Project N°: 262608

Acronym: Data without Boundaries

Deliverable D12.2

(Enrichment and Conversion Tool(s) for OS data)

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**

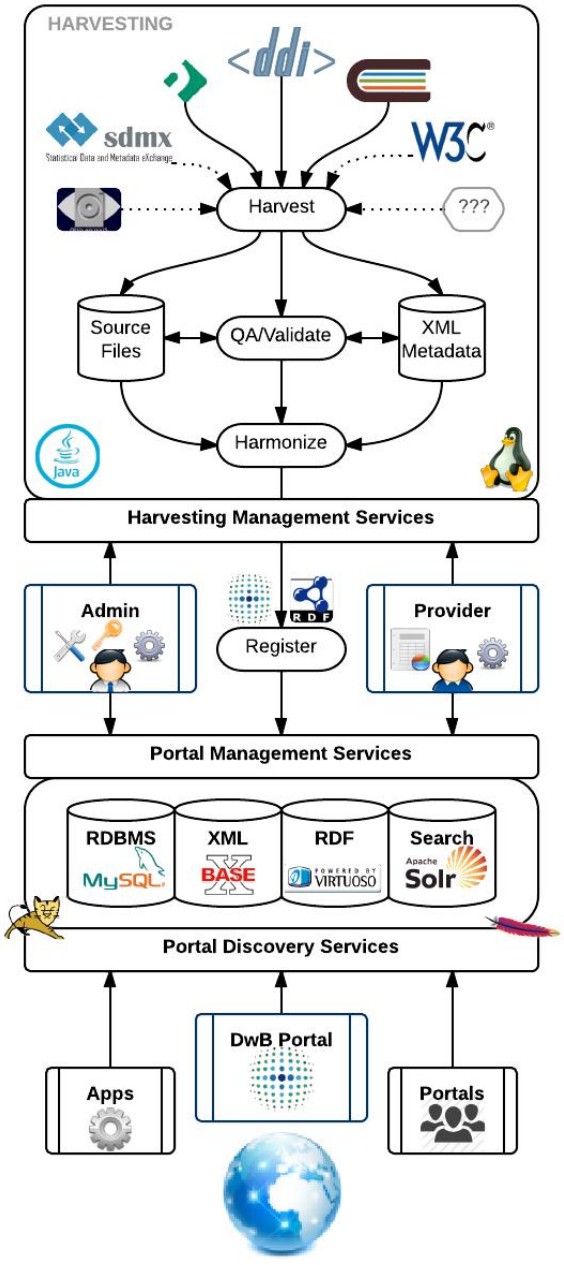
Combination of CP & CSA project funded by the European Community
Under the programme "FP7 - SP4 Capacities"
Priority 1.1.3: European Social Science Data Archives and remote access to Official Statistics

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 (DwB - Data without Boundaries).

TABLE OF CONTENTS

0. Overview

The diagram on the right provides a high level overview of the architecture of the DwB OS 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 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 their information present in the system. This include 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. The metadata is based on commonly used standards and the DwB Metadata Model.
- The Discovery Services exposing the platform and information to the outside world, and enabling integration in applications or web site, along with the Discovery Portal user interface.

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

This document focuses on Enrichment and Conversion Tool(s) for OS data and is Part 2/5 of the project deliverable.

1. Provider Toolbox

The purpose of this section is to provide potential DwB Portal participants with a collection of tools and resources to facilitate the preparation and packaging of OS data and metadata for publication and dissemination.

1.1 Tools

There are significant differences in how NSIs and data archives store, use and disseminate data and metadata. These differences are discussed in the recent report of the OECD Expert Group for International Collaboration on Microdata Access. The report recommend a closer cooperation between the two types of institutions in development of procedures and tools for promoting access to data of scientific relevance.

The software packages listed below are commonly used and popular for the production of DDI based metadata, which is the primary focus of this DwB package. Note that this list of tools is not intended nor expected to be exhaustive as several other packages are available and/or under development. It is therefore rather to provide a starting point for agencies or users unfamiliar with the DDI tools landscape and looking for guidance. Which tools to use also depends widely on the operating environment, internal capacity, or preferred metadata standard. Options range from free to use Open Source (OS) to licensed commercial off the shelf (COTS) packages.

Nesstar Publisher/Server

http://nsd.uib.no/nesstar

Nesstar is developed by the Norwegian Social Science Data Services (NSD).

Nesstar Publisher (freeware) is a DDI-based metadata editor and data conversion tool. It is typically used to add DDI metadata to imported data files (SPSS, Stata, etc). Nesstar Publisher supports data and metadata export (DDI-XML), and is currently used within data archives, data libraries and notably in the International Household Survey Network (http://ihsn.org). Nesstar Publisher may be used as a standalone tool, or in conjunction with Nesstar Server.

