

**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 (METIS)
(Geneva, Switzerland, 10-12 March 2010)

COMBINING METADATA STANDARDS: APPROACHES AND BENEFITS

Paper prepared by Arofan Gregory, Open Data Foundation

I. OVERVIEW

1. This paper summarizes the areas in which implementers are looking at leveraging several standards of interest to the METIS community, to provide support for statistical collection, processing, dissemination, and exchange throughout the data production lifecycle. The paper touches on several different standards: the Data Documentation Initiative (DDI), the Statistical Data and Metadata Exchange (SDMX) standards, ISO/IEC 11179, Dublin Core, ISO 19115, and the many “Semantic Web” standards.
2. These standards are characterized, but the most detailed description is provided for DDI 3, as it has been receiving a lot of attention recently, and is perhaps less familiar to members of the METIS community than the other standards discussed.
3. When using different standards in combination, there must be a clear idea of what benefit is to be realized, and at which stage(s) of the data lifecycle. This paper outlines several approaches for leveraging the standards in beneficial ways, while emphasizing the fitness-for-purpose of each.

II. SOME RECENT EVENTS OF INTEREST

4. There have been many events relating to the topic of cross-standards implementation, recently, as interest in actual implementation has grown. Many of these have been or are being presented either at the OECD SDMX Experts meeting or at this METIS meeting, so I will not go into great detail – this is merely a partial overview, to provide background.
5. For the past three years, there have been workshops focused on DDI 3 held in Germany, at a computer-science center called Schloss Dagstuhl. This past year, in the fall, two topics were explored

which are of relevance: the first was the discussion and beginnings of a detailed mapping between DDI 3 and SDMX v. 2.0. A paper is soon to be published by the DDI Alliance describing this initial work, and the field-level mapping itself will also hopefully be published in the coming months. The second topic of discussion at Schloss Dagstuhl that is of interest was an examination of how DDI relates to the Semantic Web technologies. While some interesting approaches for combining these standards were discussed, no conclusions were reached at that meeting, and no publication is anticipated. However, the DDI Alliance is still actively exploring this topic.

6. Several projects have recently been started to explore the relationship between the Semantic Web technologies and SDMX. These are ongoing projects, but the outcome should be interesting. One project has been undertaken at the Banca d'Italia, and another has recently been launched at the University of Tillburg in the Netherlands. Perhaps most familiar to the METIS community was a meeting hosted by the ONS in February, in which several members of the METIS community participated, along with representatives of various British governmental organizations and members of the Linked Data community. The subject of the meeting was to discuss how statistical data sets could be described using Semantic Web technologies, and after some initial discussion a collaborative project was launched to prototype the approach discussed.

7. The Australian Bureau of Statistics (ABS) has also been exploring the combined use of DDI 3 and SDMX to support the end-to-end statistical process, and has conducted a number of trainings and workshops in the past months. ABS is interested in working collaboratively with other organizations, and this type of implementation is a topic of interest in several other national statistical organizations, as well.

8. There has also been some discussion regarding classification management, in a number of areas. In Australia, many standard classifications are developed and maintained using a system based on ISO/IEC 11179, and there is discussion of a possible project disseminating these classifications also in coordinated fashion using SDMX and DDI. The Canadian RDC Network – which is doing a large-scale implementation of DDI 3 – is working with Statistics Canada, and as part of this project it has been determined that IMDB – an ISO/IEC 11179-based metadata repository – can usefully be expressed as DDI 3 for use within the Canadian RDC Network systems.

9. Census-related projects involving SDMX and DDI are also on-going. Eurostat has been working with many European NSOs around the Census Hub project, which is a large-scale SDMX implementation; for the next census in the UK, both SDMX and DDI have been evaluated, and it is likely that SDMX at least will be implemented; for IPUMS, a project to produce harmonized international census microdata, DDI is being used as a documentation model. These projects are mostly based on a single standard at this point, but some challenges which are common to large-scale census projects (eg, how to handle geographical classifications) are being addressed as a result of these projects. These challenges touch on many of the different standards (SDMX, DDI, ISO 19115) and may result in the publication of best practices or even revisions to the standards in some cases. It is easy to see how users might want access to the data and metadata according to different standard models, given that these large-scale census resources are now emerging.

10. This is not a comprehensive overview of recent events – there are probably many other on-going discussions of relevance which are not touched on here – but these should suffice to prove a point: there is a strong and growing interest in how best to implement standards in various combinations, to achieve different business goals. Further, while this topic has been discussed within METIS in the abstract in prior years, the discussions now taking place are more purposeful: people are looking at how to implement, not merely exploring interesting ideas.

III. THE STANDARDS: COMPARISONS AND EXPLANATION

11. It is not the goal of this paper to provide a summary of all the standards of interest – most of these are well-known to the METIS community already. However, it is worthwhile to briefly characterize the standards mentioned here, and to provide a bit more detail on the recent version of DDI, as this is possibly less-familiar to the METIS community than other standards discussed.

