Data models for Digital Editions: Complex XML versus Graph structures

Bruder, Daniel
dmb77@cam.ac.uk
University of Cambridge, Vereinigtes Königreich

Teufel, Simone
sht25@cl.cam.ac.uk
University of Cambridge, Vereinigtes Königreich

In terms of longevity and collation of textual data in the humanities, digital data, notwithstanding its potential, still falls short the qualities of the traditionally printed book.

To streamline the diverse and idiosyncratic Digital Editions of the time and to establish a cross- and re-usable, durable digital archive of textual cultural artifacts, in 1988 the Text Encoding Initiative (TEI) was established with the goal to present a commonly shared standard for the transcription of literary, scientific and other forms of text.

As data model, the extensible markup language XML was chosen to assure longevity and exchangeability of the data. However, it turns out that XML, and with it, the data model of the hierarchically ordered tree are questionable choices for the recording of complex texts – as they are commonly found in the humanities – by potentially rendering the data ambiguous on semantic level.

The abstract idea behind the commonly shared tag set for the description of textual data is reflected in the *TEI abstract model* (TEI Consortium 2016b) which uses XML as a serialisation format – but to which it is not bound:

> The rules and recommendations made in these Guidelines are expressed in terms of what is currently the most widely-used markup language for digital resources of all kinds: the Extensible Markup Language (XML) […]. However, the TEI encoding scheme itself does not depend on this language […], and may in future years be re-expressed in other ways as the field of markup develops and matures.

In the following, fundamental limitations of the tree data model are highlighted in spotlight fashion and contrasted with a graph based model for the sustainable recording and long-term archiving of complex textual data.

Limitations of the tree model

Paradoxically, Digital Editions as well as digital archives, tools, platforms and data repositories are not as interoperable in practice as one would theoretically expect from standardised sources. To be able to cross- or re-use data or tools between projects, in practice, serious refactoring and rededication is necessary – e.g. existing web platforms cannot readily be re-used by another project, notwithstanding the fact that the data repositories are fully validating, validating TEI-P5 sources. How is this possible?

As will be shown, this paradoxical situation of factually unattainable interoperability of editions and tools are a direct consequence of the choice of data model.

The decision towards XML and the tree data model is based on the OHCO assumption of text as an Ordered Hierarchy of Content Objects (DeRose et al. (1990); revised in Renear, Mylonas, and Durand (1993)). Contrasting the original goals (TEI Consortium 2016c) of interoperable long-term archivable data repositories with the status quo, this decision towards XML as the serialisation format needs to be critically questioned – particularly since the TEI Guidelines themselves very early on make clear that the assumption of data model behind XML is an improper simplification (TEI Consortium 2016a):

> Surprisingly perhaps, this grossly simplified view of what text is […] turns out to be very effective for a large number of purposes. It is not, however, adequate for the full complexity of real textual structures, for which more complex mechanisms need to be employed.

Already two most basic constellations can lead to a necessary departure from the tree paradigm which could be described as ‘Complex XML’.

These situations are commonly resolved by using workarounds (TEI Consortium 2016d). Although *syntactically* permissible on the level of XML markup, these workarounds establish structures beyond the data model of the tree and can lead to misrepresentation of the data on *semantic*, modelling level, seriously harming effective re-use and long-term archiving.

- Data as well as tools inevitably become idiosyncratic, i.e. they irrevocably need to be handled on individual, project-specific basis; projects increasingly develop ‘private dialects’ and couple philologists and data scientists for actually accessing the data; data and tools are inaccessible to cross- and re-use between projects; finally, the possibility of a common digital archive is lost beyond recall.

- Complex textual structures demand additional annotation to help and guide downstream tooling to not misrepresent the data. The transcription – in spite of valid, conforming data w.r.t. to the XML Schema – cannot automatically, i.e. without human intervention, be unambiguously resolved into its textual variants.
- The necessary supplementary annotation to one- unambiguously describe and model the source sets in motion a vicious circle of exponentially growing complexity in the data. Project-specific, idiosyncratic tools become necessary and must match this complexity. Moreover, such repositories typically suffer from overtagging (Hanrahan 2015), or, in the

worst case need to be abandoned entirely (Schmidt et al. 2006).