Nesstar Server is a special purpose web server designed for data- and metadata dissemination. Through its web interface, end-users/researchers can search/browse metadata collections and do simple data analysis. Nesstar Server also supports data conversion/download of data sets into a set of widely used file formats.

Metadata published to Nesstar Server is available for other systems to consume over Nesstar APIs (Java or RESTful).

More information about Nesstar Publisher and Nesstar Server is available at the NSD website (see above).

OpenDataForge (OpenDF)

http://www.openmetadata.org/dataforge

OpenDataForge (oDF) is a collection of desktop and cloud based solutions developed by Metadata Technology North America (MTNA) to address practical user needs around data management and to foster the adoption of globally recognized metadata standards such as the Data Documentation Initiative (DDI). OpenDataForge aims to facilitate the transformation and processing of statistical and scientific data by unlocking and liberating the data and metadata from proprietary formats.

OpenDataForge currently includes:

- ***SledgeHammer***: desktop product for managing data across popular proprietary, text, and database formats, and facilitating the extraction and generation of standards based metadata (see details below).
- ***Caelum***: a tool for leveraging XML data and metadata by transforming them into useful reports, documents, and other formats.
- ***Asmurex***: a simple command line Java based utility to convert multi-record fixed ASCII text data file into separate files holding only data for a specific record type.

Soon to be added to the toolkit is DataWeaver, a metadata driven package focusing on data transformations such as the production of microdata subset, data linkages, or computing aggregates for processing, distribution or analysis. Other utilities are likewise in the development stages.

OpenDataForge products are commonly available as freeware, with advanced versions of selected packages distributed under a commercial license.

SlegdeHammer (OpenDF)

http://www.openmetadata.org/sledgehammer

SledgeHammer is a core component of the OpenDataForge toolkit. It is a desktop product for managing data across several popular formats (proprietary, text and database), and facilitating the extraction and generation of standards based metadata.

At the time of this writing, SledgeHamer core features include:

- Reading data and extracting metadata from: Stata, SPSS, SAS, DDI+ASCII, and Stat/Transfer+ASCII
- Transforming data into ASCII text format (fixed, csv, delimited), with various optimizations
- Producing standard metadata from input files. Supported specifications include DDI-Codebook (1.0-2.1 and 2.5), DDI LifeCycle (3.1-3.2), and Triple-SSS (1.1, 2.0)
- Computing summary statistics at the variable and category levels for inclusion in DDI or for other purposes. Summary statistics include minimum, maximum, average, standard deviation, count of missing values, weighted and unweighted frequencies
- Generating scripts for reading ASCII data into R, SAS, SDA, SPSS, Stata, and various flavors of SQL (MS-SQL, MySql, Oracle, Vertica, HSQLDB). The database scripts support the creation of database schemas along with bulk loading of the ASCII data.

The software is available under both a freeware and commercial license (see web site for details).

SledgeHammer if of high interest to this DwB project as a tool to extract and generate metadata out of statistical data files. This is one of the typical first steps necessary for data producers or custodians for the preparation of variable level metadata or packaging datasets in open formats for dissemination.

OpenMetadata Survey Manager

http://www.openmetadata.org/surveymanager

The OM Survey Manager is a DDI 3.1 based open source metadata editor developed by MTNA deriving from similar tools implemented for the Canadian Research Data Centre Network (CRDCN), NORC at the University of Chicago, the US National Science Foundation, and other projects or agencies.

The OM Survey Manager aims at providing researchers and data administrators with a tool to familiarize themselves with DDI and survey metadata management. It is designed to work with metadata produced by DataForge SledgeHammer.

The OM Survey Manager is a standalone application used to view, edit, and enhance survey metadata saved in local or shared repositories using DDI and other related standards.

Colectica

http://www.colectica.com

The Colectica Platform offer comprehensive solutions for statistical agencies, survey research groups, public opinion research, data archivists, and other data centric collection operations that are looking to increase the expressiveness and longevity of the data collected through standards based metadata documentation.

Colectica is build around the DDI-LifeCycle and therefore aims at broadly supporting the metadata specifications' features. It is particularly strong around questionnaire/instrument documentation. The platform is composed of:

Desktop based applications

- Colectica Reader is a free tool to view DDI 3 metadata
- Colectica Express is a tool to view and edit DDI 3 metadata.
- Colectica Designer allows documenting data collection, survey specification, and dataset descriptions. It can be used with Colectica Repository in large organizations.
- Colectica for Excel brings standards-based data documentation to Excel workbooks.

Server based software

- Colectica Repository is a centralized storage system for managing data resources, enabling multi-user workflows, and providing automatic version control.
- Colectica Portal is a web application, powered by Colectica Repository, which enables data and metadata publication and discovery.

- Colectica Workflow Services provides a publication workflow for data documentation.
- Colectica RDF Services is an addin that allows querying DDI 3 as RDF using a SPARQL endpoint on top of Colectica Repository and Colectica Portal.
- Colectica Fusion is a customized package of Colectica software and services.

