Project N°: 262608

Acronym: Data without Boundaries

DELIVERABLE D7.4

Software development and data/metadata integration

WORK PACKAGE 7

Standards Development

REPORTING PERIOD:	From: Month 37	TO: Month 48
PROJECT START DATE:	1st May 2011	DURATION: 48 Months
DATE OF ISSUE OF DELIVERABLE:	April 2015	
DOCUMENT PREPARED BY:	Partners 6 & 9	NSD & IAB

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).

ACKNOWLEDGEMENTS

This document was prepared by Ørnulf Risnes, task leader (NSD), and David Schiller (IAB).
The authors gratefully acknowledge contributions and feedback from WP7 team leader Uwe Jensen (GESIS) and team partners Stefan Ekman, (UGOT-SND), Claus-Göran Hjelm, and Hans Irebäck (SCB) as well as Mari Kleemola (UTA-FSD).

TABLE OF CONTENTS

1. INTRODUCTION AND BACKGROUND

The Data without Boundaries (DwB) project is an EU project within the 7th Framework Programme (FP7), aiming at making official statistics microdata from European countries available for researchers within the European Union.
With 29 partners belonging to the European Statistical System (10 National Statistical Institutes or statistical departments), to the CESSDA (11 Data Archives) and to the Research Community (7 universities and 1 SME involved in methodological research), the project aims at discussing and promoting common improvements, solutions and frameworks to be proposed to these communities[1]

This report describes fundamental shortcomings in the current metadata standards and especially technologies in domains relevant to DwB, as well as new and alternative approaches to overcome these shortcomings.

The audience for this report are statistical institutes, data archives, software makers and all metadata experts that are interested in metadata, metadata standards and technologies.

1.1 Description of Work

The central purpose of Work Package 7 was to create a common platform for a lasting cooperation between national statistical institutes and data archives. This Deliverable is based on Task 7.6 that was conveyed following in the Description of Work (DoW):

- "This task will discover and describe specific issues involved in software development to specific widely used metadata standards."

In the DoW, the Deliverable D7.4 was described as a report on

- "Software development and metadata standards".

The title was later changed slightly to reflect the content of the deliverable; see section 1.3 below for the rationale behind this decision.

The Description of Work for the project and for Work Package 7 was written mainly in 2010, and the DwB project started in May 2011.

The DoW for WP7 emerged from extensive discussions with several stakeholders and reflected the situation and the needs of the various communities and groups in the metadata field in 2010.

The perspectives and findings in this report have been compiled after broad participation in standards- and technology design and implementation activities and other projects within and outside DwB during the entire project period. Ørnulf Risnes (NSD) has e.g. in the period joined three "DDI Moving Forward" workshops organized by the DDI Alliance, and is part of the working group on reforming data descriptions in the DDI standard, aligning it with the Generic Statistical Information Model (GSIM)[2] as

[1] Data without Boundaries - DwB website: http://www.dwbproject.org/about/
[2] GSIM 1.1: http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model

well as upcoming technologies and trends in data and research. Risnes is also responsible for development of the DDI-based software suite Nesstar[3] and a technology architect behind RAIRD[4], a novel privacy preserving platform for remote analysis and research on detailed data from administrative registers.

Additional work done as background for this paper includes, but is not limited to:

- Organization of and participation in a DwB-workshop on Microdata Computation Centre (MiCoCe - http://www.dwbproject.org/events/workshop-micoce.html) (Schiller)
- DwB WP4 - Improving Access to Official Statistics Microdata (Schiller)
- DwB WP8/12 - Improving/Implementing Resource Discovery for Official Statistics Data (Risnes)
- Writing of CESSDA AS Work Plan Proposal - Technical Services (Risnes)

1.2 Relationship with other WPs and Deliverables

This deliverable D7.4 is complementary to the integrated deliverables D7.2/7.3 (2014), and builds largely on findings and discussions there since many aspects of "Software development and metadata standards" (the main topic of this deliverable) are indeed covered by D7.2/7.3. This deliverable D7.4 will build on that work, and report on new developments and technical achievements along the lines mapped out by D7.2/7.3.