12. SDMX is very familiar – it provides XML and UN/EDIFACT formats for describing aggregate data structures, and for formatting that data; it provides a facility for describing metadata structure and formatting metadata in XML; and it provides a registry-based architecture for working not only with SDMX-based data and metadata, but also for linking this with data and metadata described using other standard models.

13. ISO/IEC 11179 is another familiar standard to the METIS community – it provides a model for describing and managing concepts, data elements, their representations and meaning, and other related metadata.

14. ISO 19115 provides – among other things - the standard metadata model for describing geography. It is worth noting that ISO 19115 is the model used for this purpose within the DDI standard.

15. The Dublin Core is a well-established metadata standard for describing resources, whether digital or physical in form; it is noteworthy that the Dublin Core is widely-used standard in the Semantic Web, and is also natively supported in DDI 3, for representing citation metadata.

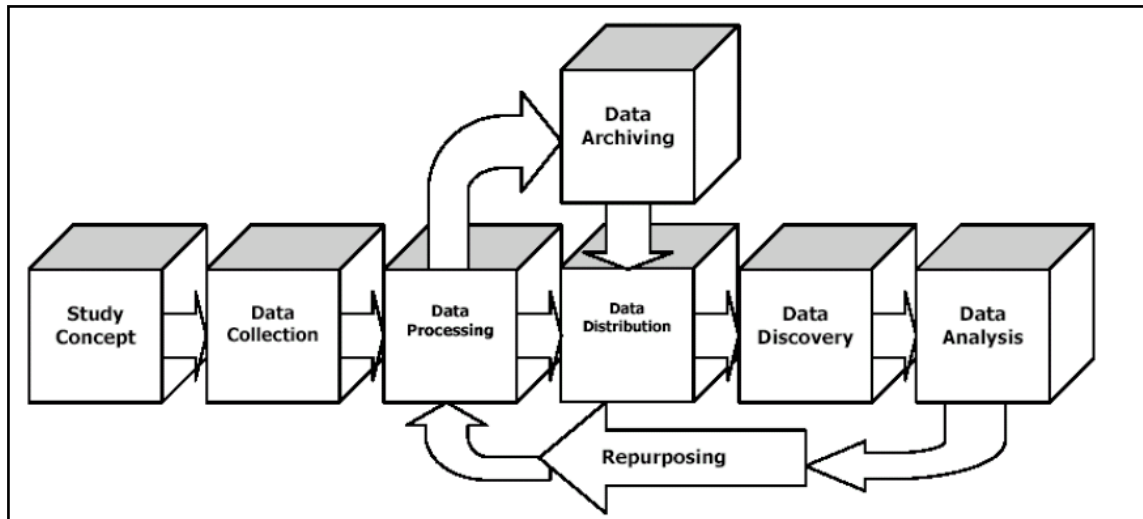
16. The paper “Open Issues on the Semantic Web” presented at this meeting has characterized the Semantic Web standards already, so I will not further describe these here.

17. I would like to give a slightly more detailed description of DDI 3.

18. DDI in earlier versions was developed and used widely by the research and data archive community world-wide. It is used as a standard metadata set by the network of European data archives, CESSDA; it has been widely introduced in the developing world as part of the International Household Survey Network’s Microdata Management Toolkit; and is in common use in North America by a large number of organizations, including many data libraries and Statistics Canada. The goal of early versions of the standard was two-fold: to assist in the discovery of data in an archive, data library, or dissemination scenario, and to provide rich documentation regarding the data once found.

19. DDI 3 has a different – and more ambitious – function. It is designed to capture metadata throughout the data lifecycle, from the conceptual aspects of the data to be collected, through the data collection process itself (whether from surveys or registers), through processing, dissemination and archiving of data, and further to describe the production of aggregate data in the process of tabulation.

20. The standard DDI lifecycle view is shown below. (Note that this is very like a high-level view of the data lifecycle that METIS has developed.)



21. It is worth noting that DDI 3 is not intended only to document these processes, but to provide machine-actionable metadata to support the use of metadata-driven systems throughout the lifecycle. Metadata is intended to be captured at the time it is created, and carried throughout the data lifecycle, benefiting both producers and users of the data at every stage.

22. Other capabilities of DDI 3 include the description of comparisons between different data sets, and the ability to describe the re-coding, processing, and re-use of data as it is tabulated or integrated. Further, DDI can be used to describe geographies, or can be used to integrate geographic shape-files expressed in other formats (e.g. ESRI).

23. It is not coincidental that many of the standards here are well-aligned: in many cases, this was intentional. Some of the standards were developed earlier than others, and have been used by later ones: ISO/IEC 11179, ISO 19115, and Dublin Core were all taken into account in the development of DDI and SDMX.

