Distr. GENERAL
25 April 2013
WP.6
ENGLISH ONLY

**UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)**

**ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE**

**Work Session on Statistical Metadata**
(Geneva, Switzerland, 6-8 May 2013)
**Topic (i): Metadata standards and models**

# STRATEGIC PRIORITIES OF THE DATA DOCUMENTATION INITIATIVE (DDI) ALLIANCE SPRING 2013

**Working Paper**

Prepared by Mary Vardigan, ICPSR, University of Michigan

## I. Introduction

1. The DDI (Data Documentation Initiative) is a structured metadata standard related to the observation and measurement of human activity. It started out in the mid-1990s as a replacement for traditional archival codebooks documenting research data, and then branched off to cover the research data life cycle. Over time, as the data landscape has changed, the DDI XML specifications have evolved to add new coverage and functionality to respond to new user requirements.

2. DDI now finds itself at a crossroads, as do national statistical organizations around the world and many other players in the research data space. How do we move forward to meet the needs of an expanding set of users? How do we develop a flexible standard that can document new types of data? How do we grow and adapt as an organization to provide the foundation and support necessary to push the use and sharing of data forward?

3. This paper outlines the strategic directions that the DDI Alliance is undertaking in order to support open access to data and advance effective data-driven science and policy-making.

# II. About the DDI standard and its benefits

## A. A lifecycle approach to metadata

4. While the Data Documentation Initiative (DDI) is an international metadata standard with origins in the quantitative social sciences, it is increasingly being used by researchers and practitioners in other disciplines. The DDI specifications are also being used to document other data types, such as social media, biomarkers, administrative data, and transaction data. The specification itself is modular and can document and manage different stages of the data lifecycle, such as conceptualization, collection, processing, analysis, distribution, discovery, repurposing, and archiving (see Figure 1).

5. This lifecycle approach has similarities to the Generic Statistical Business Process Model, the documentation[1] for which highlights the correspondences between GSBPM and the DDI Combined Life Cycle Model.

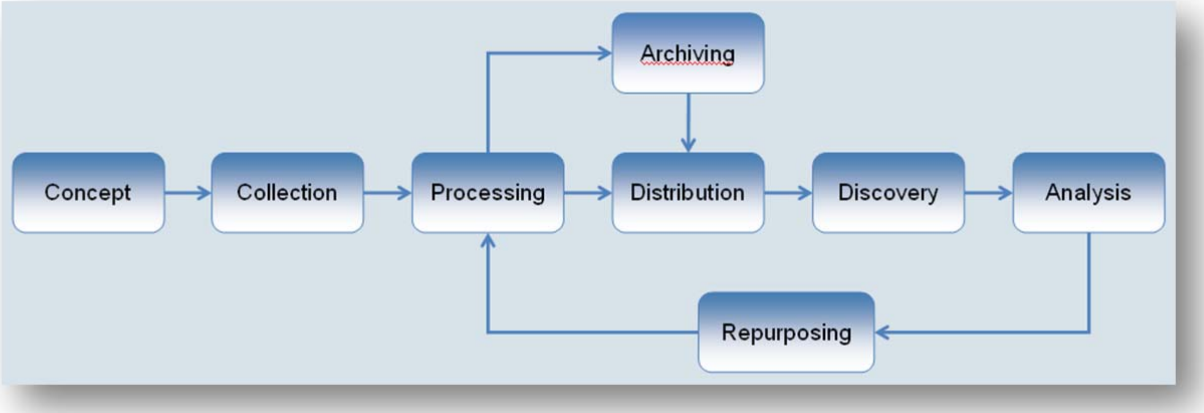

*Figure 1: Research data lifecycle*

6. DDI has two major development lines: DDI Codebook, first published in 2000, and DDI Lifecycle, first published in 2008. There is also a set of controlled vocabularies actively developed and maintained by the DDI community. In addition, an RDF discovery vocabulary for Linked Data is now being developed along with a vocabulary called XKOS, an extension to SKOS that will describe statistical classifications in RDF.

## B. Benefits of the DDI approach

7. DDI was developed with an eye toward efficiencies across the data lifecycle, and there are some salient advantages to using DDI, which we describe below.

