fhiso

CFPS 7
(Call for Papers Submission number 7)

Multi-Granularity Standards

Submitted by: Tychonievich, Luther

Created: 2013-03-17

URL: Most recent version: http://fhiso.org/files/cfp/cfps7.pdf
This version: http://fhiso.org/files/cfp/cfps7_v1-0.pdf

Description: Tools must be able to import course-grained information because existing usage patterns presume incompatible data granularity.

Keywords: Data model, requirement, conflict, multiple standards, import, export, granularity, lattice

The Case for Multi-Granularity Standards

Luther A. Tychonievich

15 March 2013

Abstract: *The data models of existing tools form a lattice. The bottom of this lattice stores just research conclusions. The top stores every research decision. Information cannot readily flow up the lattice, and different points in the lattice suggest different research styles. Hence no single data model can fully support all existing tools and use cases, so tools must be willing to import course-grained information.*

There are two ways in which one data model may be "greater than" or a "superset of" another. The first way is when one model supports a feature another does not, such as one tool including the profession of a person while another supports only their vital statistics. The second way is when one model stores more detailed information than another, such as one tool storing the evidence linking details with sources while another stores only the details and sources without connecting the two.

The first kind of superset relationship is the common partial-implementation situation that can generally be handled by some protocol for ignoring absent nodes. It is the second kind of superset I wish to discuss in this paper. For lack of a better term, I will call it the **granularity** of a tool.

My intent is to demonstrate that the varying granularity of existing tools support distinct user work flows. If end users are not to change their behaviour, the fact that granularities form a lattice will prevent any single data standard from providing full tool interoperability. Hence, we either need (1) to change end user behaviour; (2) multiple data standards; or (3) a standardised way for tools to accept course-grained information from other tools.

1 Data Granularity forms a Lattice

Consider four hypothetical data models.

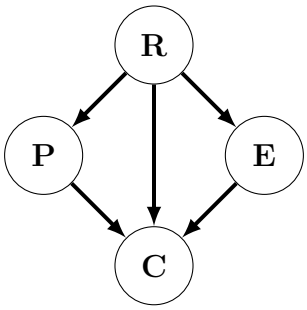
C stores conclusions. Each person as a single object with a set of attributes, relationships, and citations.

P stores "persona"s. Each person is a tree of people nodes. Leaf nodes represent the information from a single source, with attributes and relationships and a single citation. Intermediate nodes represent the decision that two other nodes refer to the same person.

E stores evidence. Each person as a single object with a set of attributes and relationships. Each attribute or relationship is the roots of an evidence tree which may be a single citation for direct evidence or a more involved derivation for indirect or conflicting evidences.

R stores research. Each and every decision of the researcher is its own node, including extraction decisions, evidence trees, matching decisions, conflict granularitys, etc.

Observe that the information in **C** is courser granularity (can be derived from) the other three and that **R** is finer granularity (can be reduced to) than the other three, but that **P** and **E** each contain some information that is not present in the other. If we draw the flow of information we get:



This structure is an example of a lattice: a directed acyclic graph with a top and bottom node or, equivalently, a partial order containing a supremum and infimum. We might easily populate it with other data models, but all will be able to export to **C** and all can be exported to from **R**. The lattice structure will remain.

2 Tool Granularity Impacts End-User Experience

Users will, of necessity, create the information their tool of choice requires. But they might easily not even create the information their tool does not store. For example, when working with a conclusions-only tool it is quite possible that a researcher never bothers to extract the information in a new source that agrees with information the researcher already knows. If the researcher switched to a persona-model tool then the researcher's process would of necessity become more involved, needing to extract all the information in each source to create a valid leaf persona.

In general, to move a user to a finer-grained tool requires them to adjust their work flow to perform more work than they previously had been accustomed to doing. To move a user to a courser-grained tool requires them to accept that some of their existing work flow will no longer be recorded in the tool itself.

Note that the difference in user experience also indicates that no amount of research logging and clever reasoning can move a course-grained tool to a finer-grained data. The user may never have even performed the work a finer-grained tool wants to record.

3 Standards and Granularity

Because data granularity impacts end-user experience, any data standardisation effort must choose between several options:

1. Multiple standards, each catering individually to one level of granularity.
2. A single standard at the bottom of the lattice; exports will be lossy.
3. A single standard at the top of the lattice; user experience will become more involved.
4. A multi-granularity standard that can support the top of the lattice but allows arbitrary portions to be stored in course-grained form instead. Exports will be perfect, but imports will be both messy and complicated.

The first two options seem unsatisfactory; the first is hardly standardised and the second doesn't meet the basic need of storing what people are already doing. The third option appears to step beyond the scope of a data standardisation process. Thus, the fourth option appears to be the most desirable one left.

In order to make a multi-resolution standard effective it is necessary to define what a tool should do when it imports material at a different granularity than it is used to supporting. Importing finer-grained information will likely be come naturally from the definition of the various levels of granularity. Importing courser-grained information is less obvious, and might have to be handled on a case-by-case basis. For example, **P** might need to treat **C** (or its user) as the source for a huge leaf persona when it imports one of **C**s person records since teasing out an accurate leaf persona for each citation is not possible. Similarly, **E** might need to treat **C** as the source providing direct evidence of each attribute and relationship when it imports from **C**.

Most discussions of genealogical data I have heard in the past have touted one particular point on the granularity lattice as superior to all others. These claims may or may not be true, but no single point will satisfy all (or even most) existing users. Hence, we need to address how to communicate up the lattice if we are to be successful.