24. It is also important to understand which parts of the lifecycle each standard was designed to support:

- SDMX is potentially useful throughout the entire lifecycle for aggregate data, but does not describe study concept or data collection for microdata specifically (the SDMX architecture could still be useful here, however.). It is widely used as a dissemination/reporting format for aggregates.
- DDI 3 supports the entire data collection lifecycle for microdata, from study concept through to the tabulation of aggregates. It does not provide an architecture, however.
- Dublin Core – citation metadata – is useful throughout the entire lifecycle.
- ISO 19115 – geographic metadata – is also useful throughout the lifecycle.
- ISO/IEC 11179 is useful throughout the entire lifecycle, in reference to concepts, data elements (“variables”), and classifications. Note that classifications can be generically represented in SDMX and DDI as well, and are also relevant throughout the lifecycle.
- Semantic Web standards are focused only on dissemination – they are not designed to support data production or management.

IV. EMERGING IMPLEMENTATION APPROACHES

25. It is easy to understand that two standards are aligned – what is not so easy is to understand the best way to combine them, to maximize the benefit in terms of the process being supported. This section addresses some of the types of implementations mentioned above, describing the basic approaches being considered, and the anticipated benefits.

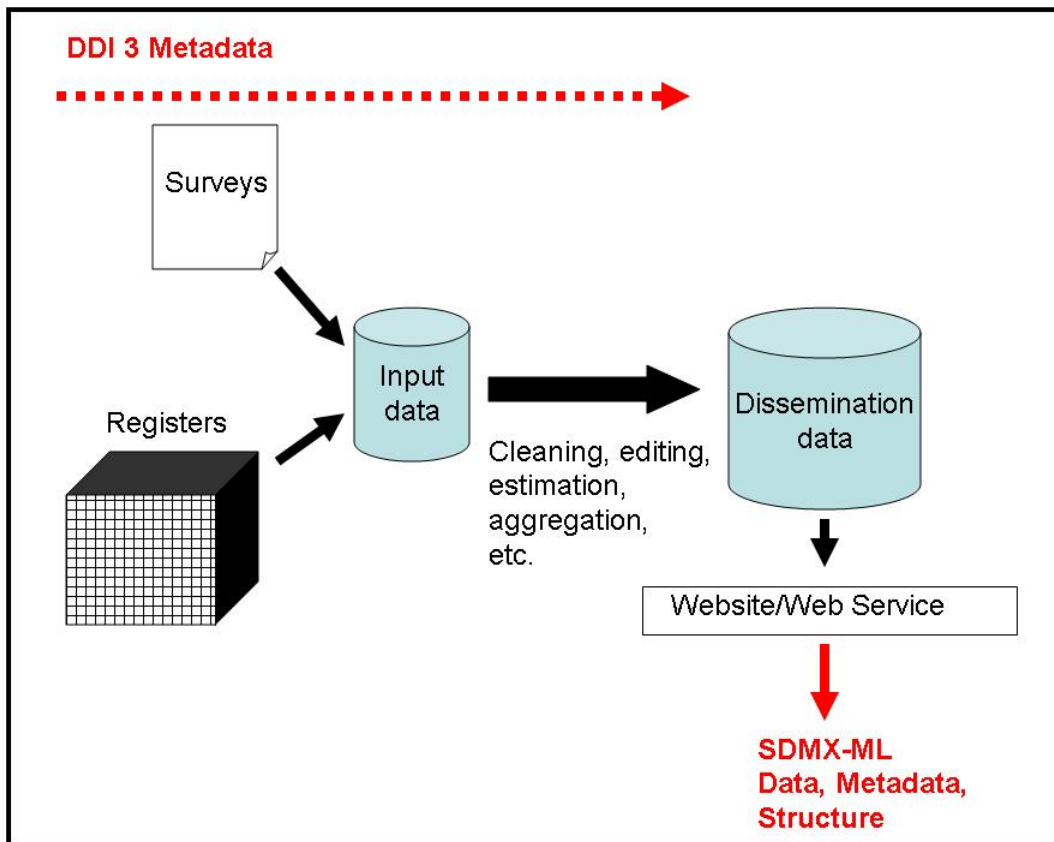
A. DDI and SDMX

26. There are two anticipated cases in which SDMX and DDI could be used in combination: the use of DDI for the collection, processing, and tabulation of microdata, which is then disseminated as SDMX; and the use of aggregates disseminated as SDMX, which will be combined with DDI-documented data by a researcher, using tools which natively understand DDI.

27. These simple use cases are probably not enough, however, to describe the full set of requirements and possibilities: additionally, there is the case where the SDMX architecture is used to support the first use case, throughout the lifecycle. Further, it is very possible that organizations which have based their systems on SDMX would source data from registers which are documented using DDI.

