[→ Knutholmen Master](https://docs.google.com/document/d/1-riSpsOHvxzZM_M4bb_8tFwkCVCKIecq7R5h6OyV3gU/edit)

Case for Developing
Full DDI Thesaurus

By Dan Gillman

This document contains the case for developing a full thesaurus of terms and concepts used in DDI-4. This thesaurus should contain definitions of terms. Currently, there is a glossary, but the terms defined are either general (e.g., abstract class) or about the development process (e.g., Drupal). Terms arising from the model itself are not contained, though the classes themselves are defined in Drupal. During discussions around using the new Codebook view, the need for definitions and explanations for those terms was uncovered. Some terms are used in class names, and it is confusing if they are used in restricted ways from their usual definitions or if the definitions are too technical.

First, no technical specification should rely just on dictionary or common definitions of terms to convey the meaning of constructs used in that specification. The need to make sure everyone has the same understanding is too important. Interoperability depends on this. The class Datum in the model is a case in point. The dictionary definition of datum (the term) neither conveys the true meaning of datum nor its usage in the model as a class. The technical definition adopted by DDI, which is not similar to the dictionary definition, is not digestible by general users. It Is too technical and needs to be explained. Further, the usage of the class in the model doesn’t quite correspond to the definition.

This situation is solvable through the development and use of a detailed thesaurus of terms and concepts, including definitions of concepts. Although, no documentation system is perfect, careful development of a DDI thesaurus can eliminate or reduce many problems. Below is a proposal for how the thesaurus should be written. Implementing the basic standard thesaurus structure and a few extensions makes the most sense as a first try. At this point, building an ontology, which would require the development of semantics for every relationship used, is just too much work.

We recommend, to start, the thesaurus is implemented as linked HTML pages. Possible additional development includes using XKOS, RDF, and other thesaurus or terminology standards.

Each specific term should be documented in the following way:

* Designation (the term being defined)
* Definition (definition of the term, not necessarily how it is or might be used in the model)
* Explanation (examples, restrictions, and additional information)
* Usage (usage of the term, especially how a class by the same name might be restricted)
* Relationship types
	+ Broader than
	+ Narrower than
	+ Related to
	+ Synonyms
	+ Homonyms (based on spelling as well as pronunciation)
	+ Antonyms
	+ Class name (names classes in the model that use the term in their names)
* References

This template loosely corresponds to the requirements of a thesaurus. There are differences here, but they are probably unimportant.

Relationships may be implemented in a simple manner, as they can be hyperlinks associated with the related terms and organized into the groups (relationship types) named above. For example, category is a specialization of (narrower than) concept. So, within the page devoted to defining concept, under the Narrower Than relationship type, the hyperlinked term category is listed.

As noted above, some terms are used in a specialized manner in the model, and they appear in class names. The class Datum is closely linked to the idea of a datum. Therefore, it is important to define datum as a term. It is required to understand the class. However, rather than defining the class separately, it is useful to explain how the class is used in a restricted manner. This will be done in the Explanation, Usage, and Class name fields.

The thesaurus (or whatever it ends up being called) should be attached as a resource to the DDI Moving Forward documents page. However, there is a more important way for it to be used, and this is to link terms used throughout DDI to the specific page where they are defined in the thesaurus. This way, users have little work to perform to find the meaning of terms. How those terms fit into the larger DDI picture are contained in the relationships associated with each term. As an implementation detail suggestion, linking to a term page should spawn a new tab in the browser rather than taking the user off their current page.

This effort will take some time. DDI-4 model is already large, and there are many classes and other terms to manage. It may be necessary to convene a team to execute the work, as the volume of work is most likely too large for one person to manage. There are already similar efforts underway in the UNECE and SDMX. They will be linked, and every attempt should be made to link the DDI thesaurus with these efforts as well. This coordination will add to the burden. But, every effort should be made to make the thesaurus available as soon as possible. Even the DDI-4 review process will be improved with the thesaurus in place.

The thesaurus development team should probably not have the authority to make changes to definitions of concepts. They need to work with the technical development groups responsible for the affected areas of the model. Changing definitions can have large effects. Arbitrary changes can be disruptive if they are not coordinated.

Given that the thesaurus will be large and its implementation complex, there will be the need to maintain and improve (both as an implementation and from the quality perspective) the thesaurus. This will be an ongoing effort, and convening a permanent team might be necessary. In any case, once the thesaurus is built, some of the people who are involved will have to be recruited to help maintain it. Both content experts and implementers will need to be involved.

Questions about this proposal can be sent to

Dan Gillman
US Bureau of Labor Statistics
Gillman.Daniel@BLS.Gov
+1.202.691.7523