- Any further annotation or commentary only ever increases the complexity: any further annotation must match the existing complexity of the amended tree structure to accordingly be integrated; data and tools suffer from a ‘Heisenberg-Effect’ in that any further, more precise description of the source makes the data only ever more imprecise.

Complex XML

In contrast to a simple edition, i.e. one of linear text without any further annotation, the need for ‘Complex XML’, on most fundamental, level arises through:

1. the edition of a non-linear text
2. the edition of a linear text, open for annotation

In essence, anything that is beyond linear text free of annotation cannot adequately be represented by a mono-hierarchical tree model and will need “more complex mechanisms” (TEI Consortium 2016a).

Complex XML through non-linear text

Non-linear text results from editorial operations such as insertions, deletions, substitutions. For instance, recording the genealogical writing process of two undecided variants within the same sentence, yields four different, non-linear potential readings.

```
                  est                              dilet
Lorem ipsum dolor sit amet, consectetur adipisicing elit
```

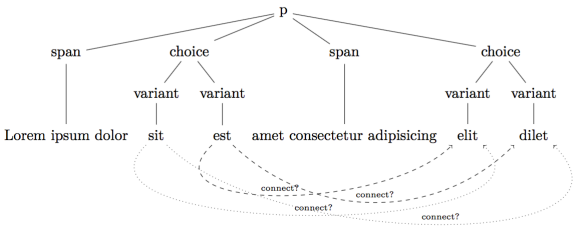
These four different readings derived from mechanical re-combination potentially are not intended and to be reduced to specific readings only.

```
> Lorem ipsum dolor sit amet, consectetur adipisicing elit
> Lorem ipsum dolor est amet, consectetur adipisicing elit
> Lorem ipsum dolor sit amet, consectetur adipisicing dilet
> Lorem ipsum dolor est amet, consectetur adipisicing dilet
```

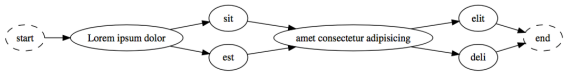
Constraining these combinatorial permutations cannot be done in general ways within the mono-hierarchical tree data model. The tree model exposes a general limitation – even without the prevalence of overlapping structures.

```
<p>
  Lorem ipsum dolor
  <choice>
    <variant id="1-a" connect-with="...">sit</variant>
    <variant id="1-b" connect-with="...">est</variant>
  </choice>
  amet, consectetur adipisicing
  <choice>
    <variant id="2-a">elit</variant>
    <variant id="2-b">dilet</variant>
  </choice>
</p>
```

While interconnecting nodes across the tree’s boundaries by (ab-)using attributes is syntactically possible it nevertheless makes the data idiosyncratic on semantic level, i.e. project-specific rules are introduced and must individually be followed when working with the data.



These interconnections to constrain the combinatorics to specific readings cannot formally be made part of the tree structure itself. To build a tree, any node in the tree must have exactly one parent. A different data model and data structure is necessary to model more than one parent for one node, namely the data model of the graph.



Complex XML through meta-data

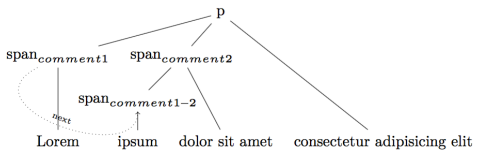
Complex XML can also result from linear text, open for annotation. The following schematic example shows a linear text with overlapping annotation:

```
Lorem ipsum dolor sit amet, consectetur adipisicing elit
`---------' comment1
      `------------------' comment2
```

