DDI Alliance Executive Board Meeting
7 November 2019

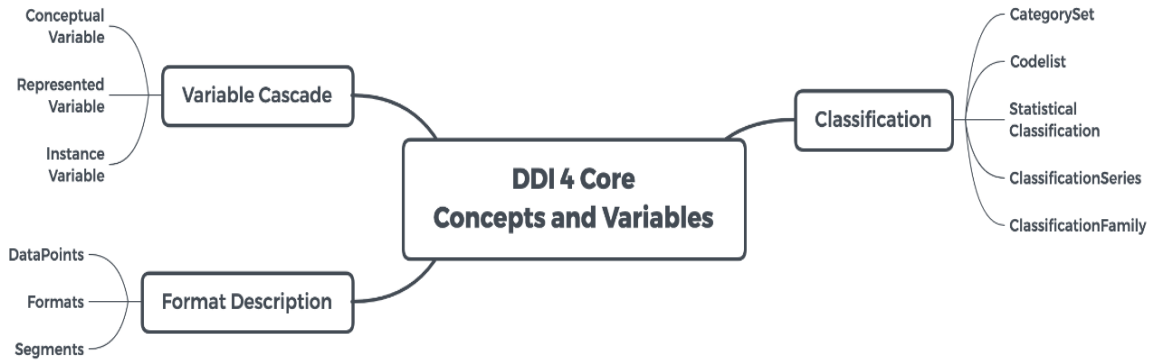
Present: Bill Block, Cathy Fitch, Jared Lyle, Steve McEachern, Barry Radler

DDI 4 Core

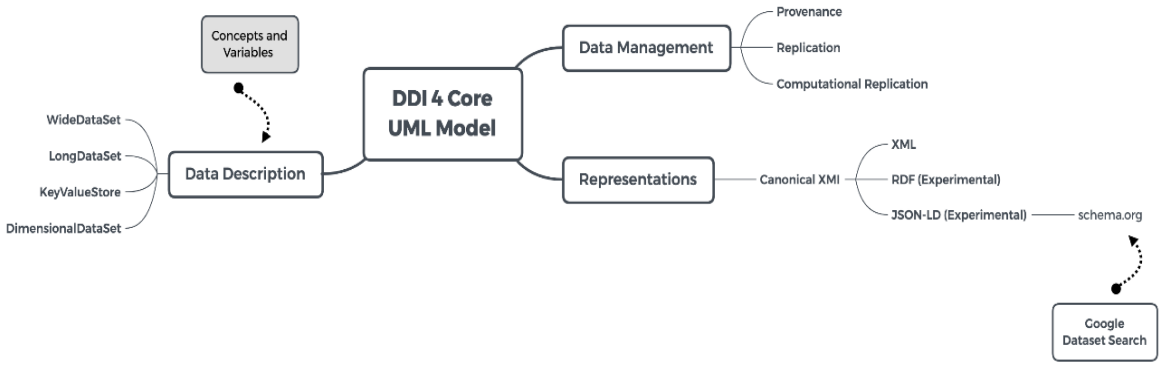
Steve presented on a way forward for DDI4, including: 1) What is the product that the DDI4 development group will actually produce? (What product are we selling?) 2) Who wants that product? (Who are we selling it to?) 3) Can/should we continue to fund the development of DDI4? (How do we pay for the development?) It was suggested to add another item: 4) When will it be ready? Steve noted that recent developments, especially since the October Dagstuhl workshops, may provide an answer.

Regarding what is the new DDI4 product, Steve referenced the document distributed by the DDI Modeling, Representation, and Testing (MRT) Working Group, "Looking Forward to DDI 4 Core" (see Attachment 1). Key messages from the document:

- The DDI4-Core product is a small, contained set of content. In the past, DDI 4 was trying to do everything. Now, the content scope is much smaller (a good thing).
- Three key content areas:
    - Concepts and Variables: Variable Cascade, Format Description, Classifications



    - UML Model: Data Description, Data Management, Representations. This means you can manage data in non-XML formats, like JSON-LD.