1.3 Outline

In a deliverable about software development and metadata standards, many readers would expect discussions on interoperability, semantic and structural mapping between standards (and versions of standards), formats, models, etc. Such aspects will however largely remain undiscussed in this deliverable, which instead focuses on a more profound metadata problem. For this reason, the title of D7.4 has been altered from "Software development and metadata standards" (as specified in the DoW) into "Software development and data/metadata integration".

Achieving robust mechanisms for dealing with metadata at the appropriate abstraction level (i.e. as close to the data as possible) and with suitable granularity, is absolutely crucial in reaching mature and cost-effective solutions for sophisticated semantic and structural interoperability and to make the most of metadata to drive data production and understanding.

Since the beginning of the DwB-project, a wide range of technologies, initiatives, projects, reports, collaborations and coordinating efforts have changed the understanding of the metadata-landscape in a profound manner. Effects of this improved understanding are already visible in many organizations and standards-communities, but the potential for improvement is enormous.

[3] Nesstar website: http://nesstar.com
[4] RAIRD project website: http://raird.no

The main focus of this deliverable is therefore to illuminate some of the most important findings during the project period, and primarily to address a fundamental problem in metadata production that transcends the choice of metadata standards: the “metadata gap”, and solutions and suggestions for mitigating this fundamental problem.

2. THE METADATA GAP AND PROVENANCE

Metadata may or may not be standardized, and metadata standards may be lightweight or comprehensive; currently they all suffer from the same underlying problem:

Metadata is currently usually not automatically recorded or updated through the data lifecycle. The main reason for this is technological; technology used in data production and data management is largely metadata-ignorant or metadata-anemic. This is true both for traditional statistical packages (SAS, SPSS, Stata, relational database management systems, etc) as well as more recent data processing technologies (R, Python, Java).

Metadata is not treated as a first class citizen in any of the technologies listed above, and metadata therefore exists in a realm outside data and data management technology:

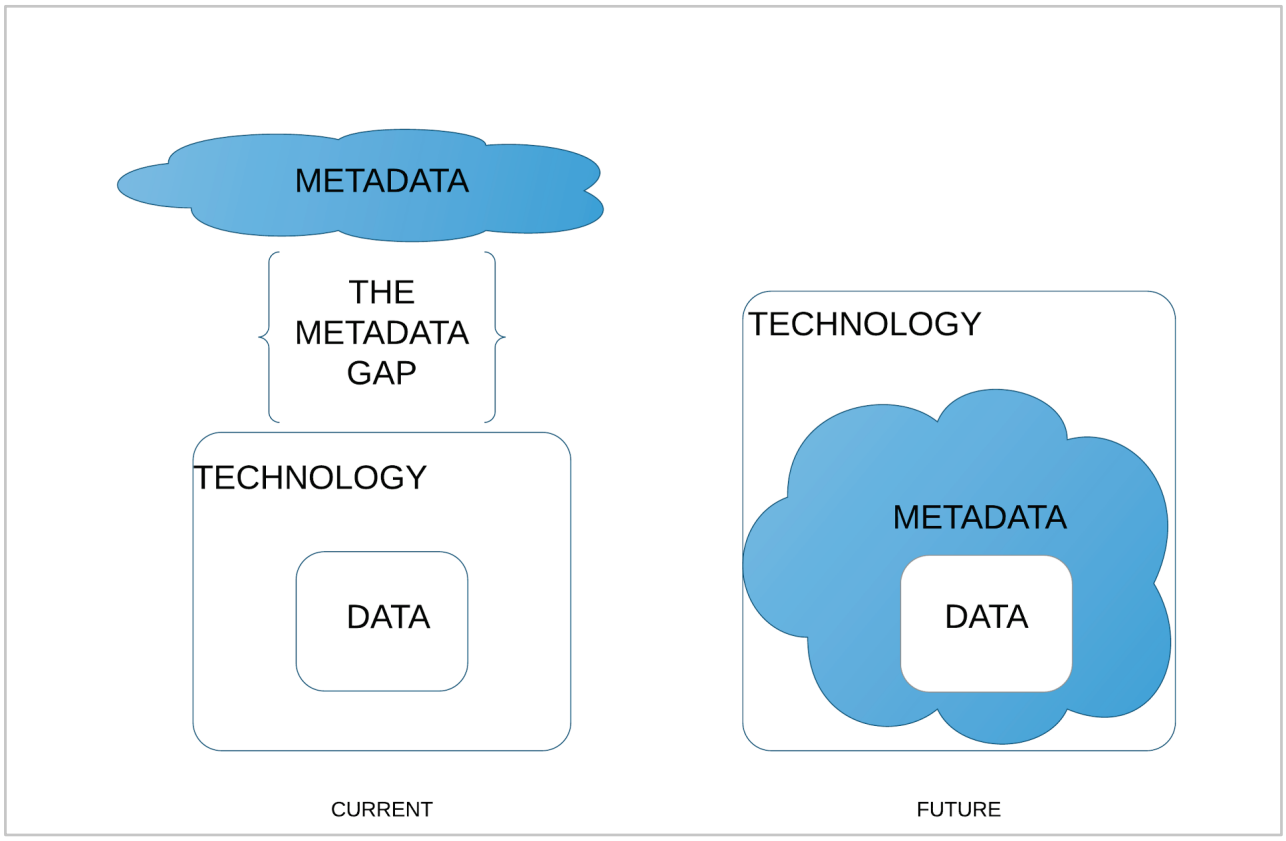

Figure 1 - Illustrating the Metadata Gap

Metadata management tools and processes do not have full access to information about the way data gets collected, structured and processed. Because of this lack of information, metadata production and management can never be fully automated, and relies on manual work even for maintaining highly structural and technical, low-level metadata.

Manual and after-the-fact metadata production and management is resource intensive, but the main problem with this approach is that metadata gets outdated, prone to error and incomplete, and that it introduces uncertainty to whether metadata reflects the current status in data - or an earlier state. Such uncertainty results in a lack of trust in downstream data and metadata processing and consumption. Complete audit-trails/provenance chains for data become unattainable.

In "A Survey of Data Provenance in e-Science" Simmhan, et. al. (2005) shows how crucial fine-grained Data Provenance is along a number of dimensions:

- Data Quality
  Provenance metadata are important factors in determining data quality and veracity.
- Audit Trails
  Audit trails are important for many reasons, including monitoring of resource usage in data production.
- Replication
  Provenance metadata are crucial to providing "replication recipes" for research or repeated use of data.
- Attribution/IPR
  Provenance metadata are important in establishing copyright and intellectual property concerns of data.
- Informational/discovery-related
  Provenance metadata for a given data set can also assist discovery of both ancestral and derived data.

The metadata gap represents a large and fundamental problem, and exists because commonly used data management technologies do not produce, understand or use metadata as an integral part of data management or data processing. As such, the metadata gap's existence is orthogonal to metadata standardization and independent of any particular metadata standard. This deliverable will therefore describe the metadata gap's underlying causes and effects, as well as currently emerging practices in mitigating the gap in different communities.

3. BIG DATA AND METADATA GOVERNANCE

Despite the metadata gap, data archives (DAs), statistical institutes (NSIs) and agencies (NSAs) have been able to maintain an acceptable level metadata quality for dissemination purposes for many years. Recently, however, new factors and actors that challenge the traditional resource intensive production processes have come into play, forcing professional data- and metadata production organizations to reorient and transform.

D7.2/7.3 includes a more detailed overview on the impact of the Big Data-trend.

3.1 The data deluge/Big Data

Data volumes have always been increasing, and as long as growth was modest, DAs and NSIs managed to keep up with the trend, partially assisted by advances in information and data management technology, and partially by scaling up staff.

Increased use of and demand for data from administrative sources as well as general trends in "Big Data" has given the rise to the data deluge, and data volumes are now increasing at a speed where even raw storage becomes a challenge. Traditional resource intensive and manual data management and documentation practices cannot easily be scaled up to handle the new data volumes.

While the metadata gap has always existed, it is safe to say that the data deluge has revealed problems that could earlier be managed by hard manual and organizational work in the relevant organizations and agencies.