28. These cases will all be considered.

29. The first use case involves using a production system where DDI 3 metadata is used to document and drive the production process, from the collection of raw data (surveys, registers) – stored in an input data repository - through the cleaning, editing, estimation, and aggregation. The resulting data is stored in a repository which drives dissemination and reporting systems (here depicted as a web site or web service).



30. In this scenario, it is very possible that public-use versions of the microdata, documented in DDI 3, might also be made available through the website.

31. However, for users who want the disseminated aggregates as SDMX – or for reporting to organizations which prefer or require SDMX as a format – the data is delivered as SDMX XML. This includes both the structural descriptions of data and metadata, and also the data itself.

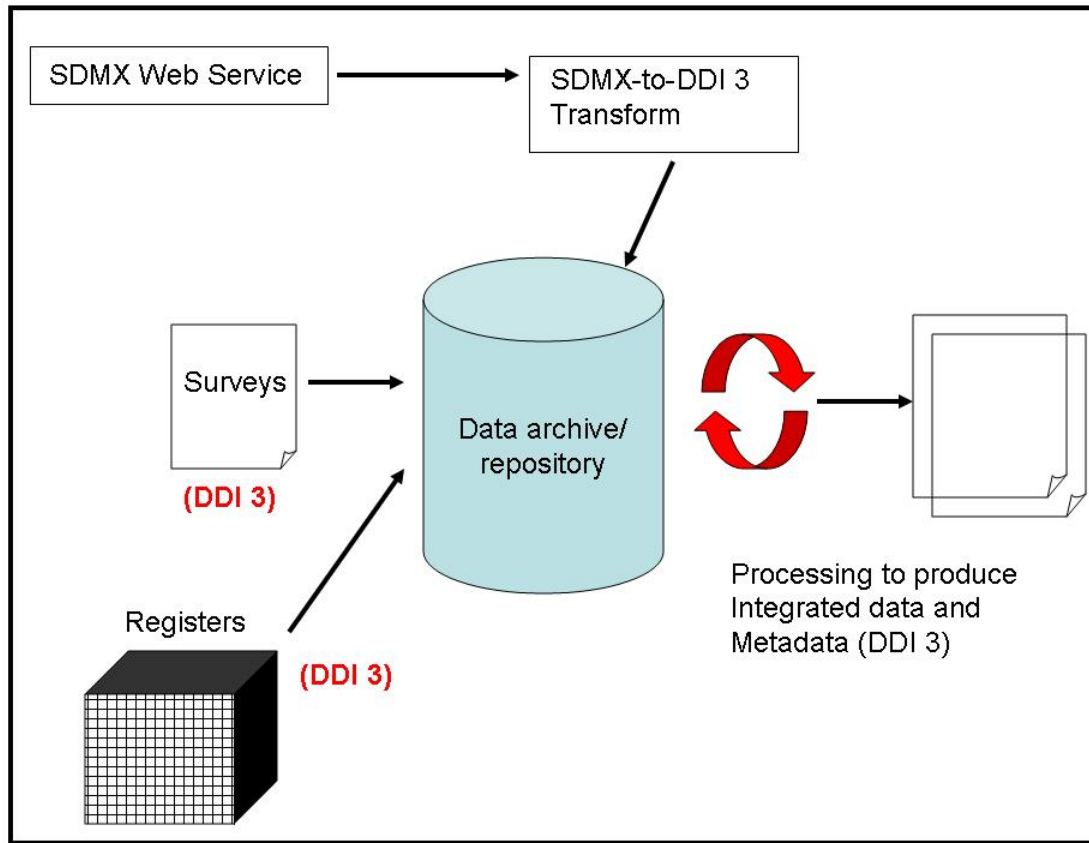
32. Note that the “use” of DDI 3 here could include the metadata structure of the data dissemination repository. While it is possible that direct transformation of DDI 3 XML-formatted metadata could be transformed into the corresponding SDMX structures and metadata reports, this is unlikely – typically, the DDI 3 XML would be an input format for metadata into the data dissemination repository. This in no way changes the needed mappings between the two standard models: both DDI 3 and SDMX are designed not only as XML formats but also as metadata models, and these can be implemented also as databases, etc., and not only as XML.

33. Note further that some DDI-described tables (“NCubes”) will not map acceptably into SDMX, because they do not have regular dimensionality. It is possible to know this from the DDI 3 instances themselves (NCubes are declared to be either “clean” SDMX-compatible data sets or not.)

34. The benefits of this approach are many, and are the typical benefits of using standard metadata models and formats:

- DDI 3 metadata helps to drive the data production process, providing machine-actionable metadata which allows for the use of “metadata-driven” systems. In essence, having a good standard metadata format for describing the data collection and production processes, and the conceptual aspects of this process, support good information management throughout the production process.
- SDMX provides formats that support standard reporting and web-service implementation of the resulting aggregates. From the perspective of the data producer, SDMX here functions as a dissemination and reporting format, and provides a standard expression which users can leverage with SDMX-based visualization tools and web-services clients. This heightens the usability of the data, and also allows the aggregates to be registered in a domain SDMX registry for heightened visibility of the data and metadata. Note that for public-use microdata, DDI 3 could also provide rich documentation for the end user.

