CFPS 63
(Call for Papers Submission number 63)

Proposal for Handling Sources and Citations

Submitted by:	Proctor, Tony
Created:	2013-04-22
URL:	Most recent version: http://fhiso.org/files/cfp/cfps63.pdf This version: http://fhiso.org/files/cfp/cfps63_v1-0.pdf
Description:	Proposal for handling sources and citations
Keywords:	Sources, Citations

Contents

Abstract

This proposal suggests a generic approach to handling sources and citations in the Data Model. This general topic is large and the proposal tries to separate-away parts that are related to processing by software in an attempt to keep the Data Model focused on what needs to be stored.

1. Analysis

When building conclusions from evidence, we must cite our items of evidence, and the sources they were obtained from. However, unambiguous and zero-loss exchange of such information is currently not possible. It is a difficult subject to approach because it is not confined to genealogy and there are many stakeholders.

Let us first consider some of the distinct parts of this support:

1.1 Presentational style

There are several citation styles in common use. For instance, in the humanities there are: Modern Language Association (MLA), Harvard referencing, Modern Humanities Research Association (MHRA), and the Chicago Manual of Style (CMOS). There are other styles commonly used in law or the sciences too. The Board for Certification of Genealogists (BCG) recommends the CMOS which utilises footnotes, endnotes, and bibliographies. Each style defines the syntax for citation of a particular source-type, although there are cultural variations in areas such as punctuation and date formatting.

Although the presentational style is what we, as human being, want to see, it is not an adequate format for data exchange and processing. In addition to the range of potential styles, and the locale-dependent variations of those styles, it is impractical for software to decompose it and work with it (as opposed to treating it as some hard-coded amorphous string).

A digested form of a citation can be used to generate any style configured by the end-user, and honouring their locale in the process. This could even include generating an OpenURL string if necessary. In contrast, a printed, final-form citation cannot be reverse-engineered. The essence of a digested form is a source-type identifier and a set of citation-element values.

1.2 Citation Elements

When citing the source of a particular piece of evidence, these are the essential elements of data that it conveys. For instance, for a book reference it might include the title, author, edition, pages, and publisher. There could be additional optional elements such as an ISBN number.

The representation of citation-element values in a model must obey the principle of locale independence, as described in CFPS 16. This includes the fact that element names are inappropriate as field labels in a form, or in any other part of the UI, since they belong to the programming locale rather than the user's locale.

Some elements may have optional values, or multiple values as in a set of page references. The supported data-types share a lot of requirements with the Custom Properties proposed in CFPS 36.

1.3 Citation Template

Using the language of BetterGEDCOM, a citation template is a component that takes a set of citation elements and generates a final-form printed citation, including required punctuation, for a given source type. (http://bettergedcom.wikispaces.com/Pending+Definitions).

There are products and tools already in this field, such as Citation Style Language (CSL) templates, and the Zotero Style Repository. If the Data Model represents the essence of a digested citation using a source-type reference and a set of citation-element values then it is dislocated from the concept of citation templates. The software product reporting off the data can apply a citation-template tool of its choosing.

1.4 Source Types

Genealogy, including family history, has a multitude of possible sources. There are recommendations for the printed style and the required citations elements for many of these in works such as *Evidence Explained* by E. S. Mills. However, it is not possible to prescribe for all possible sources, especially when considering worldwide research. Any attempt to create a finite list will force vendors and users to step outside of a standard.

Requiring a central registration scheme would be too onerous and tempt vendors and users to step outside the standard by improvising some sort of private source-type scheme. In effect, a standard requires an open scheme that avoids name clashes and yet still enables transportability between compliant products.

1.5 Data Entry

The input of a citation reference into a software product is primarily the responsibility of that product. For a given source type, the product needs to know the required citation elements, their locale-specific descriptions, whether they're mandatory or optional, their data-types, and whether they're multi-valued. However, this can be provided by a definition of the respective source type. The software will have its own UI style and so needs no more than this (i.e. no forms, templates, code, etc).

