to find and analyze the data

Data providers need to collect, document, quality-check and publish data

System developers need justifications and ability to identify technological needs for tools

System maintainers need resources and legitimacy for building and maintaining systems

Data infrastructures need legitimacy and support for storing data as a fundamental resource for science and knowledge production

Decision makers need access to cumulated relevant science-based knowledge

This indicates a formidable system, where the discovery part is sketched out in the "Overview" above and in more detail described in WP12 deliverables D12.1 - D12.4. These 4 reports describe how to develop and maintain a common resource discovery portal across a wide variety of data repositories, delivering the information needed by researchers that are looking for data. It represents a combination of principles, rules, best practices, standards and technologies with much wider relevance than just for developing a discovery portal and is based on ideas about how to work rationally towards a common aim. The "product" that is moved through the process from initial production to ultimate use is a heterogeneous combination of data, the necessary or relevant metadata and the related technologies, all this wrapped in the context of a moving legal, organizational and technical environment.

The purpose of the present report is to put these elements together as a foundation for a strategic plan and as a roadmap for a full-scale implementation of the products and technologies outlined in the 4 former deliverables.

In this report we use the term "DwB RDP" to label the DwB WP12 developed prototype for a (potential CESSDA) Resource Discovery Portal.

2. Project Roadmap / Beyond DwB

1.1 Sustainability - Integration of legal, organizational and technical aspects to enhance maintenance and access

We have indicated that the "total product" is a combination of data, its metadata and the software and other technicalities involved. To a large extent, the DWB-RDP as a discovery tool does not consider the data and focuses mainly on metadata and technologies to handle it. The indicators for a well-designed and well developed discovery portal are therefore also mainly concerned with metadata and technical systems.

Building a common data discovery portal across NSIs and data archives requires a strong focus on metadata (including its organization, content, development and maintenance). Deliverable 8.3[3] from work package 8 discusses this complex set of problems more thoroughly as a gradual merging of two perspectives; this could also be regarded as the main rationale behind the DWB-RDP development and architecture.

1. Good metadata is important for practical development of efficient work processes and efficient production of high quality statistics and traditionally this has been the main justification for the NSIs focus on organized metadata.
2. At the same time quality metadata also facilitates easier fulfillment of an expanded NSI mandate, as NSIs now face the requirement to supply social science research with more quality data. Metadata may more specifically drive many work processes, where the traditional NSI approaches were not the most efficient.

"From the perspective of an NSI, the business case for engaging with the European data archives is a straightforward one: if infrastructure initiatives such as DwB provide the mechanism for providing access to secure microdata, and can leverage the archives' experience and expertise in this area, on both the technical and legal fronts, then the NSI can achieve its own emerging goals with greater ease, and at less expense. Key to this is the fact that the NSIs are in early stages of adopting the same metadata standards for describing their statistical production as the archives use for their own internal purposes" (D8.3, p 11)

The WP12 DWB-RDP work outlines a large but decentralized system. It potentially invites integration of production, maintenance, discovery and use of data resources; however it reserves for itself only a limited role as a starting point for resource discovery. The WP12 internal work plan originally specified:

- The DWB-RDP as such will supply a catalogue of research-relevant content; it is not supposed to function as a data repository itself.
- For access to data resources, the DWB-RDP should direct users to provider-level services.
- The focus is on discovery and information about data resources. Important information should if possible also include information about access conditions.
- The way the DWB-RDP is developed with focus on structured and harmonized metadata will greatly enhance possibilities to develop other services of value for data providers.

[3] See report available here.