35. The second typical use case is one where aggregates available from an SDMX-conformant source (such as OECD.stat or similar web service) are being used within an archive, data library, data portal, or research data center, which uses DDI 3 as its native metadata model. The aggregate data are taken and combined with other sources of microdata to produce an integrated data set by or for a researcher.



36. This scenario requires a transformation from SDMX into a DDI 3 format (for the metadata) and the transformation of SDMX-ML formatted data into a format that can be used by DDI 3 tools (typically an ASCII delimited file or similar).

37. This transformation requires a mapping for SDMX structural metadata (codelists, data structures, organizations, concepts) into the corresponding DDI 3 constructs (category schemes, code schemes, NCubes, organization schemes, and concept schemes). On the data side, the SDMX-ML data must be predictably mapped into a rectangular ASCII file as described in the corresponding DDI 3 metadata.

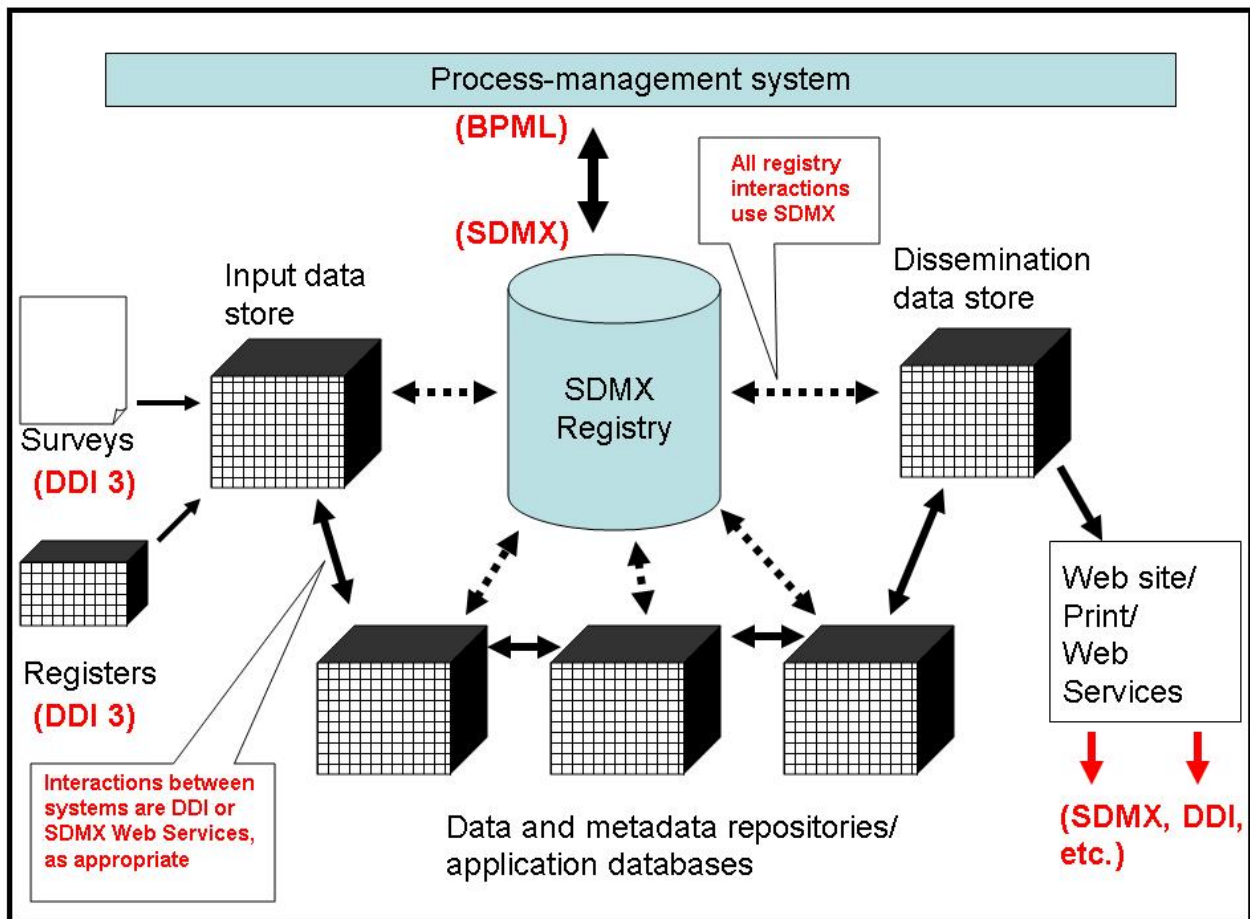
38. There is a limit in this scenario to how much DDI 3 metadata can be derived from the SDMX source. Without being supplemented by additional metadata, some of the “typical” study-level metadata expected in a DDI 3 file will not be present. However, enough will be available to process the data, and make it available in a native format for integration with other microdata. Conformant DDI 3 instances can be produced.