    - Process Model: based on the W3C PROV model, describes movements in and out. Connecting with external models means there's a lot less for DDI to have to describe. This means DDI can focus on its strength: data description.
- It does NOT replace DDI-C and DDI-L – it will complement them, and provide better capacity to provide access to DDI-enabled data via the web (particularly semantic web technologies)
- "Given the feature set, immediate implementation in some on-going projects (e.g., ALPHA Network/Alpha++, DDI 4 Core R Libraries) is expected"

Steve noted he saw a demo at Dagstuhl of DDI 4 embedded within the R package.

Steve indicated ongoing discussions with several organizations interested in contributing to development of DDI4 and other DDI work products. Dataverse, which is progressively replacing Nesstar (key tool for supporting DDI-Codebook), is interested in using DDI4 to allow use of web-based tools, including to provide better variable support. Two other projects are interested in using DDI4 and are requesting a letter of support from the Alliance.

Board members expressed enthusiasm about outside project funding, but raised concerns about expressing long-term support for a specification that has not yet been reviewed or approved. The Board expressed interest in writing the letter of support by framing it as leveraging DDI. Steve will draft the letter of support and email it to the Executive Board members for their review and approval. The Board will continue discussions about partnering with outside groups.

Board members expressed concern about establishing a third product of the Alliance. Maintaining three products will be more expensive than maintaining two. It is also unclear the integration/interoperability between these three products from the user perspective and from a tools/maintenance perspective.

The Board discussed clarifying the marketing of the separate products. There is a recognition of DDI as a brand, with DDI as a solution for documenting social science research and each of the products offering unique functionality. It was suggested to use a word rather than a number

in the DDI4 name to make it clear DDI-C, DDI-L, and DDI4-Core are separate products. It was also suggested to focus on the functionality of the product instead of the name – e.g., we're working on things that do X that support Y.

Research & Engagement

The Board members discussed the proposed DDI member feedback research project. Barry, Steve, and Ron Nakao had developed questions and topic areas that included a mix of closed- and open-ended measures that will provide the capacity to compare responses across member types while providing maximal feedback flexibility. Barry distributed the draft interview schedule in October to the Executive Board and Marketing Group and received two responses. When he distributed the information, Barry noted three points of context:

- The target audience is due-paying members. I think we're all agreed it would be ideal to include other important stakeholders (affiliates, ex-members, non-members), but we will be targeting members first
- Our budget only allows for data collection, the primary deliverables being audio-recordings of the interviews along with a methodology report. Thus, we propose creating a temporary working group tasked with transcribing (probably farmed-out) and analyzing the raw data to provide a final report. It would be great if at least one person from each of the stakeholder groups (EB, Training, TC, DDI 4, etc) represented in this effort.
- We will perform a soft launch of data collection, running a couple interviews first and having the analysis group review the data for any problems with the instrument/schedule before proceeding with production data collection on the rest of the sample.

The Board members discussed and responded to the concerns that were raised in the two responses. Joachim Wackerow, who did not attend the Executive Board meeting call, asked that his comments and questions be entered into the minutes (see Attachment 2).

Board members agreed it would be great to interview non-members, especially former members. But we're operating under constraints -- largely financial ones. It was suggested to do a soft launch (5-10) of the interviews and include in the soft launch stakeholders outside the membership. That way we can see if the interviews are working. If they're working, we expand to the entire membership. If they are not working, we can assess whether to pivot to another collection point. The soft launch might also pinpoint key questions we can ask in survey form and share more widely outside the DDI membership. We're scoping the potential for the future. We can use other opportunities, as well, like user conferences to do informal versions of the same thing.

Regarding whom to target for the interview, we could interview more than one person from an organization.

After full discussion of the concerns, all voting members on the call agreed to support proceeding with the member feedback research project. Jared will contact the survey research organization to see if they can modify the contract to allow for a limited soft launch and

evaluation before expanding interviews to all members. If they can, the Alliance will contract with them to conduct the interviews.