Although few scientific papers have been written on the subject of metadata and Big Data, the importance of metadata and metadata governance is starting to receive attention in a domain previously dominated by technology, solutions and perhaps naïve assumptions about the analytical potential in unstructured and undocumented data. In the paper "Metadata Management in Big Data", Vemuganti (2013) addresses many of the same concerns as this deliverable.

3.2 New data actors in production of statistics and research

Along with the data deluge came a range of new, mostly private actors that not only collect vast amounts of data, but that also to some degree challenge the traditional roles of NSIs, NSAs and DAs in statistics production and dissemination for research purposes.

Big Data actors may offer wonderful new data sources for research, and they may also release estimates for e.g. consumer price index (CPI) and unemployment rates much earlier than official statistical agencies.

It is not hard to argue for continued existence of both statistical institutes and data archives, with their focus on data quality, provenance, documentation and of course the NSI's role as an authoritative producer of statistics. These type of institutions still have important roles to play.

Yet the new actors, along with the data deluge, create a pressure for traditional data producers and disseminators to "do more with less", i.e. to supply new types of products, and to continue to supply traditional products, both in a more timely fashion and with less resources available to do so. This situation calls for novel and perhaps radically different solutions to data processing and curation. An important component in a battery of solutions is bridging of the metadata gap.

4. BRIDGING THE METADATA GAP

The metadata gap exists because the link between data and metadata gets broken throughout the data lifecycle. Metadata has to be carefully reconstructed, often repeatedly and manually. To close the gap, data collection, -production and -dissemination systems must honour and retain this link.

Through several initiatives and projects in both the NSI- and DA-communities, serious reforms have started in this direction.

4.1 Ongoing relevant initiatives and standards

4.1.1 GSBPM, GSIM and CSPA

The "UNECE High-Level Group for the Modernisation Of Statistical Production and Services" (HLG-BAS[5]), has fostered the Generic Statistical Business Process Model (GSBPM) and its accompanying Generic Statistical Information Model (GSIM). (See section 2.2.2 of D7.2/7.3 for more detail on the two).

The third and most recent initiative from HLG-BAS on this track has resulted in the "Common Statistical Production Architecture" (CSPA[6]). CSPA is highly interesting in this context, because it describes principles for statistical production that can eventually bridge the metadata gap. A crucial point in CSPA, is that each service (corresponding to a process in GSBPM) accepts GSIM-objects as input, and returns GSIM-objects as output. Hence, metadata is acknowledged as a first class citizen and because every CSPA service is responsible for data and metadata integrity across its processes, the link between data and metadata may be retained throughout the production process as data and metadata moves in and out of such services.

However, since GSIM is merely a conceptual model, it will be up to implementation models to make sure this link is retained in a consistent and meaningful manner.

The DDI-standard[7] is frequently listed as the most likely standard for implementation of GSIM, and as such the responsibility for keeping data and metadata synchronized may end up as a responsibility for DDI to resolve.

4.1.2 DDI Moving Forward and alignment with GSIM

In its ongoing reform program "DDI Moving Forward", the DDI-standard is developed using a model-driven approach, where the (conceptual) model is closely aligned with GSIM.

DDI is a comprehensive and powerful metadata standard, but it is important to note that earlier versions, despite their expressivity, have not really enabled a complete bridging of the metadata gap. This is of course mainly due to a lack of DDI-support in data management and processing tools - but also due to shortcomings related to granularity and abstraction levels in existing DDI-versions. These shortcomings that are being addressed by the aforementioned "Moving Forward"-program[8].

[5] http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+the+Modernisation+of+Statistical+Production+and+Services
[6] http://www1.unece.org/stat/platform/display/CSPA/Common+Statistical+Production+Architecture+Home
[7] http://www.ddialliance.org/
[8] https://ddi-alliance.atlassian.net/wiki/pages/viewpage.action?pageId=491703

4.2 Ongoing relevant projects and developments

A list of projects related to metadata concerning data processing is mentioned in section 3.4 of D7.2/7.3. This short chapter extends that list in an attempt to illustrate similarities in the approach in closing the metadata gap.