Developer tools

- Colectica Toolkit is a collection of command-line tools that enable batch metadata processing.
- Colectica SDK is a Software Developer's Kit that allows developers to create and process metadata with minimal effort.

DDIEditor

http://www.samfund.dda.dk/dditools/default.htm

The DDIEditor is developed by the Danish Data Archive, as the key tool in a framework of data processing tools and processes related to data processing of survey datasets. The end product is data documentation in accordance with state-of-the-art international metadata standards for collections of surveys. The DDIEditor produces metadata documentation in DDI-Lifecycle. A key objective for development of the DdiEditor is to provide users with a tool that is configurable, extendable and customisable, allowing users to customise their personal working environment to their needs.

Microdata Information System (MISSY)

http://www.gesis.org/missy

MISSY is developed by GESIS - Leibniz Institute for the Social Sciences.

MISSY is a scientific metadata system providing structured metadata for microdata from official statistics online. The aim of MISSY is to promote and facilitate a professional and efficient use of official microdata for the social sciences. It currently supports social scientists working with the German Microcensus (http://www.gesis.org/missy), an extension for the documentation of European microdata is under work. Within the Data without Boundaries project, MISSY is used for the documentation of integrated European OS microdata held by Eurostat (WP5, Task 3): European Union Statistics on Income and Living Conditions, European Union Labour Force Survey, Adult Education Survey, Structure of Earnings Survey and Community Innovation Survey.

Metadata is entered via the MISSY-Editor, a DDI-based metadata tool. In addition to detailed variable level metadata, the MISSY-Editor covers a comprehensive study description including country-specific information. To integrate a part of the metadata automatically into the database, several import routines are provided (e.g. from SPSS output and Excel). DDI export routines will be available to enable a reuse of the metadata in other systems. The DDI-RDF Discovery Vocabulary is used as the core data model in order to discover DDI metadata. The DDI-RDF Discovery Vocabulary is an ontology of the Data Documentation Initiative in the world of Linked Open Data. This core data model is extended by a MISSY project-specific data model. These data models can be stored using multiple formats such as DDI-XML,

DDI-RDF and relational databases.

At present, the MISSY-Editor is used for project-related documentation of European microdata and allows for off-site use by external DwB partners. It is planned to develop the MISSY-Editor as a standalone web application. Reusable Open Source software components are available at https://github.com/missy-project/.

Centralising and Integrating Metadata from European Statistics (CIMES)

https://cimes.ensae.fr/

CIMES (Centralising and Integrating Metadata from European Statistics) is a web-based application developed by CNRS-RQ for DwB. It is based on DDI and supports DDI-C XML import. CIMES allows Users to document microdata on three levels: series, studies, datasets (but not on the variable level). For now CIMES has been used only to document microdata from official statistics at the national level in 27 european countries.

CIMES is likely to take the form of an Open Source freeware. However, it is not fully decided when CNRS-RQ will make the source code available, as it requires - among other development tasks - some solid documentation before being released as OS.

More information in the WP5-T2 report at
http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d5-2_databank-national-survey_report_final2.pdf

IHSN NADA

http://ihsn.org/home/software/nada

In addition to its globally popular Microdata Management Toolkit (based on the Nesstar Publisher), the International Household Survey Network (IHSN) has developed a powerful and mature solution for data/metadata dissemination known as NADA.

NADA is a web-based cataloging system that serves as a portal for researchers to browse, search, compare, apply for access, and download relevant census or survey information. It was originally developed to support the establishment of national survey data archives. The application is used by a diverse and growing number of national, regional, and international organizations.

NADA, as with other IHSN tools, uses the DDI-Codebook XML-based metadata specification. The package is open source and based on the PHP scripting language. As such it can be easily customized and deployed to delivery an elegant and flexible solution, quickly showing the benefits and return on investment of producing standard based metadata.

Other tools

Michigan Questionnaire Documentation System (MQDS)

http://www.blaise.com/Tools

The Michigan Questionnaire Documentation System (MQDS) was designed to extract comprehensive metadata from Blaise survey instruments and render it as an eXtended Markup Language (XML) document using the Data Documentation Initiative (DDI) standard. The current version of MQDS has been updated to exploit many newly introduced features by changing the method of processing Blaise instruments into a relational database.

Questasy

http://www.centerdata.nl/en/software-solutions/questasy

Questasy is a web application developed to manage the documentation and dissemination of data and metadata for panel surveys. It manages questions and variables, including question reuse across multiple studies and longitudinal panels. It also manages concepts, publications, study information, and more.

Additional relevant DDI-related tools:
http://www.ddialliance.org/resources/tools

METIS Common Metadata Framework:
http://www1.unece.org/stat/platform/display/metis/The+Common+Metadata+Framework

New tools?

