**Title of session**: Binding

**Day**: Wednesday

**Participants**: Klaus Blomqvist, Gregg Kellogg, Barrie Nelson, Nick Car, Gary Berg-Cross, Guillaume Duffes, Jared Lyle, Steve McEachern, Dan Gillman, Dierdre Lungley, Wendy Thomas, Achim Wackerow

**Chair**: Achim Wackerow

**Note taker**: Wendy Thomas

**Background information**

* Approaches of modeling and/or generation
* Usage of other vocabularies
* The process you use to create specifications

OUTCOME: document outlining why other specs have chosen their particular binding, the process used to create them, their rationale, and approaches to new releases and versioning; how to deal with issues of identification and round-tripping of data and metadata

CSVW:

* RDF is a preferred format for CSVW
* Expressivity is important so that you need the one that is the most expressive. If you have a less expressive language, then you run into problems with round tripping. Semantic round tripping is much more achievable with RDF.
* Lossless with respect to what? Depends on how much you put into it
* Properties and attributes of relationships may be more difficult to bring across (or round-trip). A need to add an intermediate node.
* We need to look at the cases where you want to do a round trip. Full or minimum depending on the purpose. For example, some people only want the data in the cells as RDF therefore losing some of the content.

NSO perspective (GSIM)

* Statistics Sweden
  + Focus is on SDMX and DDI expressions but there are other standards for dissemination such as an output data base. We have our own world and tend to be inward looking so interested in what is happening in the rest of the world. They have to deliver SDMX by law to EUROSTAT via XML. Looking at those data structure definitions to make their National Account, data structure definitions and class descriptions.
  + GSIM is used as static conceptual model. Use RDF for the documenting the record as well as XKOS for classifications
  + Finland and Norway are implementing GSIM classification system which maps to their own metadata standard. Discussing move to GSIM and DDI as an overall metadata structure.
  + Need to be able to export to others (XML, RDF, spreadsheets, Excel files)
  + Those working with dissemination are focusing on RDF
* INSEE
  + INSEE is moving away from Excel files for quality control purposes so are using RDF as a core, canonical model in its central repository. Classifications were already centralized. Concepts were also easy.
  + XML is for exchange to EUROSTAT
  + INSEE Classifications in SKOS, XKOS, and some Dublin Core
  + XML could be limiting but XML is very important for exchange archival purposes.

Issues regarding capture tools – Excel

* BLS building a taxonomy of terms for data dissemination managed in Excel. Excel and SAS are the go-to applications.
* Excel is important for distributing to users and unfortunately a tool to capture.
* Note that Excel is strongly related to CSV. People in general are trying to move away from Excel.
* In using a simple entry such as Excel it is an effort to add on the needed semantics.

Issues regarding binding options

* XML is often a required exchange format
* XML is an important archive format
* Not all organizations use XML (One of the reasons in Statistics Sweden didn’t initially go to DDI because it was in XML only.)

Differences in specifications

* With DDI4 you have UML and bindings, with CSVW there is a specification and class diagram but not a model. You can use a URI can be associated with a row, a cell, or value.
* It can be used internally simply as CSV’s. You can use the CSVW tools to validate or move to RDF to ….?
* Having a concrete means of moving from CSVW to RDF-OWL

Looking at MethodologyPattern package

Richard C wrote rules on how to write the transformation. Another person wrote the transformation. One ontology was written according to these rules. Intent was that SHEX or SHEKEL could be used to validate.

The set of things that are people are disjoint from a set of things that are organizations

CDIS – keeping bindings in sync

With the metadata repository using Semantic Manager. Beginning to expose to developers. Export it as PDF, Excel, text, flavors of XML, []. These are produced directly from the repository. Every times changes are made these are regenerated. Also associate controlled vocabularies with specific columns which is managed elsewhere and generated by the EBS system. There is an RDF format from this but unsure of quality. Also XML, Excel, text. Released quarterly. Will be building terminology into the share repository so it is all available from the single source. Looking as use cases that would take a specific slice across the standard to support specific uses. There is still the question of how to curate this content. Looking for partners for curation (needs to be well curated and well formed).

Are you looking at RDF for SHARE – was considered when selected was made (commercial, off the shelf tool) but found no solution. Have already hit some problems with data base and XML artifacts such as speed. RDF is being tested. NCI is also looking for a replacement to CSVAR{?} and are looking at RDF. Currently is an Oracle data base.

It is the model and the standard specification with guides for implementation of the model. The class and domains are represented in the metadata repository. The model was initially formalized in Excel and are now in Bridge. High level domain analysis modes in Bridge. The Bridge model does not currently contain all that is needed for clinical research.

