Maximilian Eibl, Martin Gaedke. (Hrsg.): INFORMATIK 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 1161

Introducing NoXML for the Digital Humanities

Thomas Efer[1]

Abstract: This paper argues, that a pluralism of data models is needed in the practice of the Digital Humanities. Relevant technology for different levels of model expressiveness should be made available in the form of readily usable tools and infrastructure, together with generic and discipline-specific materials that allow an informed choice of a fitting data model for specific use cases. While it is seen as a vital (and quite viable) task to ensure the interoperability between different DH projects, community-made data encoding schemas, source collections and analysis tools, there is no striking reason to restrain all aspects of those efforts to the traditional XML ecosystem. Therefore a statement is made to venture into employing novel approaches and technologies (subsumed as *NoXML*) that augment the current tooling by providing missing modeling constructs.

Keywords: Data Models; Technology; Digital Humanities; NoXML

1 Motivation

Relational database systems have been the de facto standard for organizing, persisting and querying data for decades. The formative phase for the definition of how abstract and real-life data could be modeled within the relational paradigm began as early as the 1970es, when milestones such as the Boyce-Codd normal form [Co74] and the Entity Relationship Model [Ch76] were introduced. Subsequently, the easily comprehensible table-centric data model accompanied by the (more or less) vendor-independent query standard SQL, led to a major commercial success which resulted in a long-term technological stability of relational technology. It was not until the advent of the Web 2.0 with a rising inherent demand for a huge variety of database features, that other systems slowly moved into focus. Realtime feeds and analytics of social interactions, huge data sets that need large clusters of storage and processing units, unstructured and semi-structured data streams, highly interconnected distributed knowledge resources – all those growing application areas challenged the technical approach of classical database systems. To cope with the new requirements (mainly with respect to scalability and expressiveness) there have been numerous efforts to implement new storage and query mechanisms, each with a fitting data model. The technological progress included simple cases, such as key-value-stores, novel solutions, as big columnar stores, or continuations of older research fields, such as graph databases. All these different projects formed a heterogeneous movement to provide choice and variance among database technology, which has been labelled *NoSQL*, as an acronym

[1] Universität Leipzig, NLP Group, Augustusplatz 10, 04109 Leipzig, Germany efer@informatik.uni-leipzig.de

doi:10.18420/in2017_117

for *Not only SQL*. The wide availability of NoSQL tools has since then been a key-enabler for many modern applications running on tiny embedded systems up to huge distributed and elastically scalable systems. — So how does this all relate to the Digital Humanities?

XML technologies have been the de facto standard for organizing, persisting and querying textual (and plenty of other) data in the humanities disciplines for decades. The formative phase for the definition of how abstract and real-life data could be modeled within the markup paradigm began as early as the 1980es, when the SGML was introduced and through a lot of practical input slowly refined into the XML standard that is maintained by the World Wide Web Consortium. It was not until the recent trend of establishing the Digital Humanities as an independent field (/ a collection of best practices / a trans-disciplinary methodology / . . . [2]) and a lot of interest in digital methods from the humanities disciplines, that a variety of different modeling needs arose in the areas of edition, annotation, processing, exploration and presentation of source material. Surprisingly, the issues of XML (albeit well-known and well investigated, see e.g. [Sc10]) have not gained the wide attention of humanities scholars and computer scientists within the DH community. There is no larger technological or methodological movement to challenge the apparent monoculture of XML formats, source collections and tools. But now it seems, the time is ripe to introduce NoXML[3] into the DH: To state, that there is *Not only XML* available for modeling data. Incidentally some of the existing NoSQL technologies are good candidates for enriching the DH toolbox. For specific aspects there may still be technological solutions to be found so, that there is a need to articulate those needs and requirements in order to enable computer scientists to research fitting solutions. The plea for NoXML should not be misunderstood as an announcement of yet another of the numerous "turns" in scholarly methodology. It is rather to the opposite a call to create an environment that is supporting all the existing and upcoming alternative perspectives on sources and theories with suitable tool sets. It is plain to see, that facts and interpretations that are harder to express in a given modelling setup are more likely to be left aside, given limited resources (which applies to almost any DH project). Therefore a pluralism of possible data models liberates scholars in the long run to take the perspective most fitting instead of the ones most conveniently expressible.