(a) Rich content

8. DDI provides for documenting data at both a high and a very detailed level and offers over 800 elements to document complex datasets.

[1] http://www1.unece.org/stat/platform/download/attachments/8683538/GSBPM+Final.pdf?version=1 &modificationDate=1241066597110

(b) Metadata reuse and exchange

9. DDI is optimized for metadata reuse so that information only needs to be entered once and can then be referenced, creating efficiencies. DDI metadata can also be shared across research projects in an interoperable way.

(c) Machine-actionability

10. DDI documentation is tagged and structured, supporting automation and enabling programs to be written against it. This means that the specifications can be used to actually drive systems and can be integrated into the process of designing and implementing surveys and other efforts involving the collection of data. This approach leads to less redundant work; better, easier-to-produce data documentation; reusability of key survey components; increased data harmonization potential; and greater research integrity.[2] Quality assurance is facilitated through this type of standardization. The machine-actionable nature of the DDI specifications also supports repurposing, exchange, and reuse of information so that parts of a DDI document can be exported, for example, as a traditional social science codebook, as a catalog record, or as input to any of the major statistical packages. Fielded DDI metadata may also be used as a foundation for search engines, enabling data discovery.

(d) Data management and curation

11. Comprehensive and robust metadata are critical for sharing data, especially over the long term. DDI makes it possible to describe data transformations so that data users can better understand important issues like the provenance of data items. Further, data transformation can be documented in DDI in both human- and machine-understandable ways so that an audit trail is available.

12. Recent efforts to define and certify trusted digital repositories – e.g., the Audit and Certification of Trustworthy Digital Repositories standard (ISO 16363)[3] and the Data Seal of Approval initiative[4] – also emphasize the importance of good metadata in ensuring long-term access to data.

13. In addition, funding agencies from around the world are underscoring the value of sound data management and curation planning, and metadata are a key component of such plans. The recent memo from the U.S. Office of Science and Technology[5] highlighting open access to research results in the form of both publications and digital data stresses metadata as does the EU Recommendation on Access to and Preservation of Scientific Information.[6] DDI is a robust standard that can help researchers comply with these mandates.

(e) Support for longitudinal data and comparison

14. Large multi-wave and multi-site data collections can benefit from using DDI, and in fact a new reference model called the Generic Longitudinal Business Process Model (GLBPM),[7]

[2] Iverson, Jeremy. “Metadata-Driven Survey Design.” IASSIST Quarterly, Spring-Summer 2009. http://www.iassistdata.org/downloads/iqvol3312iverson.pdf

[3] http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510

[4] http://datasealofapproval.org/

[5] http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[6] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:194:0039:0043:EN:PDF

[7] Barkow, Ingo, William Block, Jay Greenfield, Arofan Gregory, Marcel Hebing, Larry Hoyle, and Wolfgang Zenk-Möltgen. “Generic Longitudinal Business Process Model: DDI -- Documenting the Helix.” DDI Working Paper Series – Longitudinal Best Practice No. 5, March 2013. http://dx.doi.org/10.3886/DDILongitudinal05

influenced by the Generic Statistical Business Process Model, has been developed by the DDI community to communicate best practice in documenting and managing longitudinal projects. The structured DDI approach also enables comparative analyses across geography, cultures and populations, and time.

(f) Support for preservation and platform-independent software

15. DDI is currently expressed in the mark-up language XML, which is stored in the system-independent encoding standard Unicode. This technology is well suited for long-term preservation and also enables the flexible development of software which is independent of specific IT platforms.

## C. Global reach

16. Uptake of both DDI Codebook and DDI Lifecycle has been quite rapid, and DDI is now used around the globe (see Figure 2). Projects using DDI range from the General Social Survey in the U.S. to the Research Data Centres in Canada to Statistics New Zealand. As a result of the International Household Survey Network program supported by the World Bank, DDI is also used by NSIs in over 70 countries, many in the developing world.

17. This global reach is a clear asset for DDI and a solid foundation of support. The DDI Alliance, the organization whose members shape the standard, is also international in nature.