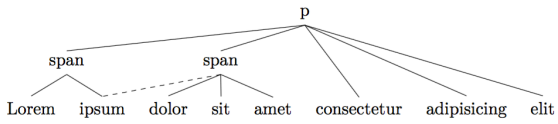
Corresponding serialisation using XML and the segmentation method (TEI Consortium 2016d):

```
<p>
<span id="comment1" next="comment1-2">Lorem</span>
  <span id="comment2">
    <span id="comment1-2">ipsum</span>
    dolor sit amet
  </span>
, consectetur adipisicing elit
</p>
```

The necessary interconnection and recombination of fragmented nodes cannot be modelled within the tree structure in general ways:



Another representation shows how one node in the tree is made the child of two parents:



The relationship between graphs and trees

Trees and graphs are closely related: An ordered tree is a special form of graph with the properties of *a)* it is a directed graph without cycles, *b)* has one designated root node and *c)* any node has exactly one parent node.

As was shown in the previous basic examples, there is strictly no possibility to interconnect nodes of the tree across branches of the tree. By trying to associate two parents to one node, the tree paradigm is effectively abandoned, and results in a permanent need for case-specific handling to resolve potential ambiguities in the data.

Conclusion

Digital Editions wanting to model more than just simple structures can – notwithstanding the syntactical possibilities of XML – not be represented in interoperable ways within the paradigm of the tree data model, making longevity and uniformly re-usable digital archives impossible.

Alternative, graph-theoretic attempts to solve this problem have been suggested and could implement the *TEI abstract model* through an adequate data structure (Huitfeldt 1994; Barnard et al. 1995; Sperberg-McQueen and Huitfeldt 2000; Huitfeldt and Sperberg-McQueen 2001; Durusau and O'Donnell 2002; Tennison and Piez 2002; Dipper 2005; Dekhtyar and Iacob 2005; Banski and Przepiórkowski 2009; Di Iorio, Peroni, and Vitali 2010; Di Iorio, Peroni, and Vitali 2011; Schmidt and Colomb 2009; Schmidt 2014; Götze and Dipper 2006; Peroni, Vitali, and Di Iorio 2009; Witt 2007; Kuczera 2016).

Yet, the question of an adequate serialisation and exchange format to any such data structure remains open. To be able to give guarantees of long term storage and archiving, any such serialisation format must be able to one-unambiguously represent the source as well as data structure. Ideally, any such serialisation format should be both machine readable as well as human intelligible and independent of existing computer hardware and software.

Previous graph-based approaches for the recording of complex textual data either did not catch on or have been abandoned for reasons of complexity in implementation or usage.

Because of the choice of data model, current repositories are idiosyncratic and tools and data must be handled on individual basis. In order to be able to build general digital archives fully interoperable data repositories are necessary. Interoperability is closely connected to the choice of data model. The TEI abstract model should be implemented as a graph structure, however, the graph structure is in need of a suitable exchange and serialisation format.

The commonly shared property between former graph-based approaches is the use of embedded markup. It is conjectured that future research on suitable serialisation formats for graph-based approaches should re-evaluate standoff based markup for the durable recording of Digital Editions.

Bibliography

Banski, Piotr, and Adam Przepiórkowski. 2009. “Stand-Off TEI Annotation: The Case of the National Corpus of P *olish.” In Proceedings of the Third Linguistic Annotatio*n Workshop, 64–67. Association for Computational Linguistics.

Barnard, David T, Lou Burnard, Jean-Pierre Gaspart, Lynne A Price, CM Sperberg-McQueen, and Giovanni Battista Varile. 1995. “Hierarchical Encoding

of Text: Technical Problems and SGML Solutions." In *Text Encoding Initiative*, 211–31. Springer.

Dekhtyar, Alex, and Ionut E Iacob. 2005. "A Framework for Management of Concurrent XML Markup." *Data & Knowledge Engineering* 52 (2). Elsevier:185–208.

DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear. 1990. "What is text, really." *Journal of Computing in Higher Education* 1 (2). Springer Nature:3–26. https://doi.org/10.1007/bf02941632 .

Di Iorio, Angelo, Silvio Peroni, and Fabio Vitali. 2010. "Handling markup overlaps using OWL." In *Knowledge Engineering and Management by the Masses*, 391–400. Springer.