In addition to the above tools that fairly broadly focus on metadata management, new tools are desirable to solve specific problems or use case. This includes:

- Office/PDF metadata extraction: many agencies have captured and organised metadata in unstructured document formats such as MS-Word, MS-Excel, Adobe PDF and similar. This severely limits the potential use of such material. Metadata should ideally be available in structured machine-readable ways for further processing and use. This points towards a need for two different types of tools:
    1. If metadata are originally stored in databases, Excel sheets or similar, functionality is needed to move material to other formats while saving the structure
    2. If metadata are stored in unstructured text-oriented formats, technology is needed to break up the material and make it more generally useful.
- Metadata quality assurance: content is in the end what matters. Metadata should be treated like official publication and go through sound quality assurance processes. Tools that facilitate such activity should be investigated, developed and implemented. This would require
    1. Compliance with international standards
    2. Versioning, documentation of development history
    3. Multilinguality, quality-checked availability of metadata in national and international language
- There is a general need for tools to store and document data transformations in a generic

format. There are many uses of such technology. Presently, if metadata are stored within one system and data are used and processed within a statistical package, there is no standard way to coordinate the two files. If changes are made to a SAS or SPSS file (e.g., a new variable is created or an existing variable is recoded), there are currently no tools to update the corresponding XML documentation automatically. Consequently, valuable information about the history of a dataset (provenance) is lost. Most of the information that could be captured from CAI systems is never used, because maintaining consistency between the data and the metadata is too difficult. This is a problem now being addressed by the data archives.

- Fine grained (i.e. field level) content visibility tools and tools to automate documentation reports.

1.2 How to guides

Tools are only one piece of the puzzle. Tools need to be complemented by guidelines and training activities around how to use or consistently integrate them in existing environment (they should in general complement existing data management applications, not compete with them). General guidelines requires a thorough understanding of at least three "phases" of the life-cycle of research data:

- The justification and production "phase"
- The management and curation "phase"
- The dissemination and usage "phase"

Metadata is at the end of the day what matters most, metadata is the glue and memory of such a system. Guidelines around content quality are essential to ensure comprehensive, relevant, consistent, harmonized metadata at the institutional and cross-agency levels.

Examples include how to assign identifiers, compose a survey title, write an abstract, or describe a sampling procedure. What belong to the metadata that directly follows the files and what should stay in documents that can be referenced? What is required, what is recommended and what is optional? Presently there is a rising concern about the long term preservation and availability of scientific data. Metadata completeness and quality is a major component in this picture and a key to work procedures and tools development.

Such good practice is for example illustrated by the International Household Survey Network (IHSN) "Data archiving and Dissemination"[1] and "Microdata Documentation"[2] guidelines, in particular the Quick Reference Guide for Data Archivist[3] that has fostered the capture and publication of consistent DDI-C metadata for thousands of surveys by national statistical agencies and archives in over 100 countries.

Presently GESIS and ICPSR are collaborating on developing a generalised set of best practices for archives in using DDI across the lifecycle of data.

[1] http://ihsn.org/home/node/115
[2] http://ihsn.org/home/node/117
[3] http://ihsn.org/home/sites/default/files/resources/DDI_IHSN_Checklist_OD_06152007.pdf

Other such references documents produced by the data archival community include:
the "Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle (5th ed.)"[4] from ICPSR and
the UK Data Archive's "Managing and sharing data - Best practice for researchers"[5], and
UK Data Archive: "Cataloguing Procedures and Guidelines"[6]
the Data Seal of Approval initiative[7]
the DataCite initiative[8]

All of the above should be widely consulted by data and metadata providers to inspire and foster the adoption of institutional good practices.

Content Oriented Guidelines

Developing instruments and collecting data are professions, but the same goes for documenting data for secondary use. It is important to secure which elements are covered and setting quality standards.

A metadata standard tries to outline a complete description of a data resource for a defined set of purposes. The need for meaning or content is related to the hierarchy of potential uses. The present portal being developed focuses on *finding* and *locating* data, problems which requires a relevant description of content, at "study" level, i.e. relative high level general descriptions. The title and a properly developed abstract usually are the specific elements used to carry this content.

There is presently an effort as part of a RDA working group to agree on metadata for discovery based on DDI.

In addition, it will be important to locate and learn about availability of data. Locate means to find out where, who is supplying the data and what spaces/populations the data are about. Availability answers the question if the data are available for use, how available they are to whom, in what way and for what purposes.

Of this, what is required, what is recommended and what is optional ?