FY2020 Work Program

The proposed FY2020 calls and topics:

- December 16 (2pm EDT, 1pm CDT, 8pm Berlin, 6am Canberra) -- DDI work products review, including DDI 4 Core outcomes
- January 21 (2pm EDT, 1pm CDT, 8pm Berlin, 6am Canberra) -- Funding sources and revenue generation
- February 18 (2pm EDT, 1pm CDT, 8pm Berlin, 6am Canberra) -- Training + Scientific Board review
- March 16 (3pm EDT, 2pm CDT, 8pm Berlin, 6am Canberra) -- Strategic planning
- April 13 (8am EDT, 7am CDT, 2pm Berlin, 10pm Canberra) -- Research and engagement project + Budget
- May 4 (8am EDT, 7am CDT, 2pm Berlin, 10pm Canberra) -- Annual meeting preparation + Budget

Attachment 1

Looking Forward to DDI 4 Core

DDI Modeling, Representation, and Testing (MRT) Working Group[1]

November 2019, draft version 1.0

I. Overview

This document describes the anticipated use of the DDI 4 Core specification, to be released for public review early in 2020. While representing a shift in direction, it provides support for functionality which is increasingly needed within the DDI community and beyond, and which is complementary to that already provided by the existing DDI specifications and work products.

II. Background

DDI 4 Core is the result of several years' work on a next-generation DDI specification. The flagship work products of the DDI Alliance play specific roles in meeting the metadata needs for archives and data producers in the social, economic, behavioral, and health sciences, and in official statistics. DDI Codebook is an after-the-fact XML representation of a data dictionary and has proven to be both popular and durable. DDI Lifecycle is a more comprehensive XML standard, supporting the entire data lifecycle. This includes upstream metadata capture, repeat waves of data collection, description of questionnaires, and other more complex features (like support for metadata-driven design approaches) than those found in DDI Codebook.

As originally conceived, DDI 4 was seen as an even more full-featured specification, based on a conceptual model which could then be represented in a variety of syntaxes (like XSD/XML and OWL/RDF). It was planned, in effect, as a replacement for DDI Lifecycle (or the next iteration of that specification). After much effort, the Prototype Review was conducted in the latter half of 2018, leading to a number of comments.

Subsequently, it was proposed that a subset of the overall model be used as the basis for continuing work, and the MRT (Modeling, Representation, and Testing) group was formed to carry this forward. The idea was that a useful production specification be made available within a year – MRT began this process at the start of 2019, holding weekly calls and conducting two face-to-face sprints. The exact requirements for the work were a subject of discussion, and the description of data and a process model for describing provenance were selected as the focus of the work. Implicit in this scope was a model for all of the conceptual pieces of the specification (variables, codelists, categories, classifications, etc.)

III. Scope and Purpose

The work began with a discussion of what specific requirements would be met. It was quickly realized that there were two needs which had to be addressed, resulting both from the comments received during the public review, and in subsequent discussions among the participants of the group. The first was that the model needed to be made simpler and more approachable. The second was that - due to cross-domain research and initiatives becoming increasingly prevalent - it needed to address the use of non-traditional forms of data, especially those coming from new sources and from domains outside the social sciences.

[1] MRT is made up of Daniel Gillman, Jay Greenfield, Arofan Gregory (acting chair), Oliver Hopt, Larry Hoyle, Hilde Orten, Flavio Rizzolo, Wendy Thomas, and Joachim Wackerow.

It became apparent that, while the developing DDI 4 model provided some improvements over DDI Lifecycle (especially the "variable cascade" that has later been incorporated into DDI Lifecycle), it was not desirable to replace it. Instead, DDI4 presented an opportunity to reach a broader audience than the mostly survey oriented one for DDI Lifecycle. Parts of the emerging DDI4 model, such as data description, are relevant for data in general. Other parts, such as the emerging data capture section were more tied to traditional DDI functionality. DDI4 Core then came about as a way to make those more general parts of the model available to the public on a short timeframe, and in a way that was complementary to the use of the existing specifications.