- The ambition is to establish the DWB-RDP as a reliable, high-quality, rich and updated service, it should be closely tied up with continuous production processes.
- A certain minimum level of information or metadata for every single resource is necessary for a data provider to publish the availability of a data resource, this necessary minimum for the DWB-RDP is specified as:
  - an abstract, a summary description at resource level,
  - specification of the temporal and spatial coverage of the data,
  - variable information, details of content to facilitate judgment of relevance for researchers, and
  - question text, actual instrument used to collect the data.
- The minimum however should not restrict more complete data documentation where available. For a more thorough discussion of the problems, see deliverable D8.2, pp 28-29 (Table 1).
- The DWB-RDP will be a point that harvests metadata from service providers and other relevant sources and organizes and stores this information. Full automation of such processes is regarded as necessary to keep the content updated with pointers to valid provider-level information.
- To be able to harvest from a decentralized set of providers and use the information for the intended purpose requires both metadata standardization, mapping between standards and systems and common service Application Programming Interfaces (APIs).
- In the longer run, this will require and stimulate content harmonization.
- The DWB-RDP will support multilingual resource discovery.
- To promote such a development, “Provider Services”, i.e. tools and recommendations, should be made available to help providers inspect their metadata and run quality checks and improvements when needed.
- Providers also themselves should control the actual presentation of their holdings in the catalogue.
- The wider use of such technologies invites integration of resource production, resource discovery and resource use.

NSIs experience a growing pressure to disseminate their data to a wider user community and to do that in a more analysis-oriented form. Adjusting to and serving this need may pay off also in more efficient and timely production of data products.

Sustainability is usually about endurance of systems and processes and requires a stable institutional framework for the work processes detailed above. In this context sustainability represent at least two different facets. This portal needs a long-term perspective on supply and maintenance of the metadata input from data providers, and it need a home and a similar perspective on system maintenance.

For the official statistical system, the legal and organizational aspects of this are well developed and NSIs across Europe presently collect a formidable amount of data of great value to social science investigation, both for descriptive and more causal oriented research. When it comes to the technical aspects of building a discovery service for such material however there is a need of greater standardization and process development. It all revolves around metadata and touches on questions of what and how metadata is produced, organized, standardized, represented, stored and communicated.

Developing a discovery tool on its own may not be enough to justify investing in such work for a NSI that has to prioritize many tasks in order of relative importance. But on this topic there are considerable additional benefits to be added as systematic work will pay off in different directions as well as in

economic terms. Metadata is the element that facilitates or links a long list of relevant applications and metadata may and should be re-used across these various applications. On the data production side, it facilitates the more efficient development of data-collection instruments and automated documentation of files; while on the data use side, it allows more automation of data repository administration, data discovery and enhanced visibility of services, on-the-fly production of dissemination-ready user files, as well as easier communication and transport of data between systems and users. In short *every data provider may concentrate on developing its data and not bother about discovery and dissemination problems*.

This can facilitate a much closer feedback loop between data providers and data users and enhance relevance and quality of data production, while supporting and enhancing legitimacy for data collecting activities from the research community. In this picture, the DWB-RDP is intended to function as a tool-kit and a facilitator. If metadata are developed to the level envisioned here, it will open a plethora of additional opportunities and it is important that all data providers organise their work on metadata in such a long-term and multi-faceted perspective.

An integrated resource discovery portal that would include OS metadata is the final manifestation of a set of challenging issues: methodologies to access a variety of metadata sources in a timely manner, sustainability and maintenance of the access points, identification and management of the disparate and heterogeneous formats of metadata, and ensuring that all is delivered to the researcher in a clear and usable form. This also entails a common interest between many types of data providers and a resourceful central point that maintains or coordinates the relevant and necessary standards, tools and software on behalf of the research community. Development and maintenance of the DWB-RDP is solidly put on the work plan for the Consortium of European Social Science Data Archives (CESSDA) as a guarantee for future sustainability of the technological aspects. However, more than anything this also requires sustainability and appropriate maintenance of the access points.

1.2 Reusability

Whilst it was never realistic to expect that this work package would or could produce a completed and fully implemented product (with all that implies) given the timescales and available budget, it was always the intention to deliver something that could be the basis for a future product if required.

To that end, design modularity and adherence to relevant software development standards have been the guiding principles, but just saying that is not enough to quantify the general reusability of the DWB-RDP software developed so far. For that purpose, we have adopted the NASA Reuse Readiness Levels (RRL) in order to help give a more objective measure of reusability of the products made available, coupled with a means to identifying areas for further work if or when a product roadmap is put in place in a more implementation-oriented follow-up. For general background on RRL, see https://earthdata.nasa.gov/esdswg/software-reuse/reuse-readiness-levels-rrls

Reuse Readiness Levels

There are nine RRL criteria, plus an overall level (which gives a general statement of reusability). Each criterion can be marked zero to nine, and the overall level is scored at zero to nine as an average result.