39. In this case, the tabulation of the data is programmatically inferred from the SDMX structure, so that a set of implicit variables are created which can be processed like microdata variables, in the fashion understood by DDI 3 systems. This process does not involve any form of disaggregation, but merely reformats the values into a structure which DDI 3-based tools can recognize.

40. The benefits of this approach are also quite straightforward: the data made available through SDMX web services from official statistics sources can be utilized by systems which natively understand DDI 3, solving many of the problems encountered when trying to integrate aggregated data and

microdata. This type of integration is fairly common for some types of research (ie, economics), where much of the micro-data is not accessible, and aggregates must be used by researchers.

41. The third case is an extended version of the first, using SDMX as a dissemination and reporting format, but also leveraging it for other purposes: the overall architecture and process descriptions are used, and possibly other of the less-known features of SDMX (ie, reporting taxonomies, structure sets). This scenario is an implementation of the SDMX architecture, but using DDI 3 to describe raw data and microdata collection and processing, and having alternate DDI 3 or SDMX expressions of many metadata resources. In addition, it allows for interactions with a process-management system, which might be using other standards such as BPML.



42. As before, the data collection process – whether from surveys or registers – is driven by DDI 3. This process – if metadata-driven – may actually be using information from a metadata repository (concepts, classifications, questions from a question bank, etc.) to drive the collection process. Incoming data would be described in DDI, and registered in a central registry. The data collection events are then visible to data processing events. There is a chain of event triggers which allow automation of the statistical processing of the data production, and this can be managed through the registry via interactions with the business process management layer.

43. It is significant that there may be many different data and metadata stores, and that these – while all interacting with the registry using SDMX Registry Services – may be communicating with each other using either SDMX or DDI, as appropriate. These interactions would use web services protocols, with whichever XML payload suits.

44. The event triggers are created using SDMX Subscription/Notification services. Non-SDMX formats (primarily DDI, but potentially anything) are represented by specific SDMX reference metadata reports, which act as placeholders for the actual DDI files, etc. and give the network location of the resources they represent. Using this technique, anything needed can be natively represented in SDMX for the purposes of registry interactions.

45. It is significant that DDI can describe non-XML data formats (SPSS, ASCII, SAS, relational data stores, etc.) because these may be the medium for transferring data between some processing applications. In other cases, XML may be the preferred data format (supported primarily by SDMX, but also DDI). This allows tactical decisions to be made about how data are described in the accompanying metadata, without requiring one standard format over the other.

46. Metadata repositories will be modeled not after one standard or the other, but must be able to produce either format on demand – they use a “superset” model. Thus, the correct mapping between SDMX and DDI is fundamental to this approach, along with any other standards which are in use (ISO 19115, ISO/IEC 11179, etc.).

47. The benefits of this approach are several:

- Because the SDMX architecture is one that emphasizes the use of Web Services communications between a set of distributed applications, it provides great flexibility in how the overall system is implemented – this can be done in an incremental way. All the applications are tied together using the event-trigger mechanism enabled by a centralized SDMX Registry. This also eases the migration path for individual applications at any point in the process. This type of architecture is well-supported in many technology platforms, being based on mainstream service-oriented architecture.
- The choice of tools based on one standard or another can be made according to the best fit – there are no compromises needed on the basis of support for either SDMX or DDI.
- All of the benefits seen in our first use case are also realized – good documentation for users, the ability to deploy metadata-driven systems, and the provision of standard formats for dissemination and reporting.
- Integration with the business process layer is facilitated by having processes described in the SDMX Registry and represented in DDI as lifecycle events. Because the registry acts as a centralized point for dispatching event triggers, these events can be detected and reported to the process management layer. This provides a strong foundation for managing process, by making all significant data-production events visible.

48. Our last use-case is a fairly minor one. In the case where a business register used as a source of administrative data is documented already in DDI 3 (this is the case in some organizations today), and the register data has to be passed to an organization using SDMX-based systems, it is possible to create a mapping between DDI 3 and SDMX. The combination of data and DDI 3 metadata can be transformed to create a specialized form of SDMX, in which the data set has only two dimensions (unit record identifier and a measure dimension). This is a cross-sectional format, typically with a large number of measures – one for each variable in the register data.