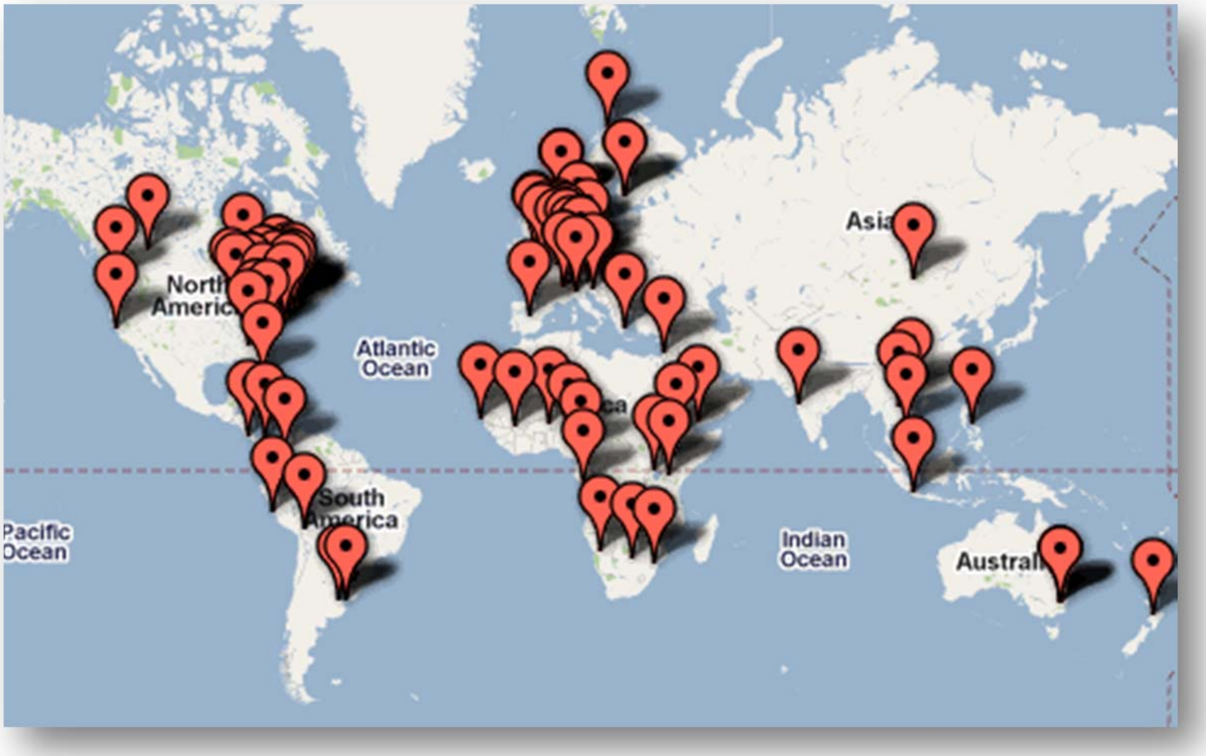

*Figure 2: Organizations using DDI by location*

## III. Priorities for the future

18. As noted above, the data “ecosystem” is undergoing rapid change and shifting in important ways. To continue to be relevant to existing and new communities, DDI must move forward in a strategic way to ensure that it addresses the metadata needs of a broad set of data users and producers. The following priorities reflect the views of the DDI Alliance leadership as of Spring 2013.

### A. Priority 1 – Restructuring for new success

19. As a project of and for the community, DDI operates as a self-sustaining membership Alliance, with each of the 36 DDI institutional members having a voice in shaping and developing the standards. Membership is diverse and includes data archives, national statistical institutes, libraries, data producers, data distributors, research centers, and software developers.

20. This membership arrangement provides for a professional organization that can grow and adapt to new challenges. However, the governance structure of the organization needs to align with the mature organization that the Alliance has become. On the advice of an external consultant, the Alliance has drafted a new Charter and Bylaws, which will go into effect in July. These new Bylaws outline an organization that is broadly representative of the membership and structured to support the effective development of the DDI specifications. There is an Executive Board elected by the member representatives, a Scientific Board that oversees the substantive development of the DDI specifications, and a Technical Committee that creates and stewards the specifications and ensures their usability.

