Using the Text Alignment Network for Scholarship on Intertextuality

Joel Kalvesmaki

Journal of Data Mining and Digital Humanities

Using the Text Alignment Network for Scholarship on Intertextuality

Joel Kalvesmaki[1]

1 Dumbarton Oaks

*Corresponding author: Joel Kalvesmaki kalvesmaki@gmail.com

Abstract
The Text Alignment Network (TAN) is a suite of XML encoding formats intended to serve anyone who wishes to encode, exchange, and study translations, paraphrases, adaptations, quotations, and other varieties of text reuse. This article briefly introduces TAN, and in the spirit of the special issue of this journal focuses on the syntax of its intertextual pointers, which are styled to be both human-readable and -interoperable. Because TAN is at present an experimental format, this report notes progress, promise, and future prospects.

keywords
XML, TAN, TEI, alignment, intertextuality, canonical references

I Intertextuality and XML

For the scholarly study of texts, encoding in XML has been critical, particularly in the Text Encoding Initiative (TEI), which is treated as a standard in the digital humanities. Its enormous library of elements and attributes (567 and 258, respectively, in the schema TEI All) covers a great number of textual feature types that scholars most want to study and annotate. Its laissez-faire approach to markup supports an enormous range of research assumptions and questions. This breadth has allowed text markup to flourish, but serious challenges remain for cross-project interoperability (Schmidt 2010, Schmidt 2014). The TEI guidelines require interpretation, and even scholars who agree on how to interpret them may validly mark the same textual features in different ways. Such pluralism means that if you wish to incorporate someone else's TEI file into your project, you must first study it, then write an algorithm to modify it to suit the needs of your project. The exercise must be repeated for every imported file, and the work that goes into this ends up benefitting only your project. Everyone else needs to conduct the same exercise on their own.

Even the humble cross-reference, a claim that passage x in work A is a quotation of passage y in work B, is not interoperable in TEI. I have argued elsewhere (Kalvesmaki 2015) that, at present, simple TEI statements expressing quotations, the molecule of intertextuality, themselves require interpretation, and are applicable only within a single project, and of specific files, and are not predictably applicable to other versions of the same work. That problem is true of other types of intertextuality as well, such as version alignment. TEI is not alone. Other markup schemes such as TMX and XLIFF offer ways to align multiple versions of the same work, but the formats were not designed to try to make data semantically interoperable across projects using the same markup method.[1]

II A New Approach to Intertextuality: The Text Alignment Network

Over the last several years I have been at work to test the hypothesis that the ordinary way we describe intertextual connections provide a syntax sufficient to semantic interoperability for

[1] For TMX and XLIFF see https://en.wikipedia.org/wiki/Translation_Memory_eXchange and https://en.wikipedia.org/wiki/XLIFF.

cross-references and alignment. This hunch has been the motivation for the Text Alignment Network (TAN), a suite of XML encoding formats, including a customization of TEI All, intended to serve anyone who wishes to encode, exchange, and study translations, paraphrases, adaptations, quotations, and other varieties of text reuse. The TAN syntax is meant to be maximally readable and editable by both humans and machines. RELAX-NG and Schematron schemas not only ensure that files are maximally likely to be syntactically and semantically interoperable, but provide feedback and help (via Schematron Quick Fixes) to anyone editing and correcting TAN files in an XML editor.

TAN is a format, not a tool. It does not try to identify quotations or to reconcile differences in alignment. Nevertheless, the library of XSLT functions that drive the validation process definitively interpret the format, and can be used to build tools. The function library has been used successfully to create applications that generate indexes, parallel editions, collations, interlinear editions, statistical analysis, and even quotation detection.

TAN is not intended to replace other formats, such as TEI. In fact, TAN transcriptions are TEI-conformant. But a number of design principles depart from those behind the TEI model:

- **Annotations should always be separated from what is annotated (stand-off markup):** Anything that is not a transcription but comments on one—e.g., remarks on quotations, morphological and lexical analysis—should stand apart from the transcription itself. Such stand-off annotation brings many benefits not available with inline annotation, among which is that multiple scholars can edit and remark on the same texts independently and collaboratively, and their annotations may overlap (one of the Great Impermissibles in a single XML file). Plus, stand-off files can be transformed into any number of combinations of derivative inline files, whereas the the reverse is not always true. A heavily marked up TEI file is like a baked cookie. If you wanted it shaped like a star, or you didn't want nuts in it, you have a lot of work ahead of you. Stand-off annotation argues that instead of baking cookies we should make the basic ingredients available, and allow others to bake the kinds of cookies they want.
- **One file per task**: The TAN schemas require an author of a TAN file to focus on a single task on a single item. If someone wishes to transcribe a work and its translation, and note word-for-word correspondences, then at least three files are required. Most important is the requirement that every TAN/TEI transcription file must be restricted to a single version of a single conceptual work found on a single text-bearing object, segmented and labeled according to a single reference system (canonical or not). Books featuring multiple versions of a work need to be broken into individual files (see the baked cookies above).
- **Metadata to focus only on data:** The TAN formats cover quite a range of topics, some of which, e.g., tokenization patterns, were never meant to be described by TEI. I have adopted the principle that no matter what the type of file, if it is to be useful to scholars it must *de minimiis* answer a common core of questions : what is the name and version of this file, what are the sources of the data, under what license is the data released, who created the data and when, and what assumptions and definitions were adopted in editing the data. Because these questions must be answered time and again across every format, the <head> of every TAN file follows a common structure, and focuses exclusively on describing the data at hand, not the metadata itself. For example, the persons responsible for editing a file must always be named, but in that file we don't need to know their age, nationality, or background. (Persons must always identified with IRIs, so those who wish to know more can go elsewhere more authoritative.)
- **All data must be human- and computer-readable :** Everything stated in a TAN file must be both human- and computer-readable. The computer-readable component is oftentimes an Internationalized Resource Identifier (IRI, roughly synonymous with URI, Uniform Resource Identifier, i.e., a URN or URL).
- **Deep validation :** TAN is highly regulated, governed by an extensive body of rules written in XSLT that probe the content more deeply than TEI schemas do. For example, validation will mark as erroneous any text that is not normalized (according to Unicode NFC). Those same rules will also provide contextual help to editors through Schematron Quick Fixes, which, when invoked with a keystroke or two, replace the text with a normalized version. The

extensive library of functions that perform such tasks also definitively interpret the format and serve as a foundation for preprocessing TAN files or as a starting point for programmers who wish to develop tools and applications that import, export, or otherwise use the TAN format.

Extensive documentation is available at the project's website, http://textalign.net. (All material is released under a Creative Commons Attribution 4.0 license, to encourage reuse.)

Rather than reduplicate that material here, I wish to highlight for readers of this special issue the reference syntax that undergirds the TAN pointing mechanism. For comparison it is worthwhile reminding us of some pointing mechanisms that are well known, and have been around a while:

- URL : http://example.com/index.htm, http://example.com/index.htm#bookmark
- XPointer : <xi:include href="index.htm" xpointer="bookmark"/>
- XPath: select="//tei:div/tei:p[ancestor::tei:div/@type = 'chapter']"

All of these pointing methods have in common a dependence upon the names of files and elements for navigation. The XML structures are deeply reflected in the syntax, which some of which is human-readable, and some of which not.

TAN has been designed with the question, can we make those reference systems even more human readable, and improve their syntactically and semantically interoperability, without terrible loss in performance? In ordinary conversation, when we wish to state that one passage is a quotation from another (a form of two-way pointing) we say, for, example, « Matthew 4:15-16 quotes from Isaiah 9:1-2. » To replicate this sentence using any of the three methods above, we would almost certainly need to depend at least partly upon unfamiliar or cumbersome syntax. We would need to use arbitrary conventions not tied to any semantics (e.g., that the string « Matthew » names a literary division defined as a book) and dependent upon the nomenclature adopted by a single file. So the statements would be applicable to only one document on each end of the cross-reference—not very readable, not very interoperable.

Elsewhere in this special issue we have read about Canonical Text Services URNs, which has been to my knowledge the first generalized attempt to address the issue of syntactic interoperability. In the case of our example statement above, we could begin with CTS URNs such as this (adopting the catalog numbering of the Thesaurus Linguae Graecae [TLG]):

- urn:cts:greekLit:tlg0031.tlg001:4.15
- urn:cts:greekLit:tlg0031.tlg001:4.16
- urn:cts:greekLit:tlg0527.tlg048:9.1
- urn:cts:greekLit:tlg0527.tlg048:9.2

These URNs reliably point to Matthew 4:15-16 and Isaiah 9:1-2 independent of any mechanism, tool, or server. In theory, the URNs point to any version you might wish of the works cited. But they are still just as cumbersome, opaque, and unreadable as the other three methods. And although they are syntactically interoperable, they are not semantically so (Kalvesmaki 2015).

In testing the hypothesis I mentioned above, I have decided, like CTS, to commit everything to unique URNs, but to model the syntax as closely as possible to what we write in footnotes. Here is how the above example would be rendered:

```
<claim verb="quotes">
  <subject work="nt-grc" ref="Mt 4:15-16"/>
  <object work="lxx" ref="Isa 9:1-2"/>
</claim>
```

This snippet, drawn from a TAN alignment file that claims to collect New Testament quotations from the Hebrew scriptures, uses human-readable codes that are defined elsewhere.

