Citations: Goals and Considerations

29 November 2021

Editor:
Luther Tychonievich

> *Note* — This entire document is a non-normative summary of exploratory work performed by FHISO. This document is not endorsed by the FHISO membership, and may be updated, replaced or obsoleted by other documents at any time.
>
> Comments on this draft should be directed to the tsc-public@fhiso.org mailing list.

One of the most common requests made of FHISO is to create a citation standard. Conversations with interested parties on this subject show that citations are often surprisingly complex: in particular, even people who have invested effort in considering citations often have incompatible visions motivated in part by different objectives.

This document's goal is to outline the breadth of the problem space, as we currently understand it, and to give a summary of possible solution approaches that have been proposed by various parties.

The concepts documented herein come from many sources, mostly unpublished, though some of the earliest conversations are documented on the sources-citations mailing list archives. Some sources, particularly those where some party explained why their effort to develop a system for handling citations failed, were shared with FHISO on condition of anonymity. Other concepts emerged as common threads in conversation with dozens of parties and cannot be clearly attributed to any one of them.

Citations

Purposes of Citations

Citations are included in documents in many fields, including family history, for multiple reasons. While rarely explicitly enumerated by citation users, we have identified the following purposes:

1. Help others locate the source. These "others" may be the citation creator's future self.
2. Acknowledge assistance rendered by others. These "others" may be individuals or organisations. Reputation and prestige are part of the currency of every research field.
3. Provide indicators of relevance and reliability. What indicates these varies by field. Academic citations indicate relevance and reliability by publication date, venue, and publisher. Family history citations indicate relevance and reliability by the provenance chain from content creation to researcher acquisition.
4. Conform to an expected style guide. Those who read many citations can understand them more easily if they are all presented the same way. Presentation style can also help reduce ambiguity and communicate target audience.

Not all users of citations care about all of these purposes, but enough care about each that a citation standard that does not serve all four purposes is unlikely to be widely accepted.

The difference in how family history and academic citations indicate relevance and reliability is one way of viewing why academic citation standards like CSL are inadequate for family history citation needs.

Citations with Provenance

A full citation in family history generally contains some provenance information. Typically, this is modelled by describing the source the research consulted, and the source of that source, and the source of that source, and so on back to the original creator of the information. Each of the individual steps in this chain is traditionally called a "layer" of the citation.

Layering occurs in two broad kinds of ways. One kind of layering is derivation: a source can be a copy, translation, transcript, abridgement, etc. of another source. The other kind of layering is transition of stewardship: sources can be held by stewards who obtained the source from other stewards. Not all family history citation styles express both kinds of layering, but enough do that both should be modelled.

Most forms of layering are linear in structure: each source derives from one other source and was in the possession of one steward at a time. However, some derivations are compilations with multiple sources; examples include a book of obituaries from multiple newspapers, a translation informed by multiple versions of a manuscript, or a conclusion informed by multiple pieces of evidence. Stewardship can also be recorded in a way that makes order ambiguous or that combines several repositories into one without tracking which source came from which repository. Thus, a citation model should be able to support branching provenance structure.

Single-Layer Citations

A single-layer citation without provenance information typically is presented as several facts about the cited source. Although often described as a set of independent properties of the source, these do contain internal structure.

Sources tend to be items that are parts of larger items. A row is part of a table which is part of a schedule which is part of a census. A section is part of a chapter which is part of a book. A web page is part of a web site which is part of the Internet. An article is part of an issue which is part of a volume which is part of a journal. And so on.

> *Note* — In many academic styles, the item containment relationship is expressed only by formatting, with a separate formatting rule for each common hierarchy of items.

Each item is generally described by a few properties of that item. Many of these properties are textual (such as the title of a book), but some are lists (such as the authors of an article) or other structured data (dates, addresses, numbers).

