DDI 3 Development at DDA

by
Jannik Jensen[1] and
Dan Kristiansen[2]

Abstract
The Danish Data Archive (DDA), a national data bank for researchers and students in Denmark and abroad, is dedicated to the acquisition, preservation and dissemination of machine-readable data created by researchers from the social science, health science and history communities. The DDA has a need to convert existing OSIRIS[3] and Data Documentation Initiative (DDI) version 2 documentation to DDI 3 and to integrate information from various metadata sources across the data life cycle. Health sciences data in particular require an enhanced metadata structure. Our efforts to make these transitions and improvements and to build useful tools have required a thorough grounding in DDI and we provide our perspectives in this paper.

Open metadata structure approach
Since there are only a few tools for DDI, we decided to DDA is a member of the Data Documentation Initiative Alliance (DDI Alliance) and has been involved in the development of the DDI 3 standard from the beginning. In general the DDA finds the DDI 3 standard a very flexible mechanism to capture the metadata structures defined by the social science and health science communities. As a conceptual model DDI 3 can provide inspiration to research communities on how to structure metadata most efficiently for reuse across the data life cycle.

As DDI 3 is an open, widely accepted standard across data archives, there is an opportunity to influence its development and to draw upon the experience of the community around it. And the community around the DDI 3 standard is reaching out to other communities. For example, DDI 3 was developed in line with other widely accepted standards, including the Statistical Data and Metadata Exchange (SDMX) standard and the Metadata Registry Standard (ISO/IEC 11179), thus facilitating metadata interoperability[4]. A proposal to define controlled vocabularies using Genericode has also drawn upon the expertise of the wider DDI community through a cross-archive working group. And work on using the Resource Description Framework (RDF) together with DDI 3 is also being undertaken.[5] All of this is leading towards more tools in the DDI tool box to apply to metadata creation and use.

Advantages for archives
The primary business model of the DDA is cleaning and enhancing the quality of data and metadata without changing the layout of the metadata and data deposited by researchers. The DDI 3 standard offers some key benefits in this process and possibilities that were not available with previous metadata models.

DDI 3 brings with it strong referential and versioning features for fine-grained metadata elements. The DDA will implement these options by reusing metadata structures across studies with the intent to focus more attention on the content of the surveys while also marking them up to fulfill long-term preservation goals. In time reuse can extend to data mining across the metadata collection.

With the OSIRIS standard not being able to document hierarchical, panel, cross-sectional, or follow-up studies, DDI 3 provides a useful alternative, especially given its focus on extended reuse and inheritance. The health sciences unit within the DDA is particularly interested in these features of DDI 3.

Regarding long-term preservation planning and storage, the DDI 3 and its community are joining forces both with technologies like FedoraCommons[6] and archiving standards such as the Metadata Encoding and Transmission Standard (METS)[7] and Preservation Metadata Implementation Strategies standard (PREMIS)[8] to incorporate additional metadata about archiving. All components come together when defining an OAIS[9] implementation of a data archive for the social science, health and history domains.

Open development
With the many standards and stakeholders in play across the current metadata landscape, no one organization can be expert in all subjects, and thus the straightforward solution is collaboration. Of course, collaboration itself is often not straightforward but more like a wave or a moving target. The same can be said for the standards themselves as they develop, mature and are constantly being enhanced with new features and as additional standards emerge; this is the joy of information technology. Over time we have seen a greater focus on information exchange and tons of exabytes of it. The strategy at DDA is to upgrade our data for this

exchange, make it available within various networks and capture its reuse and relationships.

Developing metadata upgrade and data ingest systems built upon DDI 3 technology is key in this process.[10] To execute a software project in the current environment characterized by high complexity and uncertainty, the DDA has decided upon an open source approach aligning with the community and the standard itself. This approach brings potential collaboration, knowledge exchange and product hardening informed by the feedback of others into the end product and offers wider integration possibilities with other open products in the domain, for example, the Questasy project at CentERdata in the Netherlands[11].

Even though the DDA is creating open source software, we acknowledge the closed source initiatives such as Colectica by Algenta Technologies[12] as clear assets for the DDI community. The more vendors that produce tools and applications for DDI 3 the greater the adoption is likely to be. The synergy effect of both the closed and open camps developing DDI 3 software is a real benefit for both end products and the evolution of DDI.