2 XML: One Size Fits All?

The following observations put a focus on textual data since this is still the main type of resource dealt with in the DH. Nevertheless it is likely that other types of media (and especially mixed media research objects) do possess a similar (if not higher) degree of complexity and interconnectedness so that same conclusions apply. For the scholarly digital representation of textual data (such as diplomatic and critical editions and research corpora of various text sorts) there exists a definitive guide, which is compiled by the Text Encoding

[2] A narrow definition of the DH is beyond the scope of this article and would probably be misguided anyway.

[3] The term NoXML was used before in [Le14] to argue for a decoupling of XPath from the XML data model. While this is in line with the arguments in this paper, here NoXML should be understood in a broader sense.

Initiative (TEI) [TE16]. These guidelines are the result of ongoing multidisciplinary scholarly work since the 1980es and represent thus an invaluable basis for all sorts of text encoding needs. Burnard calls the TEI in [Bu14] with good reason *one of the longest-lived and most influential projects in the field now known as the Digital Humanities*.

The modelling roots of the TEI coincide with the work on SGML and early XML so that the representation of certain concepts in the TEI is closely bound to the paradigm of single hierarchies obtained by the nesting of elements to create a tree structure of documents. The so formed textual model is called the Ordered Hierarchy of Content Objects (OHCO) – see [De90] and for a critique and relativization of the original model [RMD96]. While it is indeed possible to express facts and generally represent data in the TEI outside of the OHCO as e.g. [Br05] shows, doubts may arise as whether the resulting documents are still sufficiently human readable or even machine processible by means of standard tooling. The incredible wealth of encoding possibilities covered by the TEI grammar made it quite unwieldy as a whole (and therefore as an XML document type), so that more focused subsets have been created, especially in recent years[4]. But the direct binding to markup theory yield many shortcomings and modeling compromises, for example when mixing XML with binary data or when pointing to external resources. It is however not the aim of this paper to provide a comprehensive list of what may or may not be wrong with XML (or in-place markup and single hierarchies in general). It rather wants to announce that alternatives exist, which may be more suitable for specific use cases. For (non-standard) corpus creation and annotation there have been trends to choose a text representation with reduced complexity for easier and more flexible processing abilities[5]. Other initiatives seek to introduce greater expressiveness into textual models, mainly using graphs and Linked Data technologies with underlying data models such as Property Graphs, RDF statements as well as the (both mildly XML-influenced) OWL ontologies and Topic Maps. They cover issues such as text variants [Sc10], interoperable linguistic annotations [ZR10, Ne15], text mining and information retrieval [Ef15], stemmatology [KA16] and digital editions [Ku16]. There have also been efforts to use the abstract model of the TEI specification as a semantic web vocabulary independently from XML [Tu06, Ci16]. In addition, some approaches try to gently bridge the gap between an existing scholarly hierarchization (of both text collections and sub-work passages) and more versatile data models: With the CITE infrastructure and the CTS referencing scheme [SW09] there exists an important step towards the stand-off annotation of canonical texts via loosely coupled linked data references.

The areas in which XML seems not to be the most satisfying solution include non-destructive text normalization and multiple parallel text views as well as all aspects of versioning and source provenience. Excursions into other data models do moreover also allow to go beyond textual data towards document and author metadata, including external data managed by

[4] see e.g. the different TEI subsets compiled by
- Deutsches Textarchiv (`http://www.deutschestextarchiv.de/doku/basisformat`),
- Institut für Deutsche Sprache (`http://www1.ids-mannheim.de/kl/projekte/korpora/textmodell.html`) and
- the TEI consortium itself (`http://github.com/TEIC/TEI-Simple`)

