fhiso

CFPS 18
(Call for Papers Submission number 18)

Proposal to Accommodate Mark-up in Narrative Text

Submitted by: Proctor, Tony

Type: Technical proposal

Created: 2013-03-08

Last updated: 2013-04-20

URL: Most recent version: http://fhiso.org/files/cfp/cfps18.pdf
This version: http://fhiso.org/files/cfp/cfps18_v1-1.pdf

Description: Proposal to accommodate mark-up in all forms of narrative text

Keywords: Text, Narrative, Mark-up

Contents

1. Abstract

Narrative text (as opposed to simple titles, descriptions, etc.) is used for many different purposes. These include transcriptions, translations, abstracts or extracts, reasoning, complex citations, reference notes, To-Do lists and research logs.

This proposal is for such text to be represented as “rich text” rather than “plain text”. This would involve the adoption of a mark-up language.

2. Proposal

There are two forms of mark-up that are required for family history narrative:

- **Presentational**: This mark-up controls the layout and presentation of the text.
- **Semantic** (or **descriptive**): This mark-up provides information about part of the text without indicating how it should be handled or depicted.

2.1 Presentational Mark-up

Presentational mark-up is easy to understand. Following the XHTML precedent, it is proposed that layout attributes such as
, <p>, <ol>, <ul>, and <li> be supported, and logical visual attributes such as and . Control over explicit physical attributes such as colour, bold, italic, underline, font name, and font size is best left to the software tool presenting the text. Software products may well provide style galleries to display different data with different combinations of physical attribute, and so explicit selection of them in the data would hinder that gallery support.

NB: the XHTML strikethrough element <s> is not included in this section since it’s providing semantic information rather than merely presentational.

2.2 Semantic Mark-up

Semantic mark-up is often described as “descriptive mark-up”. However, there are really two levels of semantics and the conventional terminology does not account for the distinction. Rather than try to differentiate the semantic/descriptive terms, and so conflict with the accepted terminology, I’ll introduce two newer terms:

2.2.1 Shallow Semantics

This is the accepted meaning of semantics when people refer to the Semantic Web and semantic mark-up for Web pages. When a datum is a personal name or a place name, for instance, then such mark-up identifies the text as representing the name of a person or a place. What it does not do is identify that person or the place.

This type of mark-up, of which schema.org is a well-known example, is therefore very useful for evidence rather than conclusions. When evidence is transcribed then you want to know where such references are, but the interpretation is not part of the evidence itself.

2.2.2 Deep Semantics

This type of semantic information identifies a person or a place rather than merely the name, and so it’s more appropriate for representing conclusions. Genealogical data (including family history) is based around conclusions and so we need this more powerful type of semantic mark-up.

Attaching a deep-semantic tag to a reference is actually a superset of attaching a shallow-semantic tag. The original reference is still visible, and so may still be searched or indexed, but it also links to information about the entity being referenced – information that cannot be deduced from the reference alone.

2.3 TEI

The Text Encoding Initiative (TEI) is a type of XML-based semantic mark-up. It is mentioned separately here because it is a powerful concept that could be incorporated into a Data Model standard. Unfortunately, it is focused around the structure of a single “text document”.

It appears to have begun as a representation of shallow semantics and document structure (headings, etc) in a single document. It has since been extended to address personographies (biographical/prosopographical information) and placeographies (details of places). A TEI Personography Task Force was chartered in January 2006 to consider issues relating to the former.

In concept, this is very close to supporting deep semantics but the reliance on a single TEI Document is a hindrance. Genealogical data (including family history) involves many interconnected entities, such as Persons (with their many alternative names), Places (with their time-dependent hierarchical relationships), Dates (with granularities, uncertainties, and world calendars), Events (with possible durations and hierarchical relationships), Sources, and Citations. Narrative text is therefore part of a genealogical dataset, with a well-defined schema, as opposed to genealogical data being part of a text document.

2.4 STEMMA

STEMMA’s “structured narrative” approach is simpler and evolved through an effort to truly integrate narrative text into genealogical data.

Its narrative text can use mark-up to reference genealogical entities such as Person and Place via their ID. The references can either generate a defined canonical title (e.g. “John Smith (1812)” or “London, England”) or be used to tag an existing piece of text (e.g, “grandmother”).