Zero indicates that the criterion is not applicable and nine indicates maximum reusability.
The criteria for reuse readiness evaluation are:

- Documentation;
- Extensibility;
- Intellectual Property Issues;
- Modularity;
- Packaging;
- Portability;
- Standards Compliance;
- Support;
- Verification and Testing.

Taken as a whole, the toolchain (from metadata harvesting to discovery) has been rated as per the following Criterion/Score table.

Criterion	**Score**	**Description**
Documentation	3	Basic external documentation for sophisticated users are available
Extensibility	5	Consideration for future extensibility designed into the system for a moderate range of application contexts; extensibility approach defined and at least partially documented
Intellectual Property Issues	5	Agreement on ownership, limited reuse rights, and recommended citation
Modularity	7	Clear delineations of specific and reusable components
Packaging	1	Software or executable available only, no packaging
Portability	5	The software is moderately portable
Standards Compliance	5	Standards compliance with some testing
Support	2	Minimal support available
Verification and Testing	2	Software application formulated and unit testing performed
Overall	4	Reuse is possible; the software might be reused by most users with some effort, cost, and risk.

Citation

To cite the DwB prototype Resource Discovery Portal in publications, use:

Norsk Samfunnsvitenskapelig Datatjeneste, Universitaeta Din Bucuresti, Metadata Technology Ltd, University Of Essex, Koninklijke Nederlandse Akademie Van Wetenschappen, Centre D Estudis Demografics (2015). DWB-RDP: Data without Boundaries prototype for a CESSDA Resource Discovery Portal. URL: http://dwb-dev.nsd.uib.no/portal.

A BibTex entry:

```
@Manual{,
 title       = {DWB-RDP: Data without Boundaries prototype for a CESSDA Resource Discovery Portal},
 author      ={Norsk Samfunnsvitenskapelig Datatjeneste, Universitaeta Din Bucuresti, Metadata Technology Ltd, University Of Essex, Koninklijke Nederlandse Akademie Van Wetenschappen, Centre D Estudis Demografics
},
 organization = {Norsk Samfunnsvitenskapelig Datatjeneste, Universitaeta Din Bucuresti, Metadata Technology Ltd, University Of Essex, Koninklijke Nederlandse Akademie Van Wetenschappen, Centre D Estudis Demografics
},
 year        = 2015,
 url         = {http://dwb-dev.nsd.uib.no/portal}
}
```

License for source code

The license for the source code of Data without Boundaries prototype for a CESSDA Resource Discovery Portal is European Union Public Licence v1.1. More information about this license can be found here.

1.3 Technology / Service / Content Readiness Level

Achievements & Limitations

In terms of achieving its research objectives, we find this work package to be a success:

- A solid suite of technologies have been identified and combined to deliver an corporate-grade portal platform for preparing, harvesting, harmonizing, storing, and accessing metadata
- Through the development of a model extending beyond the DDI-Discovery specification, drawing from other WP efforts and global/community best practices, we have produced a database to support the users' discovery needs
- Initial metadata has been harvested from selected providers that expressed interest in participating in this effort and had the minimal metadata ready for use
- An initial suite of discovery services has been implemented, driven by user stories and requirements; and has been implemented as a proof of concept / starting point
- A front-end web-based portal has been developed on top of the service architecture

Using the European Commission Technology Readiness Levels[4] as a frame of reference, we grade our project as TRL Level 7.

- TRL 1. basic principles observed
- TRL 2. technology concept formulated
- TRL 3. experimental proof of concept
- TRL 4. technology validated in lab
- TRL 5. technology validated in relevant environment (industrially relevant environment in the case of key enabling technologies)
- TRL 6. technology demonstrated in relevant environment (industrially relevant environment in the case of key enabling technologies)
- **TRL 7. system prototype demonstration in operational environment**
- TRL 8. system complete and qualified
- TRL 9. actual system proven in operational environment (competitive manufacturing in the case of key enabling technologies; or in space)

While successful as an initial effort, significant challenges emerged along the way, limiting our ability to further develop the DWB-RDP and the underlying infrastructure. In particular:

- Limited availability of public metadata in terms of public accessibility or comprehensive coverage of a wide range of elements
- Challenges in harmonizing metadata across sources or languages, in particular in terms of mappings into controlled vocabularies for faceted searching
- Project limitation in terms of development resources

The following sections provide details and ideas around how these could be addressed in the future.