DDI Lifecycle was designed to support the entire data production process, and as a result it carries with it a significant overhead, as well as making some general assumptions about how data is managed across the lifecycle. This is a necessary aspect of the function it performs, but it does present the user with some initial complexities: good data management across the lifecycle often presents users with a challenge that is not only about what XML standard they will use, but about how they manage the process of data production.

DDI 4 Core – the immediate-term product of the MRT working group – was seen as a complement to this functionality, and as a tool which could potentially be used on its own. While remaining fundamentally aligned with DDI Lifecycle, it is intended to provide support for some of the new requirements faced by the DDI community.

Modern data platforms deal with data in multiple forms. Data needs to go through a number of steps before it reaches in a consumable state: concepts, variables, codesets and relevant entities need to be identified, data has to be explored, profiled and cleaned, schemas and linkages have to be defined and then shared downstream with data consumers. DDI 4 Core can be used across all these steps, in conjunction with other standards, to enable interoperable solutions and to streamline scientific data production in multiple domains.

Traditionally, data managed with DDI has been "rectangular" data as commonly seen in statistical packages such as SPSS, Stata, SAS, and R. These are unit-record data sets with a set of variables held for each observed unit. Today, however, there are many new types of data which increasingly are used in research and statistical production: "big data" (including that sourced from social media) and event data coming from administrative registers are among these, but there are others as well (data collected by sensors, etc.). In addition, alignment with some of the models for aggregate data and time series have become important. These requirements emerged from events in which the DDI specifications were viewed as part of the larger data ecosystem, in a world where cross-domain use of data is increasing.

Data provenance has been a major topic of discussion for some time now within the broader research and statistics community, with vocabularies like the W3C's PROV (provenance) becoming a focus of attention. The DDI specification developers were quick to understand these requirements but did not have a simple solution for them.

DDI 4 Core is designed to address these needs. Given its intended function, it is a more outward-looking standard than previous versions. Because it is model-based it is easier to integrate with other standards, which are themselves not always XML-based. Because it recognizes many new forms of data, yet is still aligned with earlier versions of DDI, it presents a mechanism for integrating these various types of data with more traditional rectangular forms, and to act as a bridge to existing DDI-based data management systems.

The DDI provenance model is designed to record the flow of data management processes while integrating neatly with more detailed descriptions and other standards. It embraces the two major models used for data processes and flows at a high level: the PROV model and the Business Process Modelling and Notation (BPMN) standard. It works with any lower-level description of data processing (such as the syntaxes from statistical packages, or the work on "SDTL" being done by the C2Metadata project).

IV. Availability and Features

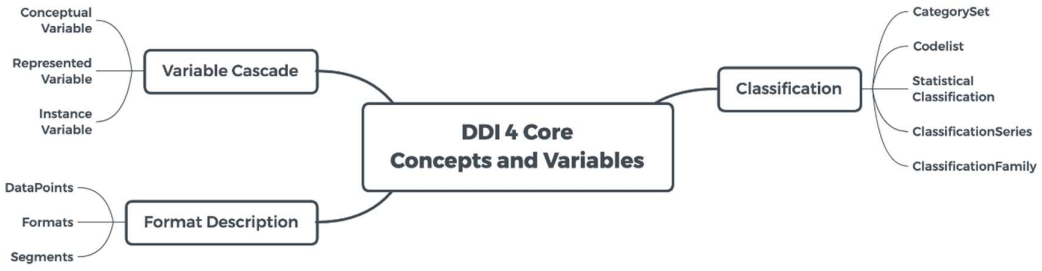
While a release date cannot be announced until after public review has been conducted, and any resulting comments addressed, it is anticipated that the initial release of DDI 4 Core will occur in the first half of 2020. Given the feature set, immediate implementation in some on-going projects (e.g., ALPHA Network/Alpha++, DDI 4 Core R Libraries) is expected. Testing of the new features of DDI 4 Core is ongoing, but the MRT group is confident that these forward-looking applications can be supported.