21. The revised Bylaws also allow for the DDI Alliance to be instantiated within the University of Michigan as an organizational host. This arrangement permits the U-M to protect the intellectual property of the Alliance and provides a home for the DDI Alliance Secretariat through the Inter-university Consortium for Political and Social Research (ICPSR).

22. Along with the Bylaws will come revamped formal procedures for making changes to the specifications. This is essential to ensure a controlled and efficient updating of the standards.

### B. Priority 2 – Developing the next-generation DDI

23. In addition to the changes in the organization supporting DDI, we also anticipate changes in the specifications that the Alliance produces. While the Codebook and Lifecycle development lines are expressed as XML Schemas, the DDI and its users will benefit by moving to a model-based approach, like the Statistical Data and Metadata eXchange (SDMX) and Generic Statistical Information Model (GSIM). The Alliance supports this move to a model-based specification as it will provide greater flexibility: the model can be expressed in a variety of technical formats including XML Schema, RDF/OWL Ontology, relational database schema, and other languages. Also, having a model will make it easier to understand the specification, to interact with other disciplines and other standards, to develop and maintain it in a consistent and structured way, and to enable software development that is less dependent on specific DDI versions.

(a) Model design goals

24. A workshop was held at Schloss Dagstuhl, Wadern, Germany, October 22-26, 2012, to focus on gathering requirements for and modeling this new next-generation DDI. Workshop

participants, a group of invited experts, formulated a set of high-level design goals intended to guide development and maintenance of the DDI Information Model [“model”]:

(iv) Interoperability and Standards – The model is optimized to facilitate interoperability with other relevant standards.
(v) Simplicity – The model is as simple as possible and easily understandable by different stakeholders.
(vi) User Driven – User perspectives inform the model to ensure that it meets the needs of the international DDI user community.
(vii) Terminology – The model uses clear terminology and when possible, uses existing terms and definitions.
(viii) Iterative Development – The model is developed iteratively, bringing in a range of views from the user community.
(ix) Documentation – The model includes and is supplemented by robust and accessible documentation.
(x) Lifecycle Orientation – The model supports the full research data lifecycle and the statistical production process, facilitating replication and the scientific method.
(xi) Reuse and Exchange – The model supports the reuse, exchange, and sharing of data and metadata within and among institutions.
(xii) Modularity – The model is modular and these modules can be used independently.
(xiii) Stability – The model is stable and new versions are developed in a controlled manner.
(xiv) Extensibility – The model has a common core and is extensible.
(xv) Tool Independence – The model is not dependent on any specific IT setting or tool.
(xvi) Innovation – The model supports both current and new ways of documenting, producing, and using data and leverages modern technologies.
(xvii) Actionable Metadata – The model provides actionable metadata that can be used to drive production and data collection processes.

(b) Core and base plus modules

25. In terms of design, workshop participants agreed that the DDI model should have a substantive Core along with a set of functions that extend the Core and are needed for specific tasks, such as documenting a simple dataset. In general, these functions would correspond to families of user stories, providing descriptors to cover functional areas. The Core of the model should contain a carefully selected set of foundational metadata objects used by many other parts of the model. To support the model a technical Base will be required in addition to the substantive Core.

(c) New content

26. The DDI effort has been gathering recommendations for new content and coverage for DDI from many audiences and sources over the past few years, and there are now several changes to the specification that we consider important for ultimate inclusion into the model-driven specification. Abstraction of data capture/collection/source, for example, is a key goal. The current data collection module is questionnaire-centric, but we should also be able to describe register data and data in the natural and health sciences (i.e., from technical devices or from laboratory analyses). We envision an abstract layer for data sources with the possibilities for “plug-ins” to handle different types of data.

27. We also want to integrate new content, some of which has already been developed by DDI working groups. New content includes information on sampling, survey implementation, weighting, and paradata; new content related to qualitative data; and frameworks for data quality, access conditions, disclosure review, and data management planning.

28. Also important is documenting process (work flow) across the data life cycle to ensure support for automation and replication. The life cycle model promulgated by DDI has been lacking this process layer, which is essential for the specification going forward.