1.4 Challenges & Enhancements

The main remaining challenges are closely linked to points discussed above. This also indicates that they are not particularly technical, but more about content, its availability, form and general usefulness.

There are many differences in how NSIs and data archives function and how they organize work and products. In that sense, two points are of noticeable interest:

- Time plays a different role in organization of data. NSIs work more with a longitudinal perspective while academic research is more cross-sectional in analytic approach. The “study” concept stems from such a cross-sectional analytic approach, while indicators and statistical concept registries are a statistical need. NSIs are not to the same degree focused on documentation of complete files or complex studies covering a broader topic.
- Traditionally NSIs are focused on production, and the end result is published as tables ready for consumption. Aggregation into tables is both an analytic tool and an anonymization technique. In contrast, research is more focused on availability and dissemination of what exists and the need to make data available so that researchers may conduct their own analysis.

This does not reflect significant differences in actual data, but makes for a distinct difference in the metadata needed.

[4] http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf

1.5 Metadata Availability

WP12 has contacted selected NSIs during the project period, with the intention of harvesting and ingesting official NSI metadata into the DWB-RDP. These contacts resulted in several constructive conversations, and WP12 picked up signs that the ideas behind the Discovery Portal would be worthwhile for NSIs to follow up on.

However, the following findings indicate that most NSIs are still in the early stages of a full-scale adoption of metadata standardization and metadata dissemination:

- General availability of official metadata (from NSIs) is sparse and currently not standardized.
- WP12 has struggled to identify appropriate (and preferably machine-readable) sources of metadata and to get commitment from potential providers.
- In many cases, informal access to internal metadata or prototypes has been suggested (albeit not followed through), rather than availability through an official service
- Metadata is not yet an integral part of the data production processes.

To improve the current state, the NSI community has in recent years established the HLG-initiative and the GSBPM/GSIM/CSPA[5] line of model products and other initiatives that points towards metadata-driven data management, reflecting also for NSIs:

- A need for guidelines for minimal/recommended/comprehensive content, and
- A commitment as data providers to produce such metadata and to integrate its use into work processes.

The DwB prototype Resource Discovery Portal is aligned with CSPA in terms of supporting business processes and delivering services, and WP12 welcomes the aforementioned initiatives as a sensible means to alleviate the challenges we have identified during the implementation of this work package.

There are several arguments or incentives that data providers may use to justify a strengthening of metadata work:

- It will greatly enhance the efficiency of production processes with enhanced timeliness of products.
- Automation, as standard scripts can be used to populate the metadata template.
- Data quality, as the use of controlled vocabularies and automated checks can ensure consistency, compliance and completeness.

Furthermore, there is a need for new (meta) data management tools. This is beyond the scope of this work package but in a nutshell:

- Work with data and metadata is using old technology and inappropriate software. Statistical analysis packages are meant for statistical analysis, they are not geared for production; they were developed in the 60's and the 70's, and have not evolved much.
- New technologies - such as R, Python, Java, etc. - have emerged and they are a much better fit for modern IT infrastructure (service oriented, XML, cloud, big data, etc.).

[5] http://www1.unece.org/stat/platform/display/CSPA/Common+Statistical+Production+Architecture+Home

Metadata Quality

- Even when available, it frequently happens that metadata lack of consistency across sources. Integrating metadata into work processes will enhance the quality of both data and metadata.
- Significant efforts must be made by data producers and archives to improve the overall quality and consistency of the metadata.
- This is not only essential for supporting the comprehensive set of features identified for the DWB-RDP, but in general for all users and across a variety of production and dissemination oriented applications.
- The DWB-RDP cannot solve the problem of content quality, but could act as a shared Quality Assessment tool, and can also incorporate tools for harmonization or addressing multi-lingual issues.
- In general, tools produced by the DWB-RDP should be usable standalone or integrated with other frameworks. Some tools are described below.

1.6 Tools / Modules Roadmap

This section captures potential technical developments for improving the DWB-RDP content and services. Many of these could broadly benefit the community and other projects / initiatives.