1.6 Semantic Tagging

A number of semantic types are defined by the Dublin Core vocabulary, e.g. 'DC.Title', 'DC.Publisher.CorporateName.Address'. There is a requirement to associate these with citation elements for the purposes of discovery. However, these should be a property of each element. No significance should be attached to the name of the citation element itself. Doing so would limit the use of genealogical semantics in the name, possibly restrict the ability to differentiate similar elements, and would be tied too strongly to an evolving independent standard.

2. Proposal

The essence of this proposal is to use a URI as the source-type. The advantages of this are that they can be generated from a domain owned by the respective person or organisation (including FHISO) without preregistration, they have visible semantics (as opposed, say, to a UUID base), and they can be extended and versioned as required.

In principle, the individual citation elements could be address as Fragment Identifiers relative to the source-type URI, e.g. http://fhiso.org/source-type/book/v1-1/#Author. However, this is an implementation detail.

The definition of a source-type would be accessed via this this URI and it would yield details of the respective citation elements, as described above under Citation Elements. A software product would likely have a local repository of source types, especially if ones have been imported along with other data contributions, or the end-user has crafted ones themselves. For more globally recognised source types then an Internet-based resource, such as a Source-type Authority, could provide resources for discovery by a software product. FHISO, itself, could provide such a discovery resource.

Information about a source type that will be displayed to an end-user, such as a title, description, and labels/descriptions for the individual citation elements (e.g. for soliciting values from the end-user), must be one level removed from this core definition of the source type. This is because it will be locale dependent. In other words, if English names and descriptions were lumped in with the core definition then it would be an awful faux pas from a design point of view.

When a final-form citation reference is to be generated then as well as the source-type URI and the end-user's locale, the software product also needs to select a presentation style (e.g. CMOS) and a citation mode (e.g. footnote, endnote, parenthetical).

3. Not Covered or Not Required

This proposal does not suggest how to deal with citation chains. For instance, citing an article separately from the journal of collected works in which it was published. This could be extended to cite the location or resource from which it was obtained, e.g., a specific archive, or an online provider.

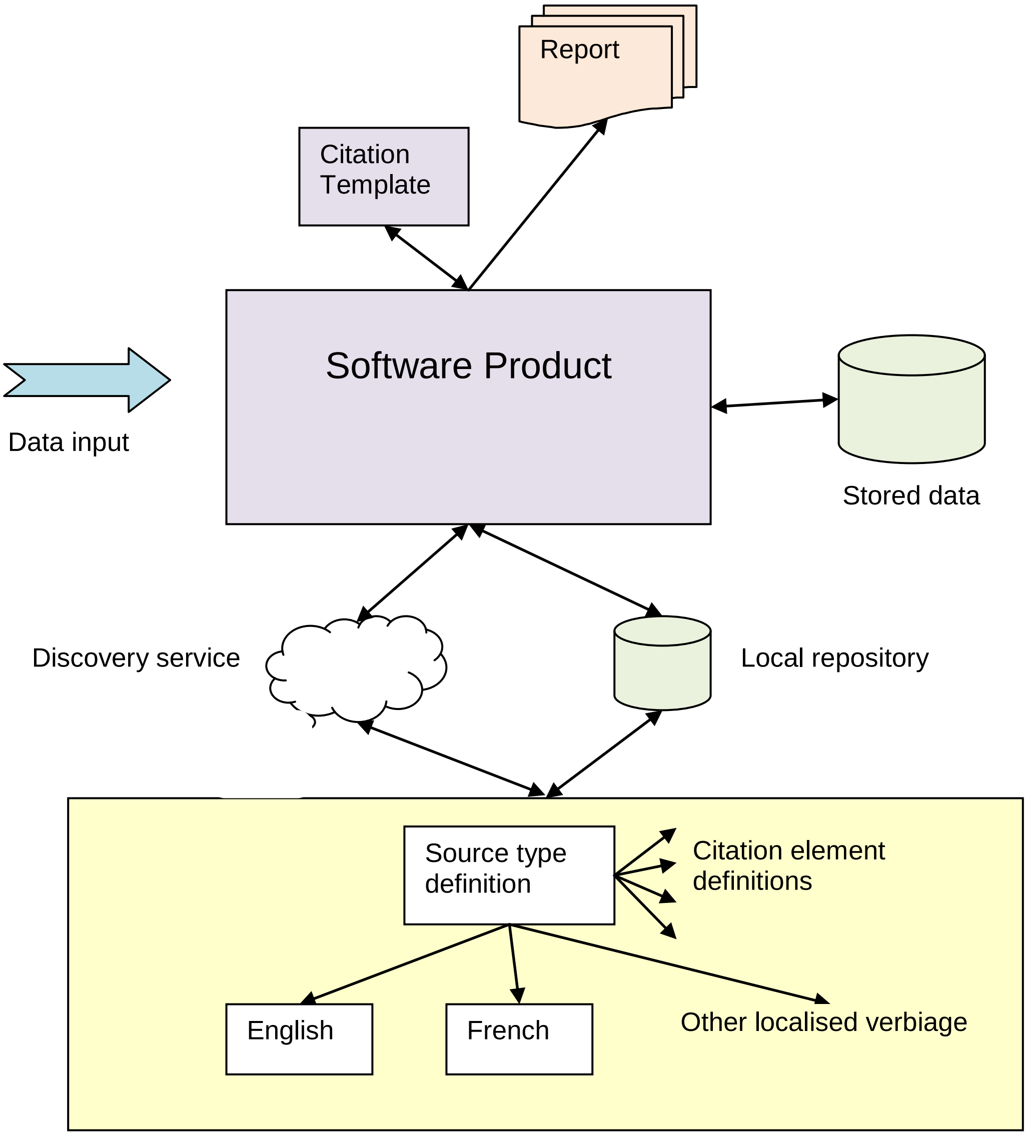
This proposal does not suggest how to cope with complex citations such as those involving author annotation and multiple, inline, simple citation references. STEMMA achieves this by allowing simple citation references to be deposited inside a normal narrative reference note.