Mandatory elements: Elements that are needed to find and locate data
Recommended elements: Elements that give us the ability to appropriately understand data, move them to analytic formats and use them properly.
Optional: The rest

Examples of such discussions:

[4] http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf
[5] http://www.ons.gov.uk/ons/guide-method/the-national-statistics-standard/code-of-practice/protocols/data-management--documentation-and-preservation.pdf
[6] http://data-archive.ac.uk/media/401477/ukda035-cataloguingproceduresandguidelines.pdf
http://www.data-archive.ac.uk/create-manage/document/resources
[7] http://www.datasealofapproval.org/
[8] http://www.datacite.org/whatisdatacite

http://www.ddialliance.org/sites/default/files/ddi-lite.html
http://www.ddialliance.org/sites/default/files/cessda-rec.pdf
http://schema.datacite.org/meta/kernel-3/index.html

Quality-oriented guidelines should answer questions like:

- How do I write a "title"
- What is an abstract and what elements should be covered
- Geography (country codes)
- How to format dates? UK, EU, ISO

Typical challenges met when a decentralised network is documenting collections of resources for common purposes:

- identifiers, naming conventions
- content harmonization
- multilinguality
- controlled vocabularies and standardization

Tools Oriented Guidelines

Effectively using data and metadata management tools entails more than just selecting a recommended package, it also necessitates good user manuals, “how-to” guides, training modules, and the likes. Most commercial packages and successful open source packages normally come with such documentation that is at least sufficient for users to get started and rapidly produce content.

In relation to the above identified tools, this includes the Nesstar Publisher Guide[9] or SledgeHammer User and Technical Guides and related videos[10]. Free or paid training services are also often available through the software publisher, conference events, or formal courses. Colectica for example offers training modules around DDI and their tools[11].

Ensuring that such materials are widely available and accessible is essential.

Developing new tools is also an important aspect of furthering the availability and quality of data and metadata.

Other Guidelines

- Guidelines for producing portal friendly/compliant metadata
  Metadata should be produced / included as closely to the data production process as possible, that enhances the quality, completeness and automation potential. This indicate a need for a two-way integrated relationship between production/instrument and metadata specification/production. Examples:

[9] http://nesstar.com/help/4.0/publisher/download_resources/Publisher_UserGuide_v4.0.pdf
[10] http://www.openmetadata.org/dataforge/sledgehammer
[11] http://colectica.com/training

- i.e PDF parser, database export tools, Blaise as CATI-system (ISR) Metadata production is not only to fetch elements from production systems or alternative systems, but it is also to guide development and functioning of the production systems. Important to enhance the usability and flexibility of metadata.

2. Provider Portal

Organizations that provide metadata to the portal, will get access to the portal's "provider services", a.k.a. the Provider Portal.

The Provider Portal will give access to the organization's portion of the harvested/indexed metadata holdings, and to a set of tools that can help improve quality and consistency in the metadata. Using the Provider Portal, organizations will also be able to manage visibility of their metadata in the portal, and to initiate re-harvesting of their local metadata repositories into the portal.

2.1 Quality assurance, usage, and other reports

Pilot project on quality reports:

In the period May-September 2013, data curation staff at NSD was given a set of pilot reports generated based on metadata harvested from NSD's Nesstar server for Norwegian survey data (http://nsddata.nsd.uib.no). The pilot reports were crude and quite static - but were still extremely well received by the data curation group. The reports enabled them to view the (otherwise file-based) collection in an integrated fashion. Despite continuous efforts to keep metadata quality and consistency on a high level, the pilot reports immediately revealed errors, inconsistencies, typos and structural problems in many of the metadata sets in the collection.

After a few improvements to the report generator (based upon feedback from the group and implementation by Pascal Heus/Metadata Technology), the data curation group at NSD could very easily correct the metadata records in their local Nesstar-based repository. Then, after reindexing, they would check the reports again to verify that quality was improved.

The conclusions from this minimalistic pilot project was that a Provider Portal (with powerful and dynamic reports and other tools to instantly improve and test) can become a highly valuable addition to the in-house tools and processes used by data providers.

The Open Language Archives Community (OLAC) produced an informational note on metadata metrics in 2009 (http://www.language-archives.org/NOTE/metrics.html) that may act as a useful guide, both quality-scores and completeness-scores may act as very useful information for serious data publishers.

User-Story-driven development:

The functionality and tools of the Provider Portal need to be specified in detail, and according to best practices, the design and development should be driven by user stories and interaction between

developers and representatives from data providers (e.g. data curation groups at Data Archives).

The basic template for a User Story is:

As a <type of user>,
I want <to perform some task>
so that I can <achieve some goal/benefit/value>

and it requires one or more scenarios of the form:

Given <precondition>,
(And <precondition> ….) OPTIONAL
Then <outcome>
(Otherwise <outcome>) OPTIONAL

The scenarios can be used as acceptance criteria (because of their 'if, then, else' structure), and tools such as Cucumber (http://cukes.info/) will convert them to runnable acceptance tests.

It is important to note the quality tools in the portal should be developed in a way that enables the tools and rules to be extracted into standalone tools that later can become part of the in-house quality assurance toolkits. The ambition is that most quality checks on metadata can be run locally before exposure to the Portal's harvesters.

