Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing

Schriften des
Instituts für Dokumentologie und Editorik

herausgegeben von:

Band 15

Schriften des Instituts für Dokumentologie und Editorik — Band 15

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing

edited by

Elena Spadini, Francesca Tomasi, Georg Vogeler

2021

BoD, Norderstedt

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über **http://dnb.d-nb.de/** abrufbar.

Digitale Parallelfassung der gedruckten Publikation zur Archivierung im Kölner Universitäts-Publikations-Server (KUPS). Stand 5. Dezember 2021.

Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-7543-4369-2
Einbandgestaltung: Stefan Dumont nach Vorarbeiten von Johanna Puhl und Katharina Weber
Satz: LuaTEX, Bernhard Assmann

Contents

Projects and Editions

Appendices

Infrastructures and Technologies

“Standing-off Trees and Graphs”: On the Affordance of Technologies for the Assertive Edition

Georg Vogeler

Abstract

Starting from the observation that the existing models of digital scholarly editions can be expressed in many technologies, this paper goes beyond the simple opposition of ‘XML’ and ‘graph’, It studies the implicit context of the technologies as applied to digital scholarly editions: embedded mark-up in XML/TEI trees, graph representations in RDF, and stand-off annotation as realised in annotation tools widely used for information extraction. It describes the affordances of the encoding methods offered. It takes as a test case the “assertive edition” (Vogeler 2019), in which the text is considered in a double role: as palaeographical and linguistic phenomenon, and as a representation of information. It comes to the conclusion that the affordances of XML help to detect sequential and hierarchical properties of a text, while those of RDF best cover the representation of knowledge as semantic networks of statements. The relationship between them can be expressed by the metaphor of ‘layers’, for which stand-off annotation technologies seem to be best fitted. However, there is no standardised technical formalism to create stand-off annotations beyond graphical tools sharing interface elements. The contribution concludes with the call for the acceptance of the advantages of each technology, and for efforts to be made to discuss the best way to combine these technologies.

1 Introduction

The debate surrounding the best technology stack for digital scholarly edition is ongoing. Recently it has focussed on an opposition between XML and graph technologies. Formally, this opposition does not exist, as you can use XML as a serialisation of graphs (RDF/XML being the best example), and you can express the XML meta-model as a graph (the XML tree is just a rooted and ordered graph). The debate is also a debate between *established* XML users, backed by the wide availability of suitable technologies, and the more recent graph database users[1], experimenting with new solutions, and finding help from a very supportive software company (neo4j.com), for instance. This paper goes beyond the social context and the mathematical models in

[1] See the contribution of Sippl et al. in this volume for an example.

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing. Ed. by Elena Spadini, Francesca Tomasi and Georg Vogeler. Schriften des Instituts für Dokumentologie und Editorik 15. Norderstedt: Books on Demand, 2021. 73–94.

use. It will explore the implicit contexts of the technologies under debate by examining the metaphors used for the meta-models, the serialisations, and the tools applied to create and process data in scholarly editions. It focusses on one type of scholarly edition (for which I coined the term "assertive edition" (Vogeler 2019), which could also carry labels like *historical edition*, *content-oriented edition*, or *semantic edition*) but will try to transfer observations from this type to other editorial genres.

The assertive edition is a type of scholarly edition that focuses on the content dimension of text in Sahle's text wheel (2013, III, 45–49). The assertive edition tries to represent the information that the authors of the texts want to communicate, or that the readers expect to be communicated, and thus the *real* world described by the text. At the current stage of development, it includes the annotation of terms and named entities, and the addition of descriptive metadata traditionally applied to scholarly editions, while extending editorial practice to add a data layer representing the assertions made by the text. The method has precedents in scholarly editions created by historians. The idea of taking textual documents as information carriers reporting historical facts is, of course, widespread in the context of historical research, but has also been applied to scholarly editions of philosophical texts (Pichler & Zöllner-Weber 2013, Pichler 2020). Indeed, *kleio* databases (Thaller 2003–2009), or relational databases/spreadsheets inserting transcription into column-based data structures, can be considered as early examples of the assertive edition. The development of W3C standards for formal representations of graph-based data models made the method more viable. The assertive editions created at the Zentrum für Informationsmodellierung at Graz University in the context of its humanities research data repository and publication platform *GAMS* (Zentrum für Informationsmodellierung 2014–2020), and those created in the context of *symogih*-infrastructure in Lyon (simogih.org 2012–2020, Beretta 2020) mix representations of the data in RDF with TEI/XML. The question is, is this a good choice? What effect does it have?