Normalization / Vocabulary Resolution Services

One of the most challenging aspects of providing data discovery mechanisms across multiple sources lies around the normalization of the highly-diverse content into harmonized controlled vocabularies, for delivering consistent faceting filters to the user. This covers a wide range of elements, from critical search dimensions (such as geography, time, topics), to less commonly used or sparsely populated ones (like sampling or collection methods), as well as portal or domain-specific system properties (like agency, metadata collection or development procedures, version, etc.)

The first question tackles the availability of a list of facets to map to. This should not be a very significant problem and could easily be remedied through a more cautious focus on details in procedures or tools, as international classifications (e.g. country codes) and common vocabularies have been developed (ELSST, DDI).

The main challenge is about picking a term or short sentence and finding its best match in a standard vocabulary, taking into account or using the fact that recorded variable values consist of a combination of codes and labels that can vary across languages or cultures, and may contain typos.

Such problems tend to become very variable specific and different service implementations with supporting databases are required for many particular metadata elements, but could build upon common technology. At the end of the day the objective usually is to convert/map a short text string into a code. Given that such exercises are essentially specialized searches (finding an entry in a vocabulary that matches some text), building a solution based on search technology like Apache Solr could provide a solid starting point. This could be complemented by automatic translation tools, language detection utilities (see below), spelling checkers, etc.

Such research and development effort could be worked on and led by a small expert team. Resulting services could be made publicly available and jointly or community maintained.

Integrated Language Detection

Determining the language(s) of various metadata elements being harvested across sources is not always a trivial process. It is not uncommon for example for a survey to have its description (abstract) in (more than) one language, and its variable-level metadata in another. The geographic location of a data source is not always indicative of the language of the underlying metadata. In some cases, the language documented in the incoming metadata does not match the actual content.

Enhancing the discovery platform by integrating language detection tools in the ingestion process would therefore increase the quality of the language identification at the metadata element level. Such tools would build upon open source libraries or public services (e.g. Google language services). This could be packaged as a standalone library that can be reused by many projects (not only DWB-RDP)

User Management

The DwB prototype for a CESSDA RDP portal is currently an anonymous facility. There is no user registration or authentication system. Adding such a facility, or preferably, integrating it with a European wide single sign-on framework could become an important question. As a one-stop-shop data discovery tool, the portal should be as visible, open, rich and available as can be. However, it could be highly desirable to be able to personalize user experience and enhance security. Note that this should extend into application level authentication and authorization at the underlying service level. Throughout this project, we have held to the principle that the Resource Discovery Portal only gives metadata access, while data access conditions are defined and decided at the decentralized provider level. This should open the general possibility that providers may allow federated access to data resources.

DDI-C Harmonizers

The DDI-Discovery RDF builds on the DDI-Lifecycle models and supports reusable metadata elements. The DDI Codebook on the other hand, uses the variable element as a container for descriptive elements such as questions, categories, and concepts; which often result in the same information being repeated, even if shared by multiple variables. In addition, this misses the capture of semantic relationships across these elements (e.g. knowing that two variables are related to the same question and consequently often share the same concept).

Simple hashing and text analysis techniques can be applied to detect such commonalities, and through that improve the quality of the discovery metadata when ingested from DDI-Codebook base sources. While there is a limit to how much can be achieved through automated harmonization, this has been very successfully used in other projects.

Such additional levels of processing could be added to the current harvesters.

Multilingual-aware Nesstarvester

When a survey is documented in more than one language and published on a Nesstar server, each language is exposed as a different DDI document. The Nesstarvester process step therefore results in multiple DDI-XML documents being retrieved, though these essentially represent the same study. It

would be beneficial to perform a post-harvesting "merger" into a single DDI document, using the xml:lang attribute to differentiate across languages. Note that this is not particularly trivial, for example:

- Not all elements are filled out in all languages; there might be severe discrepancies between the “original” and the “translated” version of metadata, both in terms of which elements are present and in their contents.
- Language agnostic elements and attributes (e.g. summary statistics) would need to be extracted for a "default" language version.
- Not all elements are translated and if missing may be copied from the default version.

Enhancing the Nesstar API / library to take this into account and produce a single multi-lingual DDI document "at source" would be an alternate approach.