Deliverable D8.4 "Final report proposing portal resource discovery functionality for a search/ browse portal interface" has provided an abundance of user story input to the portal development work. However, this is focused on the end/scientific user of the portal. Because D12.2 has metadata enrichment and conversion as its main focus, we have to conclude that the availability of user-story type guiding information from this provider perspective is not that easily available.

Expected features (examples):

As stated above, the development should be user story driven. However, it can be useful to provide a few examples of quality assurance tools that *could* be added to the Provider Portal.

Content Quality Assurance

- Identifier analysis: ensure that all studies or other metadata elements that are expected to be uniquely identifiable do not have duplicate values and comply with expected formatting (e.g. not starting with a number)
- Geospatial Analysis: Ensure that country and other geospatial related elements are present where required, based on standard codes (such as ISO-3166), and/or are cleanly and consistently labelled
- Time analysis: Ensure that date/periods and other time related elements are present where required, based on standard codes (such as ISO dates), are cleanly and consistently formatted, and chronologically consistent (e.g. start date <= end date or within expected time period)
- Naming conventions and consistency - establish and apply rules for use of case, spacing, hyphenation etc for key attribute values
- Controlled vocabulary analysis and spelling variations: generate a list of distinct term for

particular metadata elements to ensure that it is consistently spelled.

Content/Usage Analysis

- Variable bank analysis: produce list of unique variable names and their frequency of use. This is often useful to support establishment of variable banks or standard naming conventions
- Single variable analysis: generate a report for particular variable names to document values and variations of its common metadata element (e.g. how it has been labelled across all studies, what are the associated universes or question texts, etc.).
- Classification analysis: for a specific categorical variable, report code/category variation and frequencies.

2.2 Control what is visible (public)

We believe that metadata providers would prefer to be able to control which of their metadata that becomes available to the Portal's end users, especially when the metadata is in the quality improvement phase. The goal is therefore to develop functionality that gives metadata providers reasonably fine-grained control over how their content is exposed.

Details remain to be specified, but functionality to remove/add metadata sets from/to the index is planned. Another potential functionality could be for a provider to control what categories of users get access to what levels of metadata detail.

2.3 'Harvest Now' functionality

Metadata providers should also be able to initiate re-harvesting of their local metadata repositories. This is thought to be particularly useful during periods of quality improvement in the metadata. Findings from the pilot report project (see 2.1 above) indicate that metadata providers will want to verify their changes/fixes frequently. The Provider Portal should therefore support 'Harvest now' functionality, enabling metadata providers to cause their repository to be re-harvested.

2.4 Minimal study level editor

In cases where the local metadata repository (or homepage) doesn't have much metadata to expose to the portal, a minimal study editor could be useful to have in the Provider Portal. In a minimal study editor, data providers could manually register a minimum of metadata (title, abstract/subject/topic, geographic and temporal coverage) along with links to more information about how data sets may be obtained or accessed.

A web-based minimalistic editor for study level metadata is relatively straightforward to develop, and could be added to the Provider Portal with reasonable effort.

3. Metadata Processing & Enhancements

Gathering metadata to feed the portal content in a two stage process consisting of

1. Collecting metadata in various formats from external providers and
2. Converting these metadata into a harmonized/consistent format for publication in the database and indexing by the search engine.