Mark-up can be attached to a date reference to associate a machine-readable version. This is ISO 8601 for Gregorian dates but it also accommodates other world calendars. The references can include an element of granularity and/or uncertainty. This form is equally as much a conclusion since the original evidential text (e.g. “Last February”) may not provide the full context.

STEMMA also has entities for sources, citations, and resources (e.g. images or scans) and these can similarly be referenced by their IDs. The scheme also allows other segments of text to be referenced in order to create reference notes.

3. Not Covered or Not Required

While this proposal uses XML example syntax, and even cites XHTML as a precedent, the proposal does not mandate XML as a serialisation format. The same principles apply to all such formats associated with a standard.

The proposal does not cover indefinite transcription characters. Some text storage schemes use a type of ‘regular expression’ syntax to indicate characters that may be unreadable or which may be one of several alternatives. This is a requirement of the Data Model since not all evidence can be transcribed with total confidence, and any search must be aware of the uncertainties, but it will be covered in a separate proposal.

4. Illustration

This very simple illustration uses STEMMA syntax to link a person reference to a Person entity in the data, and generate a place reference from a Place entity. It also identifies a date

My <PersonRef Key=’pAEProctor’> grandmother </PersonRef> was born at <PlaceRef Key=’p3PoplarTerrace’/> on <Date Value=’1903-03-17’>St Patricks Day, 1903</Date>.

Note that the person reference does not involve a name and so it would not be searchable without deep semantics. The place reference is automatically generating a hierarchical reference using an appropriate database. The date reference is an informal one supplemented by the real date.

This might generate the following text on the screen or in a report:

> My grandmother was born at 3 Poplar Terrace, Thorneywood Rise, Nottingham on St Patrick's Day, 1903.

If this were presented on a screen then those known entities could be selectable links, as implied here, that would take you to some relevant data. For instance, part of your family tree, some place information held by a Place Authority database, or some timeline from your collected data.

5. Use Cases

The adoption of rich text in all forms of narrative increases the power of a model in so many ways. The use-case I'll provide is one of an old family letter.

Scanning the letter and including the image file along with your data adds no value from that letter. Transcribing the text into a document is slightly better, although the text is not fully integrated with the rest of your family history data, including your family tree. Furthermore, searches of the transcription are going to be plain-text searches, and so will be looking for a specific representation of someone's name, or a place reference, or a date. Note that there's a huge difference between searching for a person/place/date and searching for a specific representation of a person/place/date reference.

Now let's assume that we have transcribed the letter using a tool that allows us to add mark-up. If this mark-up supported shallow semantics, as with schema.org, then we would at least know which pieces of text constituted a personal name, or a place reference, or a date. However, you need mark-up that supports deep semantics in order to know that a "grandmother" reference corresponds to a specific person, or that "Saturday" corresponds to a specific date based on the context of the letter.

Searching is then not only unambiguous but much richer. For instance, searching for a particular person would automatically use all the acceptable names recorded for that person, and irrespective of whether the reference used a name at all (as in the "grandmother" scenario).

6. Recommendation

FHISO should establish communication with the TEI group in order to agree on a way of properly solving these issues in family history data. Without some change, it does not appear that the current TEI specification is adequate for real-life family history data.

7. References

Terminology for mark-up languages. http://en.wikipedia.org/wiki/Markup_language.

Semantic mark-up for Web pages. http://www.schema.org/.

Text Encoding Initiative (TEI). http://www.tei-c.org/index.xml.

TEI – Getting Started. http://tei.oucs.ox.ac.uk/GettingStarted/html/.

TEO Report on XML mark-up of biographical and prosopographical data. http://www.tei-c.org/Activities/Workgroups/PERS/persw02.xml.

STEMMA ‘Importance of Narrative’. http://www.familyhistorydata.parallaxview.co/research-notes/importance-narrative.

STEMMA Narrative format. http://www.familyhistorydata.parallaxview.co/home/document-structure/narrative-structure.

STEMMA Worked Examples. http://www.familyhistorydata.parallaxview.co/data-model.

FHISO cfps 34, Evidence and Conclusion. http://fhiso.org/files/cfp/cfps34.pdf.