fhiso

CFPS 112
(Call for Papers Submission number 112)

Desirable Citation Properties

Submitted by: Tychonievich, Luther

Created: 2014-08-20

URL: Most recent version: http://fhiso.org/files/cfp/cfps112.pdf
This version: http://fhiso.org/files/cfp/cfps112_v1-0.pdf

Description: A list of properties that a citation model should strive to achieve. It is unlike that any single model could achieve them all.

Keywords: Citations, sources, properties

Desirable Citation Properties

Luther A. Tychonievich

20 August 2014

Abstract: *Citations, meaning references to real-world documents and sources, are a fundamental component of family history data. There are many desirable properties of a citation, some of which appear to be in conflict with others. This CFPS provides a list of properties to consider.*

1 Definition and Introduction

In this paper, as in CFPS 65, I use the word **citation** to mean "descriptions of real-world things, which descriptions identify the things primarily by publication, location, or origin rather than by content, appearance, or meaning." I do not use it to mean formatted strings, but rather the data those strings are intended to convey. I use the word **source** to mean the things citations describe.

Three previous CFPS have addressed citations. In CFPS 63 Tony Proctor proposed an extensible source-type solution using URIs [1]. In CFPS 65 I described how citations can be thought of as lists of key:value pairs [2]. In CFPS 78 Louis Kessler emphasised that source data, not formatted citation text, is the appropriate medium of communication [3]. All three suggest that citations are stored and transferred as computer-understood data rather than as human-targeted text (though such text could be derived from the data).

Anytime data exists, it is worth considering the properties of the data and the operations that might reasonably be executed on it. In Section 2 I attempts to enumerate a set of desirable properties of citation data without suggesting and particular data model that meets these properties. In Section 3 I offer some observations on potential difficulties a data model might face in trying to achieve these properties.

2 Citation Properties

There are many possible properties a citation model might be expected to satisfy. I attempt to list these in approximate order of obviousness, though such an ordering is difficult to define.

In this section I use the convention that capital letters X, Y, and Z are sources and lower case letters x, y, and z, with or without subscripts, and citations generated to refer to their upper-case variants. For example, if X were a physical grave marker then x, x_1, x_2, and so on would each refer to a citation describing that grave marker.

2.1 Referential

Given a source X and a citation y, it should be easy to determine if y refers to X or to a different source.

Citations that are not referential are of little if any practical use.

2.2 Identifying

A citation x should refer to at most one source. That is, if x refers to X and $X \neq Y$ then it should always be the case that x does not refer to Y.

A slightly different way or putting this property is

$$(X \neq Y) \Rightarrow (x \neq y).$$

See also Section 2.8 for a more nuanced version of identification.

2.3 Constructable

Given a source X, an amateur family historian should be able to create a citation x that satisfies all of the properties that a citation is supposed to satisfy with very little chance of error and without needing to invest any significant effort.

Citations may be more or less constructable; it is not a binary property. The less constructable they are, the more likely people are to either omit them or create them incorrectly. A good tool or user interface might significantly increase the constructability of a citation.

2.4 Comparable

Given two distinct citations ($x \neq y$) it should be possible to determine if both citations refer to the same source ($X = Y$) or not ($X \neq Y$). This should be possible from the citations alone and not require appealing to the the sources themselves.

See also Section 2.8 for a more nuanced version of comparability.

2.5 Coverage

For every conceivable source X, there should exist at least one valid citation x. This includes documents, conversations, monuments, user memory, and possibly even sources of dubious validity such as hunches. That is,

$$\forall X \; \exists x.$$

See also Section 2.7 for another aspect of coverage.

2.6 Locating

If the cited source is one that can be consulted repeatedly* then the citation should provide sufficient information in order for a new researcher to locate and consult the source themselves.

A locator is generally also an identifier; however, you can have locators that are not comparable, readily constructable, or even referential. Locatabiltiy is also an analog property: a citation might make finding a source possible without making it particularly easy.