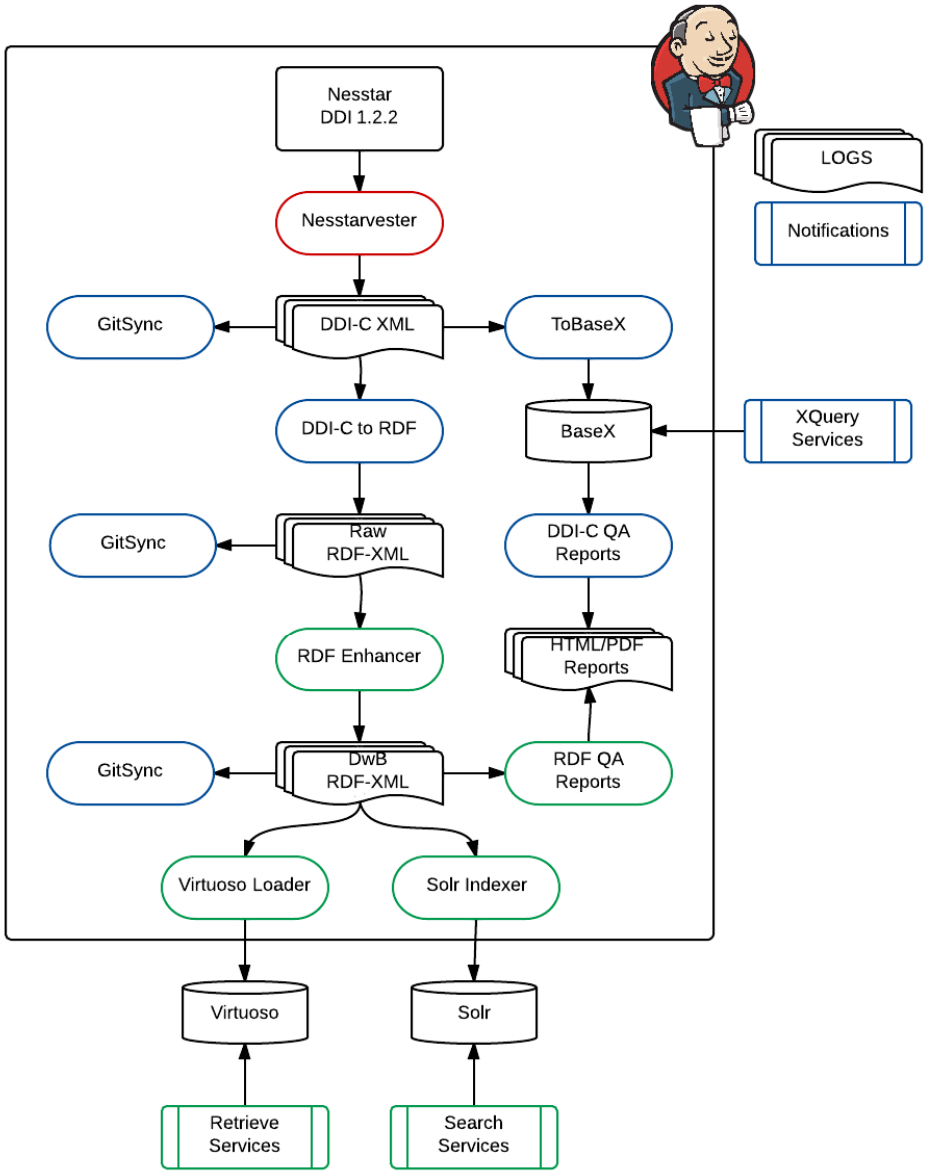
The first stage is the responsibility of the harvesting/ingestion system which is responsible for retrieving metadata from external catalogs or sources. This is the focus of the Deliverable D12.3 report which describes these components in further details. A good example of such an operation is the full or incremental retrieval of DDI-Codebook XML files from a Nesstar server.

Once the metadata has been collected, we need to further transform it into DwB-RDF format in order to publish in the catalog. This ensures that the metadata is presented in a consistent fashion to users, independently of its original format or standard. To support this, we are implementing a collection of tools for the post-processing, publication and indexing of the metadata in the DwB Portal databases.

These processes, illustrated in the diagram below, include:

- Cleaning, converting and harmonizing external metadata into DwB-RDF
- Enhancing metadata to facilitate search faceting and publication
- Generating internal quality assurance, and provider reports

The related tasks present various technological and content management challenges whose resolution will lead to the production of new tools & utilities along with new knowledge & expertise around these topics. Note that an original version on the incoming metadata will always be preserved and as much as possible carried over to the front end services and portal.



Gathering and aggregating metadata from diverse sources typically presents the following challenges:

1. *Metadata standards and format diversity*: within a domain, in our case official statistics, many metadata standards and management platforms exists. For example specifications like DDI-C, DDI-L, SDMX,, etc. and platforms like Nesstar, Colectica, and other open and proprietary systems. A content aggregator like the DwB Portal must be able to cater for each particular use case.

2. *Variation in standard practices*: while a prime objective of establishing and adopting a standard is to harmonize practices and produce consistent information across sources or agencies, it is often subject to interpretation and practically, we find significant variations in content across sources or even within agencies (e.g. departments). Such diversity tend to grow with the complexity of the standard (as more elements/attributes/options are available). DDI for

example falls into this category and not all DDI elements are used in the same same way (agency specific definitions, required/optional, attribute/element, free or rich text content, etc.). This is further complicate as standards version and change over time.

3. *Content variations*: The metadata content itself is subject to a wide range of variations, even when supposedly representing the same information. An element is essentially a piece of text that can vary in length, be open or based on keywords / controlled vocabulary, contain spelling variations, subject to typographical mistakes, be upper/lower cased, and be expressed in different languages. Interpreting this in a consistent fashion across sources can be extremely challenging. A very simple but good example illustrating this is an element identifying a country. This could be based on an ISO code (which has 3 flavors and varies over time) or other international or internal code lists, or a typed country name, which is subject to numerous variations (i.e. 56, be, bel, Belgium, Belgique, België, Royaume de Belgique, Belg., Blegiqe, etc.). Even for such simple case, consistently interpreting this into a standard county code can rapidly become a significant problem.

The DwB Portal harvesting and ingestion is the layer responsible for answering the first challenge by providing mechanisms for collecting metadata from diverses sources.

Once the metadata has been retrieved, we then need to attempt to address the remaining issues. This will be implemented by taking the metadata through various cleaning and transformation processes, orchestrated by a workflow management engine (Jenkins). The end result will be metadata expressed in a DwB-RDF format and ready for publication in the database and indexing by the search engine.