4.2.1 Statistics New Zealand - developments in statistical data editing

In the paper "On Tap: Developments in Statistical Data Editing At Statistics New Zealand" [9] (Seyb, et.al., 2012), staff from Statistics New Zealand describes their work to restructure and streamline the production pipelines in order to obtain "[...] more value from official statistics, and creating a responsive and sustainable organisation".

The work pre-dates CSPA, but is aligned with the recommendations and suggestions in the Common Statistical Production Architecture Version 1.1 released December 2014[10].

One of the most important achievements of the work on the reformed platform at Statistics New Zealand can be summed up with the following statement from the paper (page 6):

> *All survey response data and derived data can be viewed in the survey portal along with an audit trail of any changes.*

Complete and accessible audit trails are crucial in closing the Metadata Gap, and in the platform in question, this is made possible by adopting the following paradigms:

1. **Cell-level accessors and version control**
2. **Immutable/persistent data storage solutions**

These are fundamental properties to any platform or system aiming for complete audit trails/provenance chains, and for the ability to manage and document data through production pipelines. A more detailed discussion follows under section 3.3.

4.2.2 Statistics Norway/NSD - the RAIRD-project

In Norway, Statistics Norway (SSB) and the Norwegian Social Science Data Services (NSD) collaborate on a solution that will lower technical and financial barriers in using data from administrative registers for research purposes.

In RAIRD, researchers can never see data directly (as opposed to in traditional Remote Access-solutions); they only interact with data via metadata and via analytical output checked for disclosure risks.

The RAIRD Information Model[11] (Linnerud, Risnes, Gregory, 2014) shows that RAIRD shares many features with the platforms designed in New Zealand (see above), including:

1. **Cell-level accessors and version control**
2. **Immutable/persistent data storage solutions**

The two architectures are designed to server very different purposes (in-house data editing VS research on administrative registers), yet they share these two common features, as well as the ambition to provide complete audit-trails to selected groups of data consumers.

[9] http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/15_New_Zealand.pdf
[10] http://www1.unece.org/stat/platform/display/CSPA/CSPA+v1.1
[11] http://www1.unece.org/stat/platform/display/gsim/RAIRD+Information+Model+RIM+v1_0

Cell-level accessors and cell-level metadata will be discussed in section 3.4 below. Immutable and persistent data storage solutions, and their relationship to metadata and cell-level access and version control will be discussed here.

4.3 Important changes in technology

4.3.1 Immutable/persistent data storage solutions and their relevance to the data and metadata lifecycle

Most programming languages and paradigms, and most database and storage technologies have their roots in a time where computation power, available memory and storage were extremely limited. The same holds true for the most widely used statistical software packages. All actors needed to optimize for the fact that large datasets could not be processed, kept in memory or even stored for very long.

A consequence of this heritage is that most of today's software applications and database solutions (including most of the so-called "NoSQL"-offerings) see data as mutable: any data value may be overwritten at any time - at the expense of the previous value. This concept of shared, mutable state is a very common source of software bugs, something that represents a huge problem in itself. Here we will however discuss other types of implications relevant to the data production, curation and dissemination community.

When older data values get deleted (either overwritten by new values or simply erased), we lose the capability to manage change over time. This has obvious implications for the provision of audit-trails/provenance chains across the data lifecycle.

Today, when the price computation power, memory and storage has decreased dramatically compared to both historic levels as well as compared to cost for human labour, we see an increased proliferation of programming languages (Clojure, Erlang, Scala, Go, Rust, etc), database technologies (Datomic, Apache Samza, Git, etc) and software patterns that leverage the concept of immutable data in different ways.

In our context, it is natural to focus on generic software patterns that can be implemented in a variety of ways and thus integrated into existing system portfolios - rather than going in depth in any particular language or technology.

One important software pattern relevant for immutable data is "event sourcing". In an article on event sourcing, Microsoft says the following about the pattern[12]:

"[The event sourcing] pattern can [...] improve performance, scalability, and responsiveness; provide consistency for transactional data; and maintain full audit trails and history [...]"