2.7 Multilingual

Citations should be easily generated by speakers of any language about sources in any language or languages, including obscure or dead languages. This should be true even if the person generating the citation does not know the language(s) used in the source.

See also Section 2.9 for a stronger form of language-independence.

2.8 Multigranular

It should be possible to cite sources at various levels of granularity. For example, it should be possible to cite a book as a whole, a particular chapter or page from from the book, or even a single word on a page.

Multigranularity is a refined version of identification (see Section 2.2).

Multigranular constructability (see Section 2.3) suggests that a courser-grained citation can be constructed from a finer-grained citation. Multigranular comparability (see Section 2.4) suggests that we can tell if one citation is a sub-citation of another. With both constructability and comparability we can be given a set of citations and either generate a super-citation of all members of the set or assert that no such shared super-citation exists.

In all multi-grained citation systems that I have seen so far the finer details are locators (see Section 2.6) within the coursely-cited source. Fine-grained data could in theory be non-locational (e.g., I could say "my second-favourite paragraph") but I am unaware of existing models with non-locational sub-citation fields.

---

* A document, monument, or recording can generally be consulted repeatedly; a conversation, memory, or private or destroyed document cannot.

There is an argument to be made that all sources should be repeatably consultable and that what you should cite is not a conversation but rather a document describing the conversation, even if you have to create such a document yourself. Conversely, if provenance is traced, many sources would be seen to derive from (and thus presumably cite) a transient source like the conversation a census taker has with a resident of a home.

If transient sources exist, they cannot meaningfully have locating citations.

2.9 Translingual

Excepting those portions of a citation that are extracts from the source (and thus in the source's language), all portions of the citation should language- and locale-transparent, easily presented in any language a user happens to desire. This means that neither the language used by the creator of the source, the language(s) used in the source itself, nor the language(s) known to other users accessing the citation should impact any of the other properties of the citation, including comparability (see Section 2.4) or locating sources based on citations (see Section 2.6).

See also Section 2.7 for another aspect of language-independence.

2.10 Canonical

There should be a single one-to-one mapping between sources and citations. That is, $x = y$ and $X = Y$ should be equivalent statements.

Canonical citations imply identification (see Section 2.2) and comparability (see Section 2.4).

2.11 Provenance

There are many possible provenance properties, allowing citations to store information about the chain of sources that led to the cited source. I do not enumerate them here in part because I believe that provenance is not part of a citation but rather the conclusions of research about the origin of the source; I thus assume that provenance should be stored as a set of research decisions connecting various citations.

However, I acknowledge that other opinions on this topic exist and that I am barely a novice at provenance work. A citation standard should almost certainly at least consider provenance.

3 Discussion

There is natural tension between constructability (Section 2.3) and just about every other property I have identified. I thus do not believe that a useful citation model can have a high level of constructability in the data alone; a smart user interface will be needed to help users create citations with other desirable properties.

It is likely that no citation scheme can achieve canonical, constructable citations with good coverage (see Section 2.3, Section 2.5, and Section 2.10). Even if canonical citations could be made constructable, achieving coverage would mean having a canonical form for a huge set of possible citation types and the likelihood that most users would never get confused and use the wrong type appears to me to be quite small.

References

[1] Tony Proctor. "Proposal for Handling Sources and Citations." *FHISO Open Call for Papers* CFPS 63. `http://fhiso.org/files/cfp/cfps63.pdf` Retrieved 2014-08-18.

[2] Luther Tychonievich. "A Gradual Path to Standardised Citations." *FHISO Open Call for Papers* CFPS 65. `http://fhiso.org/files/cfp/cfps65.pdf` Retrieved 2014-08-18.

[3] Louis Kessler. "Nine Nedessities in a GEDCOM Replacement." *FHISO Open Call for Papers* CFPS 78. `http://fhiso.org/files/cfp/cfps78.pdf` Retrieved 2014-08-18.