Di Iorio, Angelo, Silvio Peroni, and Fabio Vitali. 2011. "A Semantic Web Approach to Everyday Overlapping Markup." *Journal of the American Society for Information Science and Technology* 62 (9). Wiley Online Library:1696–1716.

Dipper, Stefanie. 2005. "XML-Based Stand-Off Representation and Exploitation of Multi-Level Linguistic Annotation." In *Berliner XML Tage*, 39–50.

Durusau, Patrick, and M Brook O'Donnell. 2002. "Concurrent Markup for XML Documents." In *Proc. XML Europe*.

Götze, Michael, and Stefanie Dipper. 2006. "ANNIS: Complex Multilevel Annotations in a Linguistic Database." In *Proceedings of the 5* th *Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing*, 61–64. Association for Computational Linguistics.

Hanrahan, Elise. 2015. "Over-Tagging with XML in Digital Scholarly Editions." In *DHd2015 Conference – von Daten Zu Erkenntnissen. Book of Abstracts.*, edited by Various, 162–65. Graz, Austria.

Huitfeldt, Claus. 1994. "Multi-Dimensional Texts in a One-Dimensional Medium." *Computers and the Humanities* 28 (4-5). Springer:235–41.

Huitfeldt, Claus, and CM Sperberg-McQueen. 2001. "TexMECS: An Experimental Markup Meta-Language for Complex Documents." *URL Http://Www. Hit. Uib. No/Claus/Mlcd/Papers/Texmecs. Html*.

Kuczera, Andreas. 2016. "Digital Editions Beyond Xml – Graph-Based Digital Editions." In *Proceedings of the 3* rd *Histoinformatics Workshop on Computational History (Histoinformatics 2016)*, edited by Johannes Preiser-Kappeller Marten Düring Adam Jatowt.

Peroni, Silvio, Fabio Vitali, and Angelo Di Iorio. 2009. "Towards markup support for full GODDAGs and beyond: the EARMARK approach." https://doi.org/10.4242/BalisageVol3.Peroni01 .

Renear, Allen H, Elli Mylonas, and David Durand. 1993. "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies." Oxford University Press.

Schmidt, Desmond. 2014. "Towards an Interoperable Digital Scholarly Edition." *Journal of the Text Encoding Initiative*, no. 7. Text Encoding Initiative Consortium.

Schmidt, Desmond, and Robert Colomb. 2009. "A Data Structure for Representing Multi-Version Texts Online." *International Journal of Human-Computer Studies* 67 (6). Elsevier:497–514.

Schmidt, Thomas, Christian Chiarcos, Timm Lehmberg, Georg Rehm, Andreas Witt, and Erhard Hinrichs. 2006. "Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources." In *6* th *E-Meld Workshop, Ypsilanti*.

Sperberg-McQueen, C Michael, and Claus Huitfeldt. 2000. "GODDAG: A Data Structure for Overlapping Hierarchies." In *Digital Documents: Systems and Principles*, 139–60. Springer.

TEI Consortium. 2016a. "A Gentle Introduction to XML" In *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, by TEI Consortium, Version 3.2.0. Last updated on 10th July 2017, revision 0fcf651. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html#SG152 .

———. 2016b. "About These Guidelines." In *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, by TEI Consortium, Version 3.2.0. Last updated on 10th July 2017, revision 0fcf651. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/AB.html .

———. 2016c. "Design Principles." In *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, by TEI Consortium, Version 3.2.0. Last updated on 10th July 2017, revision 0fcf651. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/AB.html#ABTEI2 .

———. 2016d. "Non-Hierarchical Structures." In *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, by TEI Consortium, Version 3.2.0. Last updated on 10th July 2017, revision 0fcf651. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html .

Tennison, Jeni, and Wendell Piez. 2002. "The Layered Markup and Annotation Language (LMNL)." In *Extreme Markup Languages*.

Witt, Andreas. 2007. "Guideline: Multiple Hierarchies." In *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.