Support additional data sources

The current version of the DWB-RDP primarily supports DDI-C and DDI-Disco as ingestion / harvesting formats. Adding additional standards / format is not a particular work-intensive task, though during the lifespan of WP12 we neither had compelling use cases nor metadata sources to work with this specific implementation problem. Support for DDI-Lifecycle, SDMX, OAI, and related metadata standards should be planned for future implementations. Note that mapping these into a DDI_Discovery model could be an interim and more efficient approach.

Support for DCAT-AP

The DCAT Application profile for data portals in Europe (DCAT-AP) is a specification based on the Data Catalogue Vocabulary (DCAT) for describing European public sector datasets. Its basic use case is to enable a cross-data portal search for data sets and make public sector data better searchable across borders and sectors. This can be achieved by the exchange of descriptions of data sets among data portals.

The elaboration of the DCAT-AP was a joint initiative of DG CONNECT, the EU Publications Office and the ISA Programme. The specification was elaborated by a multidisciplinary Working Group with representatives from 16 European Member States, some European Institutions and the US. From April until July four Virtual Meetings were held and the specification has gone through a public review period of two months.

Adding DCAT-AP support to the suggested portal would be desirable. Such effort could be coordinated with the DDI Alliance as it involves developing mappings between DDI-Disco and DCAT-AP.

For more information, see:
https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-application-profile-data-portals-europe-final

Portal UI Enhancements

The current Resource Discovery Portal prototype User Interface could be enhanced through the addition of the following features:

- Better support for progressive/dynamic metadata loading.

- Variable search / view / comparison
  - Improved variable view on study details page.
  - Note that variable level metadata is limited to DDI Disco.
  - Landing pages for questions, concepts, etc.
  - Handling of longitudinal studies (DDI Disco study groups).
- Access to underlying source metadata (not just DwB-Disco RDF).
- Extended search facets (based on availability).
- Improved linking to metadata/data provider and source.

The REST service stack could also be extended to support additional search and retrieval options, as well as proxy access and querying of the back end BaseX XML repositories and Virtuoso RDF / SPARQL services.

Provider Portal

As suggested and discussed in WP12 D12.2 section 2 and D12.3 section 5, adding a secure and private provider area to the DWB-RDP is highly desirable. This aspect of the DWB-RDP has however not been explored in depth, and a number of questions need to be clarified in order to have a streamlined and well-functioning provider portal with relevant tools.

Such questions include a communication and outreach strategy. The following information should be available in potential DWB-RDP information material:

- How to become a provider?
- Why become a provider?
- What are the prerequisites to be a provider?
  (e.g. machine consumable metadata, preferably in a standard format, adherence to minimum content standards, etc).
- How much effort does it take to join?
- Provider level agreement (what are the responsibilities of a provider?)
- Benefits (quality assurance for provider and across providers, enhanced visibility and discoverability).

The DWB-RDP will also have a set of functional and nonfunctional requirements, and needs to fulfill various responsibilities in order to function as a trustworthy service for both providers and consumers. Development- and support policies will need to be produced, possibly along with formal and legal agreements.

3. Conclusions

Europe is data-rich, but the abundance of potential resources is not readily available for research and systematic knowledge production. The DWB-RDP is a tool to significantly improve this situation.

There are many potential providers of research data, but no doubt the most significant are the national statistical system of each country and research itself, of which interests are often represented by the CESSDA members, the Data Archives - the challenge is to build integrated discovery and dissemination systems across them.

The present DWB-RDP is a prototype, focused on solving the discovery problems. However, the solutions developed are robust - and deserve implementation - and have much wider applicability than discovery only. The basic argument behind the DWB-RDP is that systematic work on metadata development in machine-readable forms represents much needed groundwork for both systematic data *production* and data *dissemination.* The work of DwB WP12 is solidly aligned with the ideas of enriching and supporting the HLG-initiative of the NSIs in Europe.

The DWB-RDP has a formidable potential, but needs to be cultivated in order to achieve it. The product needs a home that actively meets the developmental challenges outlined is this summary roadmap. But the roadmap also has a much wider focus. By working in much more organized and standardized ways, data providers would reduce the burden of producing high quality metadata, improve their own performance as data providers and increase their contribution to systematic development and scientific reuse of resources.