[5] see for example the work on "mARkdown" by Maxim Romanov: `https://alraqmiyyat.github.io/mARkdown/`

norm data providers. They allow flexible connections with gazetteers and the GIS ecosystem as well as the realization of diverse temporal models – preferably including some kind of "events" as first-class modelling constructs. Maybe some aspects which are currently deemed very difficult to model turn out to be well manageable given different underlying data models. All this is neither a statement for more generic or more specific nor for simpler or more complex models – just for creating a larger choice among them. When regarding XML less as a model or mode of interaction with data and more as a serialization format (applicable for arbitrary data structures), it is obvious that it is a very potent technology with a vast and sturdy existing tool support. Without good reason this should not have to be changed at all. Good reasons could lie in the issues of compression and processing efficiency that may ultimately favour binary formats such as Protocol Buffers[6]. In the case of web services there has been a recent tendency to aim for lightweight JSON solutions for reasons of processing cost and security concerns (e.g. stemming from XML entity resolution). But other than that, there is nothing wrong with exchanging data through XML and even archiving it in that form – as long as this does not imply a strong binding to a single-hierarchy model for the data itself.

3 Desiderata and Outlook

As stated, an advance towards NoXML should not be seen as an attempt of revolt or revolution but rather as a liberating augmentation of the existing DH toolbox. No past resource has to be changed (unless deemed beneficial) but future projects are provided with more flexibility to devise an adequate modelling environment. As the raw technology seems to be already in place or at least underway, it is a valid question to ask: Where is the point of this article? Well, there are still a lot of outstanding tasks until a critical mass is reached:

First and foremost it is important to raise awareness for the matter of technological pluralism and the existence of different modelling approaches (with all their advantages and disadvantages) among the diverse DH communities. Data models matter! Considerate effort – both through theoretical and methodological works as well as with open source implementations and refined infrastructural services – has to be put into creating materials for that purpose. This should allow to gradually replace naïve and often quite unreflecting uses of existing technology with informed choices. According to McCarty a sound reflection of data models and data representation ultimately *[. . .] does not matter to the person interested only in output or effects. But it is crucial to the person, [. . .] who wants to know what is lost in translation, and more importantly what that loss illumines.* [Mc13]

Second, a systematic interdisciplinary exploration of the currently available technology and state-of-the-art modelling research is needed in order to create guidelines of "when to use what" that have to be based on relatable and somewhat paradigmatic humanities use cases. The spectrum of investigations should – in the tradition of NoSQL – well go over

[6] https://developers.google.com/protocol-buffers/

data model questions and therefore consider also technological trends from the surrounding areas such as hypermedia, semantic technologies, business process modelling, metadata repositories and the like. In any case, a wide range of tools and services should be made available to the community of practitioners as the result of those investigations. Ultimately, more research has to be conducted on the implications that different data models have on the practice of modelling and the role (in terms of expressive freedom as well as cognitive load) that a choice among multiple modelling constructs has on performing research tasks. It is likely that those insights show that the relationship between scholar and modelling tool is quite complex and surely goes beyond Maslow's popular hammer.

A likely point of criticism towards the aforementioned ideas and proposed initiatives would be, that they would lead to a blurring of the community and a fragmentation of workforce. While it is in generally not clear why the practice of out of the box thinking, that the humanities usually excel in, should stop when it comes to data models, the point is really not as strong as it may seem at first: The fragmentation is already there, inherent in the different communities. Epigraphy, literary studies and linguistics all deal with texts but do it in such different ways that even when they all use the TEI model, they rarely ever come in touch with one another. The unifying power of this model is also questionable, given the many efforts of creating smaller, better manageable TEI dialects. The question is nevertheless pressing: Can we afford to provide three or four different modelling technologies in parallel? This would indeed be extremely cumbersome and quite irresponsible unless there would be a reasonable level of interoperability between them ensured. On the technical level that may be done via importers, exporters, and completely transparent interfaces (which should still somehow be explicit about what is lost in a conversion). On the semantic level good practices for interoperability have still to be found. There is a certain trend to have a lightweight interlinked rather than a monolithic and centralized approach for this task. But in the end, the use of distributed and interlinked specialized vocabularies together with a semantic alignment via shallow upper ontologies, such as the CIDOC-CRM [Do03], can also be performed parallel to each other – without reducing the overall value. What remains to express, is the hope to inspire a dissemination of new modelling approaches and technology in the DH communities as well as to promote informed data modelling.