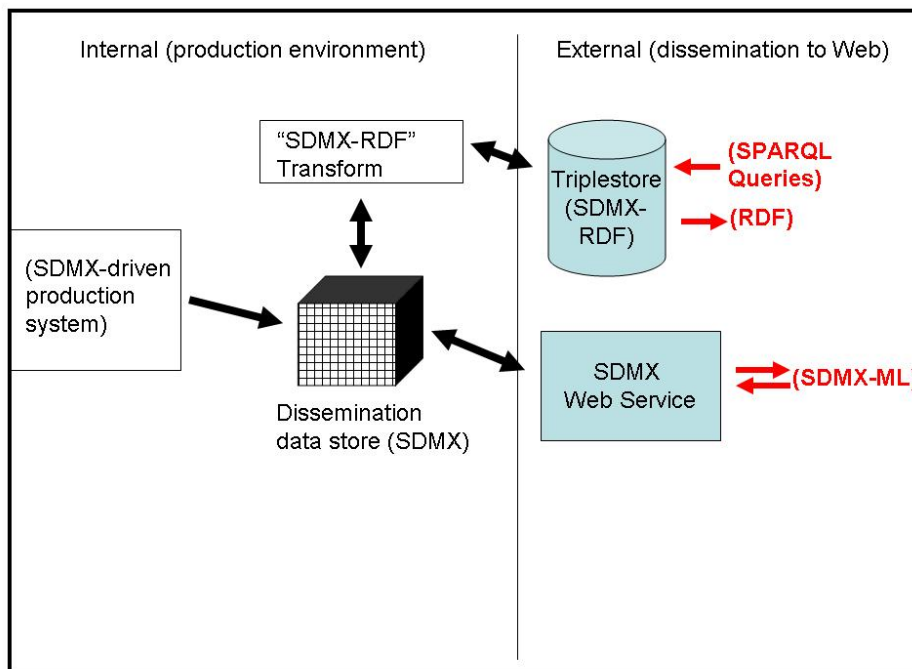
49. The only benefit to this approach is that it allows the use of DDI on one side to enable the easy creation of SDMX for the organization collecting the data. The data collector does not have to implement support for a standard which may not be otherwise used within the organization. This case is one which has some limitations – it does not carry all of the knowledge about the register data captured in DDI very effectively – but it may be useful in some implementation scenarios.

B. Relationship to the Semantic Web Technologies

50. This section will look at how SDMX (and potentially also DDI) could be used to help drive the publication of data sets and accompanying structural metadata using Semantic Web technologies and standards. The approach described here is one that has been envisaged at the ONS meeting on the subject, which confined itself to discussion of the SDMX model for aggregate data, but a similar approach could also be used with DDI metadata and microdata.

51. It is important to note that the Semantic Web standards are concerned only with the publication of data and metadata in the Web environment – these technologies are not designed to support data production or reporting. The emphasis is on being able to discover resources on the Web, and to be able to know enough about how that resource relates to other information to be able to use it effectively.

52. Also – a note on terminology: when the terms “data” and “raw data” are used in a Semantic Web (or “Linked Data”) context, they do not mean the same thing as when used in the context of a statistical data producer. “Data” is – in Semantic Web terms – anything that can be described using RDF (which is a very flexible tool for describing things!). The term is not used to refer only to numeric data. “Raw data” is another term which unfortunately has caused a lot of confusion. For a statistical office to publish raw data – that is, the numbers as collected directly from a register or through a survey – would in most cases not only be irresponsible (since it needs to be cleaned and edited to be useful) but also either unethical or illegal, given the codes and laws about data disclosure and confidentiality. When used in a Semantic Web context, this term means something approximating “public-use microdata”. This is an issue that perhaps deserves further discussion with members of the Linked Data community, but there has been some confusion caused by these terms, as understood in their respective communities.



53. In this scenario, an SDMX-capable production system is using an SDMX Web Service to disseminate SDMX-ML data and metadata. It is only important here that SDMX be implemented as an output – the database itself need not be a “native” implementation of the SDMX model (as pictured here), so long as it can express data and metadata as SDMX-ML.

54. This case assumes the creation of a standard vocabulary or ontology that describes the SDMX Information Model for use by RDF-based technologies. It is worth mentioning that some parts of the SDMX Information Model have corollaries already within the Linked Data community, so not every part of the SDMX model would need to be standardized in this form: it is very much a part of the Semantic Web culture to use existing vocabularies when they already exist. In this case, vocabularies for describing time, geography, concepts, etc. already exist and may be useful. The name we have suggested for the SDMX-based vocabulary here is “SDMX-RDF”, but that is just a shorthand for the purposes of discussion.

55. Given the existence of such a standard vocabulary for describing SDMX in terms meaningful to the Semantic Web technologies, it would be possible for a standard transformation to be implemented, as an additional part of the typical SDMX Web-Services packages. Such a transformation into RDF could be used to populate a “Triplestore” – a database designed to publish RDF triples, and to act as an end-point for SPARQL queries, which are the normal means of finding RDF resources (there are many different flavors of this type of technology - this is used as an illustrative example.)