Currently the standards are managed independently, (analysis, etc.) goal is to harmonize and bring together. SHARE will give them the discipline that they haven’t had before so that changes in one part will analyze the impact across the board. Versioning is done at the standard level. Looking at an alternative at the domain or class level. Specific changes are noted as dif files and are published as part of package.

Talking to FDA about using the SHARE as part of their review process.

SHARE = Semantics Manager (Oracle Database)

Have not gone back to the beginning in SHARE but versions listed in FDA requirements are in SHARE. Earlier versions will be maintained in perpetuity. Ex. Adam 2 versions, others with 2 versions and 4 versions. Versions take 2-3 years to publish but this should be faster with SHARE…perhaps an annual release strategy.

What is the reasoning on versioning? The class library and functional views and these get versioned but there are dependencies. (DDI). CDISC has no clear answer on this. Organizations must tell FDA exactly what version they’re delivering. Have added capability to support versioning at the class level. The FDA are publishing a set of rejection criteria regarding quality of data and conformance to standards.

Example: Within Adverse domain variable severity. There are identifiers, topic, timing, additional qualifiers.

CDISC work is easier because you have the FDA being prescriptive which pushes people to adopt the standard because FDA requires it.

GSIM

INSEE - Took the XML formats and transformed to RDF-OWL. There are 2 papers for CENSTAT (Connect these) describing difficulties of going from XML to OWL 95% is easily convertible but 5% is not. Providing recommendations to update the UML improve and correct their UML usage. All the XSLT transformations are published. Similar work was done for GSBPM so there were already an XSLT was already done. XMI to simplified XML and then to RDF-OWL (taken from EA). The second step was done just recently. Rules are imbedded in XSLT as it was a one-off action. It is well documented so you can read the comments. The document contains some mappings between GSIM to RDF-OWL with deeper descriptions.

Main issue is the possibility in UML of giving the same name to different relationships. The property names have to be unique (cross usage must have same property name, datatype, cardinality). You can set a restriction on an OWL class within a domain. Restrictions are OK.

GSIM is conceptual so this means a leap directly from conceptual to implementation so is it detailed enough?

GSIM still stands as a strong marker as to how we should think of the structure of metadata and the DDI moving forward work is following this in a pretty strong way. So it provides a good guide. Should concentrate on the G in GSIM which is generic. Regarding the work done in EUROSTAT with GSBPM this has been very useful at the management level. What LIM is trying to accomplish is very, very difficult because of the differences at the local level. How can we build a CSPA service for what’s in GSIM. DDI could be seen as a way of implementing CSPA services. For GSIM to be useful it has to be generic. You need national level specification. Makes it much easier to talk to each other. Could use the GSIM to implementation to identify the common sub-set of broader DDI or other implementation models that need to be expressed in a common way to support CSPA/LIM goals. We need to continue to develop these in a synergistic way.

In the generated RDF is there use of additional vocabularies. They added languages like FOAF as necessary in the transformation process. See appendix of table for attribute mapping.

Are there any tools available to transform from UML to RDF-OWL? No. UML modeling is not as prevalent in the RDF community. Simon Cox has had similar experiences in going from UML to RDF-OWL.

Some people say model in OWL but you can’t generate the pretty pictures. FIBO (Financial Industry Business Ontology) is being produced in OWL. It’s not easy to read. There is a Visual OWL but simply a visualization. CMAP (Hayes).

*[Links to any relevant documents*]

An OWL Ontology for the Common Statistical Production: Architecture

[**http://semstats.org/2016/content/an-owl-ontology-for-the-common-statistical-production-architecture/article.pdf**](http://semstats.org/2016/content/an-owl-ontology-for-the-common-statistical-production-architecture/article.pdf)

An OWL Ontology for the Common Statistical Information Model GSIM Design and Implementation

[**http://semstats.org/2016/content/an-owl-ontology-for-the-generic-statistical-information-model-gsim-design-and-implementation/article.pdf**](http://semstats.org/2016/content/an-owl-ontology-for-the-generic-statistical-information-model-gsim-design-and-implementation/article.pdf)

**Decisions**

*[Please note any proposals agreed]*

**Issues requiring further discussion**

*[Please note pros and cons for different positions and the reasoning for them]*

**Is there a need for a continuation of this discussion?**

*[Possible options are further session, plenary, or evening session]*

**Is there a need for a longer document to continue this discussion?**

Barrie – PDF sent to Arofan (get and add as reference) – also wiling to contribute to an overview of the different approaches or are links sufficient?

Could work with CDISC to coordinate on a live version of the CDISC guide for academic users.