The relative simplicity of the model – and the flexibility provided by having a canonical form in UML – will make it a useful tool for many different types of adopters, and one which does not necessarily carry a large overhead.

A. Conceptual Metadata

The use of conceptual metadata in DDI 4 Core is largely unchanged from what was seen in the Prototype Review, covering models for many foundational metadata items (concepts, classifications, variables, etc.) as they have evolved over time within the DDI family of specifications.

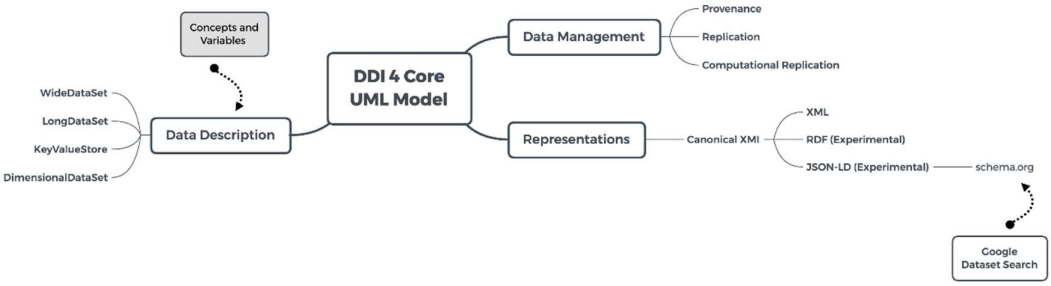
Two items are worth noting: the "variable cascade" and the model around "data points". These can be seen in the diagram below.



The variable cascade gives a sophisticated model of how variables are used in different aspects of data description: at the conceptual level in design activities (Conceptual Variable), at the logical level for the purposes of reuse (Represented Variable), and as a construct within an actual data set (Instance Variable). This is a more nuanced model than that found in many other specifications, and it better supports the many demands placed on the Variable in different scenarios.

B. Data Description

The diagram below shows how data description – the heart of the DDI 4 Core model – relates to other parts of the specification.



In describing data, the DDI 4 Core model recognizes four main data structures:

- Unit-record ("wide") data (e.g., traditional rectangular data)
- "Long" data (event data, spells, sensor flows, etc.)
- Multi-dimensional data (aggregates, indicators, time series, etc.)
- Key-value data ("big" data, data lakes, unstructured data, etc.)

Data description is implemented at the conceptual level by assigning structural and specific roles to the most atomic parts of a data set. (The term "datum-centric" is often used to describe this approach.) These atomic parts (datums, data values, table cells) are then assembled into the higher-level structures which are used by processing systems. The same data set can be seen in different ways by different systems, without the manual intervention such transformations have traditionally required.

The power of this approach is that data can be viewed as needed by the systems which process it: there is no guarantee that a given data set will conform to the requirements of a specific processing system, but the available data are understood, and the needed forms can be programmatically assembled if the requisite data are available. No other metadata specification in common use today provides this all-encompassing functionality – this is a new requirement stemming from changes in how data are used in modern research and statistical production. (An example which may be familiar to those in the DDI community is the RAIRD project in Norway, which employs many different data structures at different points as it transforms sensitive register data into a form which can be safely accessed by researchers remotely.)

DDI 4 Core has also been designed in a way that will be extensible to types of data beyond numbers and text, such as binary objects like images or sound clips. This may facilitate future extensions into describing qualitative data in a more straightforward fashion than would be possible with Codebook or Lifecycle. In addition, DDI 4 Core supports schema-on-read and other flexible approaches to describing data in data lakes and other modern data environments.

C. Process Metadata

The DDI 4 Core process model supports three main uses, which are of growing importance in research, especially as increased emphasis is placed on transparency. The first is a description of provenance at the level of the systems which manage data; the second is reproducibility (e.g., does the information exist to reproduce findings?), and the third is computational reproducibility (can the reproducibility be performed by computers?). All three uses require a similar set of information, of increasing detail.