Acknowledgements: The author's work is partially funded by the German Federal Ministry of Education and Research under project ScaDS Dresden/Leipzig (BMBF 01IS14014B)

References

[Br05] Bradley, John: Documents and Data: Modelling Materials for Humanities Research in XML and Relational Databases. LLC, 20(1):133–151, 2005.

[Bu14] Burnard, Lou: What Is the Text Encoding Initiative? – How to Add Intelligent Markup to Digital Resources. Encyclopédie Numérique. OpenEdition Press, 2014.

[Ch76] Chen, Peter Pin-Shan: The Entity-Relationship Model – Toward a Unified View of Data. ACM Transactions on Database Systems, 1(1):9–36, Mar 1976.

[Ci16] Ciotti, Fabio; Silvio, Peroni; Francesca, Tomasi; Fabio, Vitali: An OWL 2 Formal Ontology for the Text Encoding Initiative. In: Book of Abstracts of the DH. pp. 151–153, 2016.

[Co74] Codd, Edgar F.: Recent Investigations into Relational Data Base Systems. In (Rosenfeld, Jack L., ed.): Proceedings of the IFIP Congress. volume 6, pp. 1017–1021, 1974.

[De90] DeRose, Steven J.; Durand, David G.; Mylonas, Elli; Renear, Allen H.: What is text, really? J. Computing in Higher Education, 1(2):3–26, 1990.

[Do03] Doerr, Martin: The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. AI Magazine of the Association for the Advancement of Artificial Intelligence, pp. 75–92, 2003.

[Ef15] Efer, Thomas: Text Mining with Graph Databases - Traversal of Persisted Token-Level Representations for Flexible On-Demand Processing. 842, pp. 157–167, 2015.

[KA16] Kaufmann, Sascha; Andrews, Tara Lee: Bearbeitung und Annotation historischer Texte mittels Graph-Datenbanken am Beispiel der Chronik des Matthias von Edessa. In: Konferenzabstracts der DHd. pp. 176–178, 2016.

[Ku16] Kuczera, Andreas: Digital Editions beyond XML – Graph-based Digital Editions. In (Düring, Marten; Jatowt, Adam; Preiser-Kappeller, Johannes; van Den Bosch, Antal, eds): Proceedings of the 3rd HistoInformatics Workshop at the DH. pp. 37–46, 2016.

[Le14] Lee, David: NoXML: Extending the relevance of XPath by breaking the chains of the DOM. In: Proceedings of Balisage: The Markup Conference 2014. volume 13 of Balisage Series on Markup Technologies, pp. 157–167, 2014.

[Mc13] McCarty, Willard: The essential contradiction. In: 6th International Conference Innovative Information Technologies for Science, Business and Education (IIT). p. 12p, 2013.

[Ne15] Neumann, Arne: discoursegraphs: A graph-based merging tool and converter for multilayer annotated corpora. In (Megyesi, Beáta, ed.): Proceedings of the 20th NODALIDA. Linköping University Electronic Press, pp. 309–312, 2015.

[RMD96] Renear, Allen; Mylonas, Elli; Durand, David: Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies. 4, p. 263–280, Sep 1996.

[Sc10] Schmidt, Desmond: The inadequacy of embedded markup for cultural heritage texts. Literary and Linguistic Computing (LLC), 25(3):337–356, Apr 2010.

[SW09] Smith, D. Neel; Weaver, Gabriel A.: Applying Domain Knowledge from Structured Citation Formats to Text and Data Mining: Examples Using the CITE Architecture. Report TR2009-649, Department of Computer Science - Dartmouth College, 2009.

[TE16] TEI Consortium, ed. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.0.0. Text Encoding Initiative, März 2016.

[Tu06] Tummarello, Giovanni; Morbidoni, Christian; Kepler, Fabio Natanael; Piazza, Francesco; Puliti, Paolo: A novel Textual Encoding paradigm based on Semantic Web tools and semantics. In: Proceedings of the 5th LREC. pp. 247–252, 2006.

[ZR10] Zipser, Florian; Romary, Laurent: A model oriented approach to the mapping of annotation formats using standards. In: Proceedings of the LREC Workshop on Language Resource and Language Technology Standards. pp. 7–18, 2010.