Each step will typically take XML/RDF metadata as input and generate outputs, stored in files or database, and ready for the next step. The general philosophy will be for each task to be a simple specialized script or program with no or little coupling with other ones.

Secondary processes will also be part of the workflow to produce along the way various outputs and reports. A subset of these outputs will be exposed as documentation and manuals through the provider portal.

3.1 Clean-up / harmonization

The primary objective of the clean-up and harmonization processes will be to transform the incoming metadata into a normalized DwB-RDF. This will involve chaining processes such as the one listed below:

- Source specific post-processors: to be applied right after ingestion to perform any structural or content adjustment applicable to a particular metadata source or provider.
- Standards/flavour specific post-processors: these will be applied after the source. specific processors to perform further adjustments as needed and facilitate subsequent steps. This may include sub-processes such as ID generators or upgraders. Some may be flavour specific (e.g. Nesstar, Colectica, CIMES, MISSY, etc.)
- ID generators: to attach DwB portal consistent and unique identifiers or other tags to relevant metadata elements
- Content Quality Assurance: to ensure that the metadata content conforms to the portal

requirements. This is both from the perspective of being able to process the incoming information and ensure that it delivers relevant information to the user. Note that failing some of the QA test may interrupt the processing for the particular document.

- DDI-C 2.5 upgrade - a tool to turn any previous version of DDI-C into version 2.5
- DwB-RDF converters: to clean/harmonized DDI-C 2.5, DDI-L 3.1, DDI-L 3.2 tinto DwB-RDF

The DwB-RDF will then be send through the faceting and other enhancements processes to finalize the metadata.

3.2 Faceting / Enhancement

3.2.1 Overview

Faceting is a a powerful feature of the Solr search engine to be used to support the portal.

Faceted search, also called faceted navigation or faceted browsing, is a technique for accessing information organized according to a faceted classification system, allowing users to explore a collection of information by applying multiple filters. A faceted classification system classifies each information element along multiple explicit dimensions, enabling the classifications to be accessed and ordered in multiple ways rather than in a single, pre-determined, taxonomic order.

Facets breaks up search results into multiple categories, typically showing counts for each. It in turns allow users to 'drill down' or further restrict their search results based on those facets and rapidly mine their way to items of particular interest.

The following web sites are good illustration of faceting in action:

- http://www.kayak.com/

Faceting for the DwB portal will focus on the fundamental search dimensions that have been identified through the WP8 activities and further discussed during WP8/WP12 joint meetings. These includes:

- Geospatial coverage (by country/sub-national region)
- Temporal coverage (single date/time period)
- Kind of data (e.g. census, household, see WP5 T5.2)
- Topic (ELSST)
- Sampling method
- Time method
- Agency (producer, provider)
- Language
- Mode of collection
- Access policy
- Contains variable metadata (flag)

Note that the other major search aspect is free text search, but this is fully handled by Solr and does not

require further processing of the incoming metadata. However, this requires that these elements are reasonably well covered in the actual metadata, and this will requirement will influence the data quality enhancement procedures.

3.2.2 Challenges

To be effective there needs to be a range of approaches to navigating the data, requiring:

- Multi-lingual faceting;
- Global facets (common across all collections);
- Specialised facets (particular to a collection);
- ELSST thesaurus-driven search (e.g. query expansion with synonyms, broader/narrower terms).

3.2.3 Strategy

To be able to develop a consistent data location apparatus, some strategic questions have to be answered. These are

- Use a Solr core per collection (some thought is needed to determine the criteria for collections, but they should be high-level discriminators in order to clearly divide the search space);
- Make sure the global facets are high-level discriminators that are orthogonal to the collection criteria);
- Specialised facets are used to refine search within a collection, reducing the number of results returned whilst increasing their relevance.

Some examples might be helpful to distinguish between collections and facets:

- Collections - social science research; NSI outputs; restaurants;
- Global facets - country, language, date, producer;
- Specialised facets - cuisine, price range, reviewer ratings.

3.3 Controlled vocabularies

As discussed above, an essential feature of modern search engines if to deliver mining and filtering capabilities through faceting. For non-open ended list (like keywords) or continuous dimensions (like time), this implies establishing controlled vocabularies that can then consistently be used across sources and languages.

For the purpose of this project, we expect to need and establish such list based on:

1. Commonly used international standards (e.g. ISO, Eurostat, Newchatel initiative)
2. Outputs from the DwB Work Package 5 and practical working on integrated data collections.
3. The Multilingual European Language Social Science Thesaurus (ELSST), in particular its potential to function as a synonums repository.
4. Recommendations from the DDI Alliance Controlled Vocabularries working group (http://www.ddialliance.org/controlled-vocabularies)

Additional sources may be identified as the project progresses.