The values of @verb, @work, and @ref are defined elsewhere, either in the <head> or in the underlying sources, and can be algorithmically converted to IRIs. In the spirit of the Semantic Web, the snippet says, «(New Testament) Matthew 4:15-16 quotes from (Septuagint) Isaiah 9:1-2. » The statement is interpreted to be true for any source that share the same IRI definition for the works mentioned.

Any TAN/TEI file or fragment, and any TAN <div-ref> may be algorithmically converted to a CTS URN. (Unfortunately, the reverse is not true.[2]) Because the underlying TAN/TEI transcriptions are devoted to a single version of a single work in a single, standard reference scheme, the values in @ref are unique and are therefore just as reliable as any other method of pointing. Further, because every division in a TAN/TEI transcription must be defined as a particular kind of division, a computer can draw semantic inferences, that Mt and Isa are labels for books, and that the numbers correspond to chapters and verses.

The TAN reference syntax, which weds human readability with computability, has many other features and offers interesting implications, but they all cannot be explored here. An immediate concern to some readers should be cases where transcriptions use different labels (e.g., « Matt » for « Mt ») or where there are competing, overlapping, or nonexistent canonical reference systems (e.g., the works of Plato and Aristotle or papyrus fragments). These common problems have been anticipated, and are treated at length in the official TAN guidelines and examples.

3. Progress and Prospects

I must emphasize that at this time TAN is still an experimental format, under development. It has been used successfully in service to three different projects : the Guide to Evagrius Ponticus (GEP), the Chrysostomus Latinus in Iohannem Online project (CLIO), and a private project that is translating a fourth-century Christian work from both its original Greek (where it is extant) and from the ancient Syriac translations made of it.[3]

The GEP relies upon TAN files to make available transcriptions of select primary sources. It also relies upon a TAN-c file, which contains RDF-like claims, in conjunction with a Zotero bibliographic database, to generate an extensively documented master checklist of writings from the fourth century monk.

CLIO relies upon a small library of TAN-TEI files to populate a website devoted to the making available the three Latin translations of the eighty-eight homilies by John Chrysostom. Those same files are used to help document differences between different editions of the same translation.

The private translation project has been able to create a library of TAN files and develop a number of XSLT-based tools, including :

(1) Master HTML pages that collate every version (Greek, Syriac, the working English translations, etc.) not just with each other but with juxtaposed commentary and quotations from Aristotle and the Bible. The number and sequence of the versions are easily configured, without touching the underlying TAN files.
(2) Master indexes of quotations from the Bible and Aristotle, sortable in the order of the quoted or quoting work, and including contextual snippets
(3) Statistical profiles (e.g., word counts, hapax legomena, most frequent words).

[2] TAN requires individual things to be defined with IRIs, and a CTS URN is a compound of items that each deserve and require their own URN. For other comments on CTS URNs, see Kalvesmaki 2014.

[3] GEP : http://evagriusponticus.net; CLIO : via https://www.chrisnighman.com/; the private translation project is under contract with Oxford University Press.

(4) Reports that suggest how words in one particular language version are translated by the others.
(5) Quotation detection.

Such progress is promising, but tentative. At the present, TAN needs a few more projects in the digital humanities willing to use it, and actively contribute to the development of the schemas, examples, and documentation. That prospect has low risk, since making data available in the TAN format does not preclude making it available in any other format. By expanding the network, to see what rules work, what rules don't, and what can't be reduced to a rule, the TAN format will become more servicable to scholars everywhere. TAN promises to be the beginning of what could be a new web of primary sources.

References

Kalvesmaki, J. Canonical References in Electronic Texts: Rationale and Best Practices, *Digital Humanities Quarterly*8.2 (2014), http://www.digitalhumanities.org/dhq/vol/8/2/000181/000181.html.

Kalvesmaki, J. "Three Ways to Enhance the Interoperability of Cross-References in TEI XML." Presented at Symposium on Cultural Heritage Markup, Washington, DC, August 10, 2015. In *Proceedings of the Symposium on Cultural Heritage Markup*. Balisage Series on Markup Technologies, vol. 16 (2015). doi:10.4242/BalisageVol16.Kalvesmaki01. http://www.balisage.net/Proceedings/vol16/html/Kalvesmaki01/BalisageVol16-Kalvesmaki01.html

Schmidt, D. Towards an Interoperable Digital Scholarly Edition, *Journal of the Text Encoding Initiative* [Online], Issue 7 | November 2014, Online since 12 November 2014, connection on 24 March 2015. URL:http://jtei.revues.org/979; doi:10.4000/jtei.979.

Schmidt, D. The Inadequacy of Embedded Markup for Cultural Heritage Texts, *Literary and Linguistic Computing* 25 (2010): 337-356. doi:10.1093/llc/fqq007.