In short, event sourcing involves storing all events that operate on data, in order to be able to reconstruct every state on demand - and to produce complete audit trails.

There are other software patterns for achieving immutability, most of whom can be realized both with new "immutable-by-default"-solutions as well as using technology that does not enforce immutability (e.g. Java, relational databases, No-SQL-stores, etc) when implemented with care.

[12] https://msdn.microsoft.com/en-us/library/dn589792.aspx

4.3.2 Granularity concerns for immutable data in a data lifecycle context

The current view of the data lifecycle, and the corresponding metadata abstractions are highly influenced by file-based data organization; data is organized in “variables” (or columns), “records” (or rows) and “datasets” (or matrices/files).

Characteristics (simplified):

- Record
  A set of values/properties for a given unit (e.g. individual, region or business)
- Variable
  A value/property with common conceptual domain and value domain for a set of units
- Dataset
  A collection of records/variables

Generally speaking, there have traditionally not been appropriate and widespread abstractions available for the individual values (cells). This is problematic for several reasons:

1. Variables are typically not as homogeneous as earlier models assume. Frequently, we see that different values for a given variable:
   a. Have different sources/provenance chains
   b. May be the result of a recorded observation, an imputation or a derivation
   c. Refer to different time periods or collection times/modes
   d. Have different footnotes or annotations
2. When preserving a complete audit trail, a variable-level or dataset-level granularity becomes cumbersome as this may result in repeated storage/copying of large data structures (like datasets and variables) even for small edits/corrections of individual values.
3. Recombining data for new/other purposes, linking data, or supporting production of highly customized datasets for research quickly becomes work-intensive when the abstraction level for access is too high (i.e. on the variable or record level and thereby above the cell-level)

These factors (and possibly additional ones) have led to the introduction of the “Datum”-object in GSIM, a metadata abstraction on the level of individual data values in datasets. The same granularity exists in the two projects listed above (from New Zealand and Norway), and is being considered for adoption in future versions of DDI. Section 3.4 covers the Datum-construction, its rationale and its potential usages in more detail.

4.4 Datum-oriented approaches in GSIM and DDI

Note: The Datum-construction exists in various published GSIM-documents, and in several working documents in DDI. While the construction is considered valid and relevant and very much needed in both communities, the integration points between the Datum and the surrounding model are still subject for debate - and may therefore change in the future.

4.4.1 Datum in GSIM

The Datum-object was presented already in GSIM 1.0. It shares certain features with "facts" in the data-warehouse-domain, but should be understood much more broadly, and in the context of the domain covered by the GSBPM and a data lifecycle perspective.

GSIM 1.1 defines Datum like this[13]:
"A Datum is the actual instance of data that was collected or derived. It is the value which populates a Data Point. A Datum is the value found in a cell of a table."

The context of the Datum-object in GSIM is shown in Figure 2 below. Note the relationships to (Instance) Variable and Data Set.

