DDI 4 as a Cross-Domain Work Product:

Lessons from the 2018 Dagstuhl Workshop, Week 2

DDI Sprint, Berlin, 27 November, 2018

Introduction

This document attempts to summarize the lessons (to become the basis of business requirements) coming from the workshop at Schloss Dagstuhl regarding how DDI could be used for cross-domain data integration. While this is a specialized application of the standard in some senses, it is an important one in a world where cross-domain data sharing is becoming increasingly of interest to many scientific domains, the social sciences only one among them.

At the same time, it is clear that the focus on data management which DDI encompasses is unique among the various metadata standards considered at the workshop, and that there are nuances in how data is described across the life cycle which other metadata standards have not analyzed completely or understood, and which they do not support. In short, this document focuses on those areas where DDI has something to offer the world in terms of cross-domain data integration.

If DDI 4 is to become a useful tool for the purposes of sharing data between domains, the requirements of this audience become good requirements for the further development of the standard, not to do everything for every application, but as a tool for working with other applications which often already have their own standards for domain-specific use. DDI becomes one of the standards which form a cross-domain model used to bridge between domains.

It should be noted that these business requirements are for the standard as a product, aimed at developers, system architects, and others with that level of skills. These are *not* requirements for those to whom we might try to market the standard (researchers), most of whom cannot (and cannot be bothered to) understand such technical artefacts as UML models, XML schemas, or RDF vocabularies. Our primary audience consists of those individuals who specify, design, and build infrastructure for the end users, the researchers. (A marketing plan is different from a requirements document for a reason!)

Important DDI 4 Features

1. **DDI 4 as a Simple UML Model**

The fact that DDI is expressed and documented as a UML model which does not rely on the more complex features of UML, but uses mostly common ones, is a major benefit. Other domains may be far more focused on technologies other than XML (RDF being the prime example). There are some very different perspectives on what data is and how it is managed. Regardless of these domain-specific aspects, UML serves as a lingua franca for communicating the ideas around data found in DDI 4.

At Dagstuhl, with experts from domains as diverse as astronomy and geography, we found that the core aspects of the DDI 4 model were immediately apparent to those who looked at it in this form. UML is a precise and widely-understood formalism which is much more useful than XML Schema or OWL/RDF-S in conveying the exact ideas of a design.

The style of UML used as a canonical expression of DDI is therefore important: it should be simple, and as useable as possible within different UML tools, as these vary from domain to domain. Special rules regarding the meaning of the model (the specialized "implements" relationship for patterns, for example) should be avoided.

A standard based on a conceptual model in UML can be understood as taking a sustainable approach. In the same way that current representations in OWL/RDF-S and XML Schema are generated on this basis, other and future technical expressions can be derived from the UML model as well.

2. DDI 4 as a Library

In earlier versions of DDI, the whole was treated as monolithic, which was extremely confusing to some users. Once the learning curve had been scaled, the monolithic whole in fact may prove simpler to use (Colectica has voiced this opinion as regards the XML schema expression of the model in the past), but as a conceptual model it needs to be modular. DDI 4 is currently organized primarily according to function, and although the organization may not be optimal in some ways, the fact that you could look at a single piece of the model and understand how it would work for a specific function proved to be sound.

The idea of having Views into a Library is a solid one as defining a subset of the Library for a specific purpose, even though its implementation up to this point has been problematic. The business requirement for it is clear, however – for cross-domain purposes, the ability to use specific functions from a library of them, and not the whole of the standard, is the correct approach. It should be noted that many other domain standards use this approach to a much greater degree than in the current DDI 4. This shows us that standards packaged as libraries are feasible and, going forward, it is the implementation of this idea in DDI 4 that requires additional work.

3. DDI 4 Datum-Driven Data Description

At the heart of the DDI 4 model is the idea that data can be viewed in an incredibly granular way – as a single datum surrounded by the points of information which give it context. These granular objects can then be assembled into different higher-level structures for different cases (rectangular microdata, event data, aggregate data, etc.).

This idea makes DDI very powerful in a cross-domain data sharing scenario, especially since domain perspectives on data vary so widely. The traditional DDI culture of use is to see data as managed datasets, which are frozen and versionable snapshots taken of the data as it moves through the lifecycle. This is a powerful perspective. It does not help in domains where data is seen as evolving streams coming from services or sensors, however. It also does not work well for scenarios where individual data points are attached as supporting information to transactional structures, or structures whose main focus is on describing something else (as in the case of some applications in the geographical domain).

The datum-driven approach is flexible enough to support all of these cases and was seen at Dagstuhl as very useful. That different higher-level structures in different domains might be needed is true, but agreeing on a flexible low-level description of data provides a good starting point for bridging between such structures.

Moving ahead, it would be significant that DDI 4 contain as many different higher-level structures as possible in the form of standard formalizations. One formalization that we are missing in DDI 4 that may require fresh thinking is the representation of aggregate data and the relationship between DDI and SDMX. This work would need to be both modeled and also documented in a very clear way (a good example of this style of documentation was the slide deck on the Variable Cascade which was done during the first week in Dagstuhl this year).