Processes can be described both as process templates and as records of what processes have been used for specific data sets. Many data management systems today use a style of declarative process

description which sets pre- and post-conditions which are then met by the processing system, acting as a "black box." These systems often use parallel processing techniques and sophisticated modern approaches which defy easy description as traditional "linear" process flows. At the same time, many organizations continue to use more traditional systems, or use them in combination with the newer approaches. DDI 4 Core has a process model which can describe both types of processes and their hybrid forms.

The new features of DDI 4 Core have been developed in combination with testing in projects which found that the Prototype Review DDI 4 was lacking support for some types of implementations. Among these was a system which generated the provenance descriptions from a declarative ETL (extract, transform, load) platform, using DDI Codebook metadata as an input. Corresponding descriptions of data were generated, as wide ("rectangular") files used to produce event data and spell data ("long" data) in DDI 4 Core for analysis. Dissemination of aggregates and indicators resulting from the analysis was also examined.

D. Platform

It is worth considering the advantages a model-based specification brings to the table. Because the model itself is formalized using a subset of UML, expressed as Canonical XMI (a well-supported portable XML format for exchanging models between UML tools), it can immediately be used as the basis for implementation in a variety of syntaxes.

Open-source tools such as the Eclipse platform provide programmatically generated syntax representations in a wide variety of languages (Java, C++, Python, JSON, RDF, SQL systems, etc.) directly from the model. In addition, an XSD/XML syntax representation for preservation and exchange will be provided as part of the official specification, to support traditional implementation approaches. A dedicated OWL/RDF syntax representation for discovery purposes in the Web of Linked Data is also being prototyped.

Maintaining the model as a UML formalization also contributes to its sustainability. The technologies used to implement data systems are changing at an increasing rate – a flexible expression of the model provides a guarantee that the specification will not become less useful over time as a result of having the wrong set of published syntax representations. In this sense, DDI 4 Core represents an enhanced degree of "future-proofing" when compared with earlier DDI specifications, at a time when this aspect is becoming more important in terms of general technology developments.

E. Alignment with Other Standards

Given the wide range of data and functionality that systems must support today, it is understood that a description of data and process will be used in combination with many other relevant standards and models. DDI 4 Core is designed for this type of use in a fundamental way.

In the past, DDI specifications have been designed with the idea that the metadata they describe may need to be transformed into other formats, based on other models, and the specifications have been intentionally designed to support this use. That aspect of the DDI design remains also in DDI 4 Core.

What is new in the DDI 4 Core model is the direct use of many relevant models found in other standards, as the basis for DDI modeling. The fact that DDI 4 Core is expressed as a UML model allows for classes from external models to be used directly inside the DDI 4 Core model typically as a more generalized construct which is narrowed to meet specific requirements. Perhaps the single best example of this is how the PROV model from the W3C specification is integrated into the

process model in DDI 4: the basic constructs of PROV are directly implemented in the DDI 4 Core model.

Thus, integration with external standards is supported both at the level of the model and at the level of different standard formats, increasing the fashion in which systems can build on these alignments. Because DDI 4 Core is a standards-based model, using UML, this feature becomes possible.

V. Summary and Future

DDI 4 Core is a different type of specification from DDI Codebook and DDI Lifecycle – it is a new and complementary tool that many organizations will find useful to address some of the newer demands which they face in terms of data management. It is not a replacement for earlier versions of DDI, but can act as a supplement to them, being used as the basis of systems which extend functionality to cover non-rectangular forms of data, and which implement data provenance at the level of the management platform.

DDI 4 Core is designed intentionally to be aligned with earlier versions of DDI and to integrate with external metadata standards. Modern implementations must cover different aspects of data discovery, processing, and dissemination, and DDI 4 Core provides a basis for integrating with many of the standards and models in common use today.