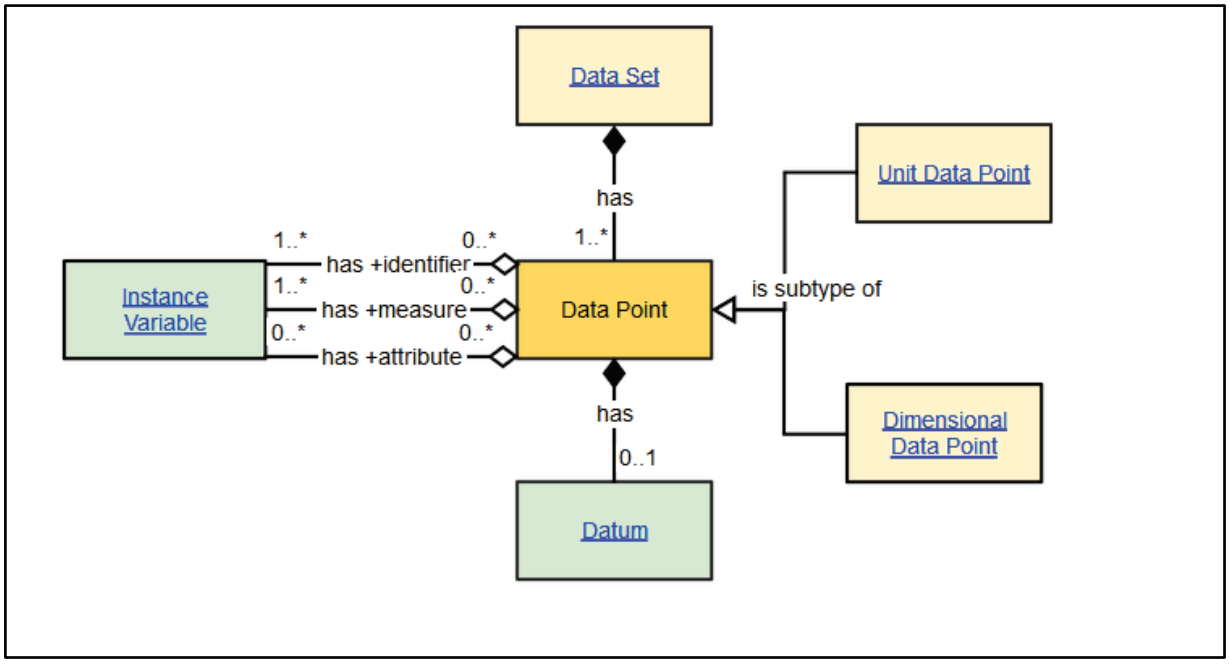

Figure 2 - Datum in GSIM 1.1

Figure 2 illustrates to some extent the "contextual richness" that can be assigned to each Datum. There are relations to *measure*, which holds the actual value of interest. There are relations to *identifier*, which identifies the unit the Datum belongs to. (In case of dimensional data, unit identifiers are composite.) Ultimately, there are relations to *attribute* which allows for relating extended information (e.g. cell-level footnotes, provenance information) to each Datum.

Depending on how the Datum-attribute-relation is interpreted, the current model either has or hasn't any concept of the Datum's temporal and spatial contexts. For handling event-history data (also known as "spell data" or "duration data") common in administrative registers, handling both temporal and spatial contexts on the Datum-level is required.

The introduction of Datum could be regarded as a first and perhaps the most important step towards a closure of the Metadata Gap.

However, GSIM is an information model, and leaves many details for more implementation centric models (e.g. DDI). This implies that the Datum and its relationship to other constructions/objects will need further refinements and clarifications in order to fulfil its promise in GSIM.

[13] http://www1.unece.org/stat/platform/display/GSIMclick/Datum

The Datum-construction has been present in DDI-discussions for a while, and after the 2014 Dagstuhl workshop[14] the Datum-related discussions really gained momentum. Section 4.4.2 below will give an overview of the findings and current status of these discussions (most of whom have taken place in the Simple Data Description View Team[15]).

4.4.2 Datum in DDI

The Simple Data Description View Team under the "DDI Moving Forward"-program has acknowledged the need for Datum-level metadata, and found this to be a fundamental building block for a modernised DDI-standard for the very reasons outlined in this document.

The DDI standard is indeed moving forward in many directions, and to ensure optimal utilization of the possibilities of the Datum-level abstraction level, careful measures have to be taken to align Datum with adjacent areas of the standard. Such areas include, but are not limited to:

- Simple Instrument View
- Physical Data Description View
- The Core Process model

The (currently unpublished) paper "Linking Instrument, Observation, Datum and Variables" by Gillman & Greenfield (2015) contains a thorough discussion and suggestions for how Datum could be fitted into the model in a way that supports a "Datum lifecycle"; the ability to create complete audit trails and complete self-contained contexts for individual Datums from their collection planning, initial capture and throughout their lifespan - regardless of what collections they are part of.

The discussion in the DDI-team has also been influenced by work done in the domain of Open Electronic Health Records (OpenEHR)[16], a domain which shares our goals of providing rich context for data as well as metadata-driven data processing and -packaging.

As far as the authors of this deliverable can predict, there will likely be published both papers and official DDI work products including the Datum-object during 2015.