Open and configurable reusable solution

In 2007 the DDA took an active role in the DDI Foundation Tools Program[13], a collaborative endeavor with stakeholders from several institutions contributing time and money to developing DDI 3-based tools using an outline of common IT development tools[14]. Specifying an operating- system independent approach based on Java[15] with incorporation of various open source projects, the project resulted in DDI 3 tools licensed under an open source license[16]. The past and current IT developments in the DDI 3 field are following an updated version of these recommendations.

The first step for the DDA was to develop a common object model in Java using the Apache XMLBeans technology[17]. Around this object model the DDA has built tools for secondary validation, URN element generation, and a mechanism to extract studies contained within a grouped structure[18].

The first DDI 3 developments led to the design and developments of a centralized suite approach instead of relying on pooling separate tools into a system. The aim of the suite approach is to minimize the footprint when dealing with large amounts of XML.

Work on an editing tool for DDI 3, which is being built using the Eclipse Rich Client Platform[19] as the front end, began in the fall of 2008 with architecture design reviews provided by the Open Data Foundation.

The editing suite is primarily designed for configuration and reuse/extension. The rationale for these design decisions arose out of the need for flexibility in the implementation of metadata in DDI 3 and the need to tweak system components for customized needs [20]. This flexibility makes it possible to, for example, change the XML persistence layer or how IDs are generated for identifiable DDI 3 elements.

Conclusion

So far the DDA has built a sound basis for its DDI 3 strategy, and the major components have been identified and constructed. The work ahead is to harden and extend these components. To lead this process, the DDA has compiled a functional requirements document following the IEEE Std 830-1998 [21] . On the agenda are several topics of interest, including functionality to help create and update longitudinal metadata and functionality to facilitate the review process of a study. To enhance the reuse functionality, an internal lookup and resolution service for DDI 3 URNs is scheduled.

Internally the DDA has allocated two additional user representatives within the DDA for the project to ensure end user collaboration, interaction and acceptance of features and the user interface layout, leading developments in a more agile direction.

As DDI 3 is growing in use and becoming part of production systems, more services will evolve around it, including services not currently identified. The software approach the DDA is taking is to deliver components that are very close to the standard but not tied to a particular vendor, thus ensuring agility in implementation and reuse in planned and future systems.

Notes

1 Jannik Jensen is a software developer at the Danish Data Archive in Odense, Denmark.

2 Dan Kristiansen is a software developer at the Danish Data Archive in Odense, Denmark.

3 Anne Sofie Fink et al. (2003). “Preservation of Knowledge- Data processing in the Danish Data Archives.” *IASSIST QUARTERLY*, 2003.

4 Arofan Gregory et al. (2009). “Metadata.” RatSWD Working Paper 57, 2009.

5 Patrick Carmichael and Agostina Martinez Garcia (2009). *Semantic Technologies to Support Teaching and Learning with Cases: Challenges and Opportunities*. University of Cambridge, 2009.

6 http://www.fedora-commons.org/

7 http://www.loc.gov/standards/mets/

8 http://www.loc.gov/standards/premis/

9 http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683

10 Jannik Jensen and Dan Kristiansen, (2010). "Building a Modular DDI 3 Editor." DDI Working Paper Series, 2010, doi:10.3886/DDIUseCases02.

11 http://centerdata.nl

12 http://www.colectica.com/

13 http://tools.ddialliance.org/

14 Open Data Foundation (2008). *Guidelines for Tools Development and Recommendations for Operating Environment*, 2008.

15 http://www.java.com/

16 http://www.opensource.org/

17 http://xmlbeans.apache.org/

18 Jensen, Jannik; Pascal Heus; Joachim Wackerow; Jeremy Iverson; Dirk Roorda; Rene van Horik. "DDI and Related Tools: Next Generation Tools for Converting, Displaying and Visualising Data." Chair Wendy Thomas. Presented at the annual meeting of the International Association of Social Science Information Service and Technology (IASSIST), Palo Alto, CA, May 2008.

19 http://wiki.eclipse.org/index.php/Rich_Client_Platform

20 Jannik Jensen and Dan Kristiansen, (2010). "Building a Modular DDI 3 Editor." DDI Working Paper Series, 2010, doi:10.3886/DDIUseCases02

21 IEEE Std 830-1998 IEEE Recommended Practice for Software Requirements Specifications – Description http://standards.ieee.org/reading/ieee/std_public/description/se/830-1998_desc.html