29. In addition, we want to align with existing standards like GSBPM/GSIM, SDMX, CDISC (the standard for clinical trials metadata), and Triple-S (used by many marketing organizations).

(d) Community-driven development process

30. Leveraging the community in moving forward to create a model-based specification is a key objective for the DDI Alliance. The modelling work should allow for a distributed development effort, with different groups able to work on parts of the model somewhat independently. Also essential is a feedback loop so that development can proceed iteratively.

(e) ISO certification

31. Concurrent with the move to a DDI model, we are also pursuing ISO certification for the DDI specifications. We are encouraged by the experience of SDMX in gaining ISO status and hope to follow the trail that SDMX has blazed.

## C. Priority 3 -- Outreach to the community

(a) Working with NSIs

32. In the past few years, there has been a surge of interest in DDI by many of the national statistical organizations around the world, and the DDI Alliance now counts the U.S. Bureau of Labor Statistics, Statistics New Zealand, the Australian Bureau of Statistics, and Eurostat as DDI members. This is a promising development as our communities can learn from each other, and the Alliance looks forward to welcoming more NSIs into the DDI membership.

33. The DDI Alliance has begun working with NSIs in some notable ways. The SDMX-DDI Dialogue project is helping to surface the similarities and differences between the two standards, and the DDI Alliance has made a formal statement endorsing a collaboration with the SDMX community to enable the two standards to work together. Interestingly, development of the GSIM model is pushing this work forward as mappings are being created to show the relationships among DDI, SDMX, and GSIM.

34. The Alliance has also supported development of the GSIM model and provided representatives to attend the HLG meetings and to participate in a GSIM Sprint. In addition, the Alliance has made an offer of support to work together on an implementation model for GSIM.

35. These projects have paved the way for new levels of cooperation and new partnerships.

(b) Developing an implementation model for GSIM

36. As noted above, one area in which we can work together to great advantage is around development a primary implementation model for GSIM, enabling expression in XML and other formats and providing the foundation for tools to be built.

37. GSIM has many correspondences with the DDI lifecycle specification, giving rise to the idea of DDI aligning closely with GSIM. More specifically, Concepts and Structures (see Figure 3) appear to be the areas where there is most intersection with DDI. This type of content describing questions, concepts, and variables and providing an understanding of what the data are measuring (Concepts) and defining the terms used in relation to data and its structure (Structures) has long been part of the DDI standard, and DDI has appropriate metadata content to offer.