Some properties more naturally belong to the relationship between items than to any specific item itself. For example, a page of a PDF document may have both an item property for the page number printed on the page and a relationship property for the ordinal position of the page within the PDF's sequence of pages.

Some properties can be posed either as properties of an item or as properties of a relationship to an omitted second item. URLs are a common example: locators within the vast digital repository known as the Internet, but often given without referencing the Internet at all.

Item relationships can branch and re-merge; for example, a paragraph might both be in a section in a chapter in a book and on a page in a printing of a volume of that book.

Data Models

Purposes of Citation Data Models

When data is present inside a computer it is stored in some type of data model. The design of that data model impacts what the computer can do with the data. Five broad categories of actions have been discussed for computers to perform with citations:

1. **Display**: Some operations on citations require a formatted presentation that can be displayed to the user and printed out.
2. **Process**: Some operations on citations require knowledge of the underlying meaning of the citations: notably, searching, sorting, validating, and de-duplicating.
3. **Generate**: Some operations on citations require the ability to generate formatted presentation from the underlying meaning: notably, report generation, changing style guides, and merging citation sets presented in different styles.
4. **Shorten**: Some style guides recommend that a citation have as many as four different levels of verbosity:
   - A full citation with provenance information.
   - A full citation of just the most important layer.
   - An abbreviated citation of just the most important items and properties of the most important layer.
   - A citation reference with just enough information to locate the citation in a related works section.

   Not all styles use all four of these, but enough use each that tools to generate briefer citations from fuller data have been requested.
5. **Customise**: Some operations on citations require the ability to let users specify parts of the presentation without modifying the underlying meaning: notably, when users have stylistic opinions that have not be fully captured by *generation* algorithms.

In this document, I refer to the formatted presentation of a citation as "citation text" or simply "text" and the underlying meaning of a citation as "citation data" or simply "data".

Text-only

Text-only citations are stored with the minimal information needed for display. If shorter versions are needed, those are separately created by the user.

Data Model	Display	Process	Generate	Provenance	Shorten	Customise
Text	Yes	No	No	Yes	by user	Yes

Text-only citations (or something very like them) are used by GECDOM 5.5.1, GEDCOM-X, and FamilySearch GEDCOM 7.0. *Text-only* citations are the format provided by most family history-oriented online repositories and archives and documented in most family history citation style guides.

Property Sets and Templates

Property set data is created by enumerating all expected properties of all expected items and relationships between items, giving each a unique name; the citation data for any given citation is a subset of those named properties with their associated values. Display is supported by processing that data with a presentation **template**. Simple *templates* list the order and styling to use in concatenating the properties of data; more advanced *templates* contain conditional reasoning where the value of one property can influence the presentation of other properties.

Data Model	Display	Process	Generate	Provenance	Shorten	Customise
PS&T	Yes	Yes	Yes	No	Yes	No

Property sets and *templates* are used by BibTeX, CiteProc, CSL, EndNote, Zotero, and other academic document preparation toolchains. *Citation data* is provided in this format by most academic-oriented online repositories and archives. *Templates* in this format are offered by most academic publication venues.

Property set design requires a trade-off between complexity and expressiveness. Academic *property set* citation standards typically have between 25 and 100 defined property names, with some sort of catch-all “miscellaneous” name to handle citations with properties outside that set. FHISO representatives have spoken with teams who attempted extending this model to family history citations who reported requiring hundreds more names to cover source types like rows in a tabular census and inscriptions on a grave marker. When they attempted to add in names for common provenance relationships, the number of names quickly passed a thousand and the projects were abandoned as leading to an unusable end.

Text with Semantic Markup

Text citations with **semantic markup** are much like *text-only* citations, but also have embedded markers to indicate to the computer what parts of the text indicate the authors, title, and so on.

Data Model	Display	Process	Generate	Provenance	Shorten	Customise
Markup	Yes	Partial	No	Yes	by user	Yes

Semantic markup is supported by various web technologies, notably including RDFa and Microdata. FHISO is aware of only a few uses of it for citations, none of which have yet gained a large user base.