[14] http://www.dagstuhl.de/de/programm/kalender/evhp/?semnr=14432
[15] https://ddi-alliance.atlassian.net/wiki/display/DDI4/Simple+data+description+View+Team
[16] http://www.openehr.org/what_is_openehr

5. CONCLUSION

Metadata can never fulfil its potential as a driver of automated processes and an authoritative, fully trustworthy source of information for understanding data as long as the metadata gap (as described here) exists.

This deliverable has attempted to aggregate selected findings, developments and technical advances over the last five years to illustrate that bridging and closing/bridging the metadata gap is now technically feasible, affordable and probably the economic choice to make in the current research data and official statistical landscape.

Several projects and organizations have already started the transition towards data production and storage with fully integrated and fine-grained metadata solutions surrounding them. These developments lean themselves on important foundational work done in the realms of GSBPM/GSIM/CSPA as well as SDMX and DDI.

While some work still remains, these converging efforts have paved the way for a long sought-after liberation from sub-optimal workflows and metadata-ignorant tools.

REFERENCES

Common Statistical Production Architecture (CSPA) website:
http://www1.unece.org/stat/platform/display/CSPA/Common+Statistical+Production+Architecture+Home

Common Statistical Production Architecture (CSPA) version 1.1:
http://www1.unece.org/stat/platform/display/CSPA/CSPA+v1.1

Data without Boundaries.
DwB website:
http://www.dwbproject.org/about/

DDI Alliance website:
http://www.ddialliance.org/

DDI Moving Forward 2014 Dagstuhl workshop:
http://www.dagstuhl.de/de/programm/kalender/evhp/?semnr=14432

DDI Moving Forward home page:
https://ddi-alliance.atlassian.net/wiki/pages/viewpage.action?pageId=491703

DDI Moving Forward Simple Data Description View Team:
https://ddi-alliance.atlassian.net/wiki/display/DDI4/Simple+data+description+View+Team

Event Sourcing Pattern (Microsoft Developer Network)
Available at:
https://msdn.microsoft.com/en-us/library/dn589792.aspx

Generic Statistical Information Model (GSIM) version 1.1:
http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model

GSIM 1.1 Datum Object description:
http://www1.unece.org/stat/platform/display/GSIMclick/Datum

High-level Group for the Modernisation of Statistical Production and Services - HLG Overview
http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+the+Modernisation+of+Statistical+Production+and+Services

Nesstar website:
http://nesstar.com

OpenEHR Foundation website:
http://www.openehr.org/what_is_openehr

RAIRD project website:
http://raird.no

RAIRD Information Model (RIM) 1.1:
http://www1.unece.org/stat/platform/display/gsim/RAIRD+Information+Model+RIM+v1_0

Seyb, A. et.al., 2012
On Tap: Developments in Statistical Data Editing at Statistics New Zealand
Available at:
http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/15_New_Zealand.pdf

Simmhan, Y.L., et al.: A survey of data provenance in e-science. SIGMOD Rec. 34(3), 31–36 (2005)
Available at:
http://www.sigmod.org/publications/sigmod-record/0509/p31-special-sw-section-5.pdf

Vale, S. 2010.
Exploring the relationship between DDI, SDMX and the Generic Statistical Business Process Model.
Second Annual European DDI User Group Meeting (EDDI).
Utrecht, Netherlands 8-9 December 2010.
Available at:
http://www1.unece.org/stat/platform/download/attachments/57835554/EDDI+paper.pdf?version=1

Vemuganti, G, “Metadata Management in Big Data,” Infosys Labs Briefings Journal, 11, no. 1, pp. 3-8, 2013.
Available at:
http://www.infosys.com/infosys-labs/publications/Documents/metadata-management.pdf

GLOSSARY OF ABBREVIATIONS

CESSDA	The Council of European Social Science Data Archives
CSPA	Common Statistical Production Architecture
DA	Data Archive
DDI	Data Documentation Initiative
DwB	Data without Boundaries
GSIM	Generic Statistical Information Model
GSBPM	The Generic Statistical Business Process Model
HLG	High-Level Group for the Modernisation of Statistical Production and Services
NSI	National Statistical Institute
RAIRD	Remote Access Infrastructure for Register Data