The datum-driven approach is not only flexible and powerful enough to describe all these different forms of data, it can also be used as an opportunity to define a single way to describe data across domains (and standards) which will support cross-domain data integration and analysis.

4. The Variable Cascade

The Variable Cascade shows the evolution of data as it moves from the stage of conception and design through to a final "published" state. This way of understanding data across the lifecycle, while still providing unifying objects for managing the many different incarnations of the variable as a single entity, is a strength of the DDI 4 model. (It has already, in fact, pointed up some of the failings of cross-domain specifications such as DCAT in how they understand data.)

Given this strength, it is appropriate for future evolutions of DDI 4 to not only continue to include it (a given) but also to consider how well it can be used in describing unfamiliar types of data in other domains. (We may find that such constructs as sentinel values might have corollaries in other domains which are not tied to the same analysis tools used in the DDI world, etc.)

It is important that DDI shares this basic construct with GSIM, coming from the world of official statistics, and that official data are often in demand in cross-domain scenarios (the disaster risk mitigation use-case at Dagstuhl placed a huge emphasis on official data sources, and all of them relied on demographic and census data, to cite one example.)

5. Concepts and Classifications

The dominant cross-domain model for describing concepts and concept systems seems to be SKOS, as we learned into our initial forays into RDF. However, even then we needed to extend SKOS to support our own domain use cases. It is not a sufficient tool for cross-domain data integration. The richness of the DDI 4 model around concepts and classifications is a strength, and one which can perhaps be further built upon. Formal statistical classifications within the traditional DDI community are well-behaved and well-understood things when compared to classifications and similar concept systems as used in other domains. DDI 4 should make sure that the constructs in the model are both general enough to support use cases in other domains, and to continue to provide support for the demands of formal classification management within the statistical world.

The multiplicity of similar constructs in the existing DDI models is perhaps something which could benefit from rationalization, or at least more complete definition and documentation. There are many different types of concept systems across the various versions of DDI work products which are potentially confusing: keywords, thesauri, concept schemes, classifications, statistical classifications, controlled vocabularies, etc. DDI 4 should have enough complexity to support needed cases, and no more.

6. The DDI 4 Process Model

The DDI 4 process model is one which has been synthesized from many different examples, including W3C specifications, BPMN, and elsewhere. This is in many ways its strength when seen from the perspective of cross-domain data integration: it is generally familiar to people from different domains. Provenance is one common case which receives a lot of attention, although a formal standard expression in the popular Prov-O vocabulary has not been defined (which we may wish to explore moving ahead). There are some other strengths, however: the DDI 4 model supports multi-threaded processes, which is a more important feature in some domains than it is within social science.

The integration of process and data is another aspect of this which is important in cross-domain data integration scenarios. At Dagstuhl, there was some discussion of disaster risk recovery and mitigation. The tools used for these purposes are sometimes similar to those used in social science research (data is fed into modeling frameworks) but in other cases is a real-time asset which is tied to process flows in ways we do not commonly see in Social Science research. These requirements might drive some changes in how we relate data and process in DDI 4.

7. Survey Questionnaires

Another strength of DDI – although arguably not yet in the current DDI 4 model – is its ability to describe questionnaires. While the emphasis in the DDI 4 development has been on describing non-survey data sources (registry data and “measurements”), other domains which actually deal with these as their primary data sources are likely to have more mature models for this. What DDI does better than other standards, however, is describing survey questionnaires.

Survey questionnaires are used in many different domains. It may be a good idea to emphasize this strength of DDI in making DDI 4 a better tool for cross-domain data integration, and to try to learn from other domains how best to address sensor data and non-survey data sources.

Summary

This document does not present business requirements as such. Rather, it suggests those areas within the DDI 4 development which might best serve as areas of focus in the immediate term. In many ways, the specific features of the model listed above are important for many different applications both within the DDI community and in external domains, and as such have already been a point of focus. This is good, but at this point we do not have a finished product.

What emerged from the workshop at Dagstuhl was that we need not only to come up with good core models, but also to apply them in a way that is immediately useful. They need to be finished and published in a way that is easy to understand by their intended audience, and one which is free of assumptions based on both domain and technology platform.

A generalized DDI 4 which would serve the needs of cross-domain data integration is not a long-term goal – it is within easy reach if we can take what we have and work it into the appropriate product. This does demand, however, that we consider the requirements of cross-domain data integration as we move toward a final first version of the model, especially as we are revisiting some of our initial assumptions in light of the prototype review.

Part of doing this will involve outreach to projects which are concerned with cross-domain data integration, and the groups present at the Dagstuhl workshop offer some opportunities for doing this. It also demands that we document these specific features of the standard in a way which makes it easy to communicate with other domains. Such documentation is probably unlike most of what we have produced in the past for use within our own community, but we do have good examples in hand.