Analysis: Annotation, Access, and Provenance

# Overview

This document considers the issues raised by the Prototype Review and by other discussions during Sprint and before regarding how annotation, access, and provenance information should be modelled in the DDI 4 Core. This is a cross-cutting issue, affecting (and being affected by) decisions on both the style and content of the model itself. This document provides an initial analysis of the topic intended to serve as the basis for further discussion within the MRT group.

# Ennumeration of Problems and Questions

Below is a listing of issues and questions which will need to be framed up for discussion and addressed by the group.

1. Annotation has more than is needed for a clear set of citation information (Title, Subtitle, Creator, Contributor, Publisher, Publication Date, Version, Identifier, Copyright, Language) – what happens to other content?
2. Annotation as a structured data type adds a layer to the model . (Is there a reason for bundling this information?) This may be annoying and unnecessary.
3. Association uses a structured data type that contain associations – Agent Association that provides the option of entering a name in-line, a role, and/or a relationship to an Agent. This may clash with the emerging MRT modeling guidelines.
4. Inheritance chains are too long: Identifiable – AnnotatedIdentifiable – Class. This is an issue for users. Reducing the length of inheritance chains is a stated goal of the modeling work in MRT.
5. Access information is separate from Annotation because:
   1. Rules vary by source (producer, distributor, user).
   2. One set of rules may apply to a range of objects.
   3. Access rules are not currently available at all levels needed (publication level, Study, SampleFrame, CatalogItem).
   4. Other classes may need this information (ex. Question, Variable of all types, Classification, etc.)

The attachment and organization of access information needs to be reviewed.

1. Explicit provenance information is provided by tracking links such as basedOnObject, sourceCapture, generatedBy, and the process model plus a textual statement of provenance. There is no means to cite a source or clarify how to interpret/navigate the appropriate provenance chain for an object.

# Goals/Needed Resolutions

This section lays out the goals of re-organizing this information, to handle the issues given above.

* Deal with the more general problem of the data type containing an association – this impacts the modeling guidelines being developed
* Implement a flattening of model structures to the extent possible (i.e. bundling only where it is required for understanding)
* Reduce the length of inheritance chains in line with the modeling guidelines
* Provide annotation and access information where it is needed but not more broadly
* Clarify the provenance chain at all appropriate levels from the datum on up

# Possible Solutions/Actions for Consideration

Some suggestions for realizing the goals given above are listed here.

* Remove the packaging of Annotation and just include properties and relationships directly, in order to flatten the model.
* Resolve conflicts between closely conceptual items such as name and title, which serve different functional roles in processing data and metadata. Clarify how and why these exist if different.
* Ensure consistency in how this content is modeled/used, and minimize duplication (in navigating/interpreting the provenance chain) where possible.