Semantic markup can bypass the property name explosion problem that *property sets* face by allowing much of the citation to be provided without markup. However, partial markup also limits software's ability to perform more advanced *processing* operations such as validation and de-duplication.

Lists of Layers

The *lists of layers* model has been proposed in various forms, but the common design is to combine some single-layer citation model (typically a variant of the *property set* model) with a linear provenance model. Full citation text would then be generated by creating the citation text for each layer and combining them in order with connectors encoded in the list structure.

Data Model	Display	Process	Generate	Provenance	Shorten	Customise
LoL	Yes	Yes	Yes	Branchless	Yes	No

FHISO is unaware of any implemention of *lists of layers* citation data.

The list-based structure of provenance provides for simple creation of citation text but is not directly compatible with branching provenance. Possible solutions whereby branches are either combined into a single layer or serialised into a list using a topological sort have been proposed but have not yet been designed or discussed in detail.

Lists of layers may also be compatible with *semantic markup*; that possibility has not yet been discussed by FHISO.

Graphs of Layers

The *graphs of layers* model has been proposed as an extension to *lists of layers* to handle branching provenance. Instead of a list of layers, layering information would be stored as a tree structure with the consulted source as the root and the sources and stewards it came from as branches. Using a directed acyclic graph (DAG) instead of a tree could also represent merging provenance, such as might arise when stewardship ordering was ambiguous.

Data Model	Display	Process	Generate	Provenance	Shorten	Customise
GoL	Yes	Yes	Yes	Yes	Yes	No

FHISO is unaware of any implemention of *graphs of layers* citation data.

Graphs of layers might also be compatible with *semantic markup*, but because it is nonlinear in structure the sequential nature of markup might make that more difficult than combining *lists of layers* with *semantic markup*.

Graphs of Items

Graphs of items is like *graphs of layers*, but with per-layer information stored in a graph-based model instead of a *property set* model.

Data Model	Display	Process	Generate	Provenance	Shorten	Customise
GoI	Yes	Yes	Yes	Yes	Yes	No

FHISO is unaware of any implemention of *graphs of items* citation data.

The primary advantage of a graph over a set for describing a single layer is flexibility and extensibility. Sets of properties require separate property names for each property of each item, meaning they must consider all possible item relationships in advance and give a different name to each property of each item in each such organisation. Graphs can instead use the same vocabulary for each property regardless of what item it applies to, with a separate vocabulary for item relationships. Both can then be used with a hierarchical ontology; for example, book title, page number, and personal name could all be subtypes of the generic “label” property and *templates* could chose to either handle them separately or by defaulting to the handling of all labels.

Summary

Purposes:

1. Help others locate the source.
2. Acknowledge assistance rendered by others.
3. Provide indicators of relevance and reliability.
4. Conform to an expected style guide.

Algorithms:

1. Display citation to user.
2. Process citation set.
3. Generate citations to conform to a style guide.
4. Shorten citations for repeated reference.
5. Customise citation text beyond what can be done algorithmically.

Models:

Data Model	Display	Process	Generate	Provenance	Shorten	Customise
Text	Yes	No	No	Yes	by user	Yes
PS&T	Yes	Yes	Yes	No	Yes	No
Markup	Yes	Partial	No	Yes	by user	Yes

Data Model	Display	Process	Generate	Provenance	Shorten	Customise
LoL	Yes	Yes	Yes	Branchless	Yes	No
GoL	Yes	Yes	Yes	Yes	Yes	No
GoI	Yes	Yes	Yes	Yes	Yes	No

None of the models by itself provides every desirable functionality of a citation. In general, the more expressive the data model, the more complicated it is, creating a design tradeoff. Before investing in a full development of any one of these solutions, it may be wise to learn what level of complexity is acceptable and which functionality is most important to the broader community.