56. One consideration for implementation is the relative file size of data sets when expressed in SDMX and RDF – in one prototype, a 4.5 Meg SDMX-EDI (GESMES/TS) file was transformed into 45 Megs of “compact” SDMX-ML and subsequently produced 420 Megs of RDF triples. This suggests that the best way to publish RDF triples is using a dynamic, queryable mechanism rather than publishing full data sets as static files.

57. Note that there is still a lot of work to be done to understand all the issues surrounding this approach, but at this stage these issues do not seem to be insurmountable. There is much more work to be done, however, before these issues are completely understood.

58. It is also worth mentioning that it may well be possible to take “SDMX-RDF” data and metadata and transform this back into conformant SDMX-ML. This is a possibility that will need to be further explored.

59. There are some obvious benefits to the approach described above:

- SDMX data sets and metadata can now be utilized by a new set of technologies in a Web dissemination scenario. The Semantic Web technologies have received a lot of attention within the technology world, and are popular among academics and in some other user communities. Note that the hard work of producing good data and documenting it sufficiently is not being done using RDF – it is being done using familiar standards and best practices – the creation of the RDF triples is an automatic by-product of having systems capable of producing SDMX-ML.
- From the perspective of a data producer – as opposed to the developers who create the SDMX Web-Services packages – no deep understanding of RDF is required. The fact that there is a standard implementation, which requires only that a system be capable of producing SDMX-ML, means that the benefits of publishing statistical data and metadata into the Linked Data Web can be realized fairly cheaply and easily. RDF and related technologies are typically not well understood by IT staff within data producing institutions, and this approach accommodates that

reality. The cost of better-exposing the data and metadata as “Linked Data” becomes relatively moderate.

- For the Linked Data Web, having the producers of official statistics publish their data in this way would be of huge benefit. Provenance is a major issue when dealing with data discovered on the Web – how can the quality of the data be known? If organizations whose mission is the dissemination of high-quality statistics make their data available, then this problem becomes much simpler: the data can be obtained from the source directly, with the accompanying assurance of quality.

C. Geography and Other Classifications/Concept Systems

60. One subject which touches on almost all of the standards we have mentioned is the publication of official classifications. Classifications are used for different purposes by different standards: in SDMX, classifications can be used as concept schemes, as codelists and hierarchical codelists to represent concepts, and as category schemes for the organization and navigation of data sets and other resources. In DDI, classifications are typically used as representations, but there are some specialized applications, as well – geography (as for ISO 19115) and as search terms associated with variables. ISO/IEC 11179 is used as a model for developing, maintaining, and modeling classifications, and such classifications also act as the value domains and representations of data elements.

61. One major issue which faces the users of classifications is provenance – users want to know that they are using the official (and often most recent) version of a particular classification. While there are many places which re-publish standard classifications in various formats, this is sometimes problematic: users want to know that they are using the correct version as published by the official maintainer of the classification. Further, if different “unofficial” versions are used, this can impact data comparability, as the unofficial versions may have been modified for specific purposes.

62. If mappings between the various standards exist, and a best-practice is established in this area across standards, then it would be possible for the classifications to be published by their maintainers in several formats: as SDMX, as DDI, as ISO 11179-conformant formats, in RDF (according to the posited “SDMX-RDF” vocabulary), etc. This would be of great benefit to the users and producers of data, both. This idea is already being explored by some organizations, but it will only be possible if an agreed best practice can be developed among the various standards bodies.

V. FUTURE WORK – SOME IDEAS

63. This paper examines several approaches for using standards together for the collection, reporting, dissemination, processing, and management of statistics. In each case, the work is on-going, but it is worth summarizing the places where there is an obvious need for more work within the METIS community and in liaison with other related communities:

- SDMX-DDI mappings are being created: these need to be endorsed by the various communities as a best practice if they are to provide the greatest benefit (off-the-shelf tools, etc.)
- An “SDMX-RDF” type of vocabulary needs to be developed and housed in a reliable standards body. This might be SDMX, or it might be some other related organization.
- A harmonized model of statistical data – both microdata and aggregates – could be created as the basis of an RDF expression. This could be a joint harmonized subset of all the standards discussed, with an emphasis on SDMX, DDI, and ISO/IEC 11179. If such a model existed, it would make the use of statistical data disseminated as RDF much more useful for those who don’t understand the various standards with the same depth as the data archives and data

- If this standard statistical vocabulary existed, then the creators of SDMX and DDI tools should be encouraged to implement it, helping to hide the complexity of the Semantic Web technologies from data producers.
- The maintainers of statistical classifications should be encouraged to publish their classifications in a set of useful standard formats (SDMX, DDI, ISO/IEC 11179, etc.) Toward this end, a best practice guide should be created, possibly accompanied by a set of free tools, which would make this publication simpler for the maintainers of classifications.