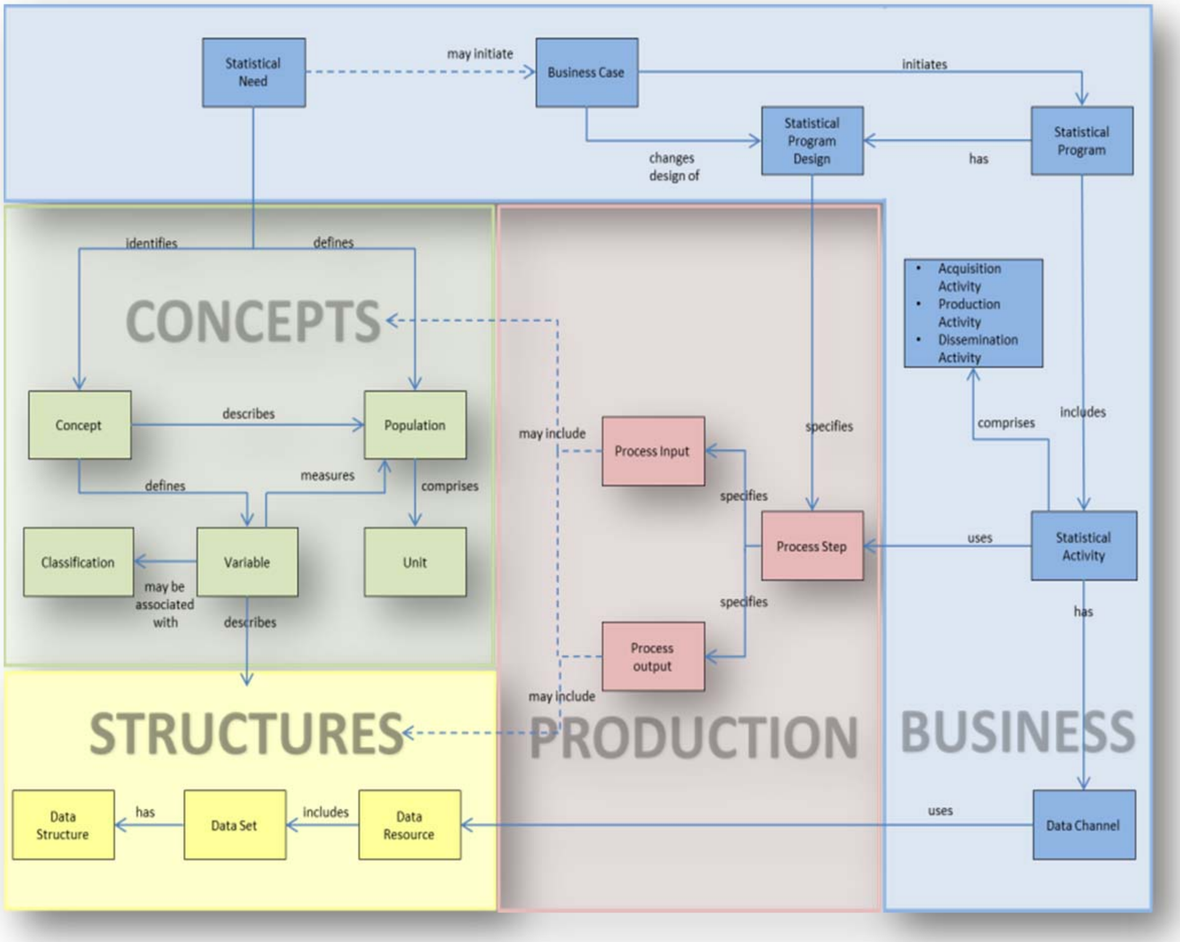

*Figure 3: GSIM coverage*

38. The GSIM Production group is used to describe each step in the statistical process, with a particular focus on describing the inputs and outputs of these steps. Production is also the group in which GSIM has the greatest connection with GSBPM (or with any other model for business processes, such as the longitudinal variant on GSBPM mentioned earlier). As mentioned above, documenting process is a key goal for the next-generation DDI, so this is a great opportunity for DDI and a clear area for synergy, where both GSIM and DDI want a simple, flexible model about processes, process flows, and information flowing into and out of processes. The machine-actionable capability of DDI complements the Production component of GSIM.

39. The GSIM Business group is used to capture the designs and plans of statistical programs. This includes the identification of a Statistical Need; the Acquisition, Production, and Dissemination Activities that comprise the statistical program; and the evaluations of them. As statistical production processes accept as input data from many sources (administrative data, survey data, big data) and result in products and services for dissemination, GSIM ends up describing information objects that "pass through the boundaries" between statistical agencies and the outside world as well as the objects that are internal to statistical agencies. This means that GSIM ends up describing, for example,

- The content of products and services that researchers (or archives) might consume from statistical agencies
- Data sources external to the statistical agency which it might be useful for researchers to harness and which are potentially of interest for the statistical agency to harness

40. Thus, the Business group is possibly the most exciting frontier. While there is currently little overlap or correspondence with DDI, there is the potential that building out this area in DDI would enhance the specification and open it up to new audiences. As we look closer, it becomes evident that we might compare what GSIM calls a statistical need to a research question, which in the academic world triggers a grant application, data collection, publishing and dissemination, and archiving. There are other potential correspondences that seem fruitful to investigate further. We look forward to exploring these parallels in greater detail.

## IV. Conclusion

41. Looking ahead, the DDI Alliance is poised to take on new challenges and to pursue a critical set of strategic directions. Ensuring that the organization supporting the DDI is on sound footing is the first priority and this has largely been accomplished through the drafting of new Bylaws and other organizational documents and policies. Also critical is the move to a model-based specification, which will offer new efficiencies and greater relevance to new audiences. Finally, the Alliance sees the potential for fruitful partnerships, especially with the NSIs around GSIM, which can move us all forward to accomplish our common goals of more efficient and standardized data production and open access to data.