4. Illustration

The following diagram illustrates how the main components in this scheme operate together. The source-type URI is used to fetch the definition of a source type, whether through a discovery service or from a local repository. That definition will also include the verbiage appropriate to various locales.

Data-input for a citation reference is solicited using the appropriate locale-specific verbiage for the individual citation elements, acknowledging their defined data-types and other properties in the process.

When generating a final-form citation reference, the software product must interface to some citation-template tool. If this has its own programming interface then it will mean packaging the citation-element values appropriately before calling on that tool. The weakest part of this proposal is how the citation template knows what the final-form should look like for the source type described by the given URI. It could be described using something like CSL (Citation Style language) and held along with the source-type definition. Note that this would mean the source-type definition must then provide a final-form for each supported presentational style.



5. Use Cases

The crux of this proposal is that the digested form of a citation is the fundamental representation. This is locale-independent and neutral with respect to citation presentational style. It is therefore the means we need for zero-loss and unambiguous data exchange. All presentational styles can be generated from this digested form.

However, this is at a cost. Whereas most proposals contributing to a Data Model standard can be viewed as something that need only be considered during data exchange, this proposal is more far-reaching. As it says above, it is not possible to reverse-engineer a printed, final-form citation reference. For the proposal to be effective, that digested form must be the form stored by compliant products.

6. References

FHISO cfps 16, Locale Independence. http://fhiso.org/files/cfp/cfps16.pdf.

FHISO cfps 36, Custom Properties. http://fhiso.org/files/cfp/cfps36.pdf.

FHISO cfps 20. Partially Controlled Vocabulary. http://fhiso.org/files/cfp/cfps20.pdf.

Elizabeth Shown Mills, *Evidence Explained: Citing History Sources from Artifacts to Cyberspace* (Baltimore: Genealogical Publishing Co., 2007)—or the earlier abridged edition, *Evidence! Citation & Analysis for the Family Historian* (1997) together with its companion *QuickSheet: Citing Online Historical Resources Evidence! Style* (rev. 2007).

Dublin Core Metadata Initiative. http://en.wikipedia.org/wiki/Dublin_Core.

Citation style Language (CSL). http://en.wikipedia.org/wiki/Citation_Style_Language.

OpenURL format, ANSI/NISO standard, Z39.88-2004. http://en.wikipedia.org/wiki/OpenURL.

STEMMA discussion of citations. http://www.familyhistorydata.parallaxview.co/research-notes/importance-narrative, section 6, Citations.