Recent discussions at the “Interoperability of Metadata Standards in Cross-Domain Science, Health, and Social Science Applications II” workshop at Schloss Dagstuhl in October 2019 made it clear that DDI 4 Core was of interest as an integration model for cross-domain and inter-disciplinary use. In addition to fulfilling emerging needs of existing DDI user institutions it is expected that it will find implementers in some domains and applications which have traditionally not seen DDI as relevant. These discussions also showed that some additional features might be desirable for complete integration with specifications such as Schema.org to expose data holdings through the major search engines.

Between a public review in early 2020, and continued discussions with the participants at the Dagstuhl workshop, it is anticipated that revisions to the DDI 4 Core model will be forthcoming. The current version of the model is, however, sufficient for immediate use in some implementations.

Ultimately, it is hoped that DDI 4 Core will provide a way forward for those both within and outside of the DDI community in addressing the new demands placed on their systems by new types of data and an increasing interest in data provenance. For users of DDI Codebook and/or DDI Lifecycle, it represents not a replacement, but rather an extension: it is an additional tool to be added to the already-powerful DDI suite of work products.

Attachment 2

---------- Email message ---------
From: Wackerow, Joachim <Joachim.Wackerow@gesis.org>
Date: Fri, Oct 18, 9:32 AM

Dear Executive Board Colleagues and Marketing Group,

I have several comments and questions on the planned member survey.

The purpose of the member survey seems to be providing more information for decisions on the direction of the strategic plan. But is this really the case? I see here following questions:

1. The members expected in the past leadership of the steering/executive board. Some members expressed in the past specific interests but most members were happy to get some guidance from the steering/executive board. Will the members really provide strong information for a strategic plan for the next 2-5 years?
2. I think asking only members results in a limited view on the DDI community. The focus seems to be even only on due-paying members (according to the website 27). Is the purpose of the DDI Alliance to serve mainly the interest of the members like in any sports club? I would doubt this. The charter says something different:
   "The Data Documentation Initiative (hereinafter "DDI") Alliance shares a commitment to meet worldwide requirements for publicly available standards and semantic products supporting the documentation and integration of social science data and other data for understanding the human condition."
3. The DDI community is much larger than the member organizations. The user conferences EDDI and NADDI show this impressively for many years. Why not getting input from the larger community?
4. A strategic plan should not only focus on the past and present. It should be a plan for the future and it should reflect current ongoing developments in the research data community at large. Why not getting input from possible users and possible partners beyond the DDI community?
5. I think the member survey results in something we already know: some members use DDI Codebook (i.e. archives), some DDI Lifecycle (i.e. institutions which deal with complicated studies like panels, describe questionnaires), some will use DDI 4 (official statistics, organizations with new data types like digital behavioural data). I don't expect any strong group in one of these interests which would make clear which directions mainly to follow. What would this mean for a future strategic plan? Does this really help?
6. In summary, I think that a market research on existing and possible users, input from possible partners would result in a much wider insight. Could this give some better information how the membership and/or partnerships can be increased?
7. Would it be not better to have an external organization to make any survey? Is it really a good idea that members of DDI Alliance and Executive Board are leading this? I could see a possible conflict of interest in a situation where people have strong opinions. The review of the DDI Alliance in 2011 was done by an external company. Even the methodology of the review was developed by this company.

8. Assuming that this member survey makes sense (from which I'm not convinced), I'm not sure that the proposed qualitative interview is the right method. The proposals say that they can collect the data for the given amount but they cannot prepare, analyze, and report the results. Having the data is only half-way. But we shouldn't spend more money on this. It is already 30% of the annual revenue. Could not another approach be taken which gives a meaningful result with the given budget amount?
9. Who will be really asked from the members? The "Representative to the DDI Alliance" or the "Representative to the Scientific Board"? The first one is often a person from the leadership, the second one is often on the working level really dealing with metadata. There are often different understandings on the use of DDI in one organization. Will it not be difficult to interpret the answers?

Maybe it would be best to ask the company of the 2011 review (Breckenhill) what they could do for the given amount (30000 USD) in terms of a market research? The contact was Ned Eustace.

Achim