The paper discusses the technical solutions to realise assertive editions by the affordances of the technology used. *Affordance* refers to the concepts of James J. Gibson (1977) and Don Norman (1988). Gibson describes the affordance of an object as the possible action of any object, and Norman restricts this to the perceivable actions. I would like to apply it not to a single object, but to a technology, in order to formalise a definition. Thus, the affordances of XML encoding, of RDF formal semantics and triple stores, of Graph databases, of annotation tools etc., are here neither only the theoretical mathematical and computational capabilities of these technologies, nor the human–computer interfaces of the individual implementations. Rather, they refer to the mental models of activities associated with the metaphors dominating the technology in daily scholarly editing practice. The affordance of these technologies can be described as the easily perceivable activities of marking up text, structuring text, connecting entities in a text, expressing knowledge as triples, etc.

Affordance, in this sense, includes prototypical and widely used tools and elements of code, but is not restricted to them.

This follows the pragmatic concept of modelling in the digital humanities forwarded by Arianna Ciula, Øyvind Eide, Cristina Marras and Patrick Sahle (2018, Ciula & Eide 2017). It recognises that epistemological work in the Digital Humanities is often based on external meta-models. This work uses non-computational representations, iteratively translates them into computational implementations, and uses the response to modify the model. This highlights the importance of serialisations and tools for knowledge creation both in and with the model, as they can trigger different metaphors.

Therefore, the affordance of a technology can be considered as the trigger for the selection of a non-technological meta-model. The technical solutions use metaphors like 'hierarchies' (e.g., as 'trees', or 'nested lists'), 'annotations', 'links', 'graphs', 'triples' etc., to describe their meta-model. Which possible use cases are brought to mind by these metaphors? When creating an assertive edition, is it easier to think in terms of hierarchical structure and embedded annotation (XML in an XML editor like Oxygen XML or XMLSpy, and using XPath- and XQuery-based querying), vertex and edge relationships (graph technologies such as RDF in triples stores and labelled property graph databases like neo4j), or separation of base-text and annotation (stand-off annotation tools)?

This approach explicitly mixes data model, serialisations and available tools. The human interaction is with a mixture of the three, and it is human interaction with technology that creates the affordance of the technology.

This paper leaves out several technologies that are theoretically available for digital scholarly editing, but have not gained much acceptance in the scholarly editing community. Relational databases, for instance, are extensively used in software engineering, but are not widely used in the field of scholarly editing, so it makes sense not to discuss affordances of XML-enabled relational SQL database systems, or the implementations of conceptual models using SQL instances. The same is true for scholarly editions based on default content management systems like Drupal or Typo3, which are not used by the wider community of digital scholarly editors. Certainly, it would be worth to study social context of technologies as a reason for their acceptance, but this study focusses on the epistemological implications of the technologies.

The same is true for technologies that were previously used to create scholarly editions, but have fallen out of use — or technologies that were only proposed, but never put into editorial practice. This teaches us again, that social context is one of the things that drive a community to adopt one or the other technology. Sometime, it is hard to distinguish if affordance or social context drives decisions, as affordance, in part, shapes the social context. Manfred Thaller's *kleio* (Thaller 2003–2009), for

instance, was a very effective way of representing complex data structures together with the original transcription. However, the programming language used for creating and manipulating these structures did not correspond to the skills taught in computer science introductions, and the software lacked a graphical user interface for a long period. Thus, the emerging digital humanists started their work with tables (and spreadsheet software) and standard relational databases. While SGML, to give another example, had mechanisms to handle overlapping mark-up in concurrent trees, and was therefore well-suited to scholarly editing problems, computer scientists preferred the strict hierarchy of XML. Suggestions for inline mark-up handling overlapping structures like *TexMECS* (Huitfeldt & Sperberg-McQueen 2001), *LMNL* (Tennison & Piez 2002), GODDAG (Sperberg-McQueen & Huitfeld 2004), or *EARMARK* (Peroni 2012–2020; Di Iorio et al. 2009) never offered enough data manipulation possibilities for the technologies to flourish beyond the academic context of proposals and single projects. Excluding these technologies that are not really used in digital scholarly editing reduces the influence of group behaviour in the analysis. It can focus on the perceived affordances of the established technologies, and the metaphorical concepts related to these affordances.

2 XML/TEI

The major technological standard for current editing practice is XML/TEI. For instance, the German research funding scheme DFG this is, for instance, recommended as standard for any scholarly edition (DFG 2015). The digital preservation community accepts XML/TEI as storage format because the definitions of the TEI to encode texts added to the documents gain a semantic explicitness beyond the individual project, and make them fit for digital archiving. Sociologically, the full set of W3C standardised X-technologies offers a well-established technological infrastructure for XML/TEI. The Text Encoding Initiative is probably the largest Digital Humanities semantic data modelling community, and has, since its foundation, focused on creating a terminology as close to the humanities tradition as possible.

Considering the implicit consequences of this technology stack, it is necessary to distinguish between XML and TEI. James Cummings (2018) has argued that the model of the TEI expressed in the TEI Guidelines goes beyond assumptions based on perceived affordance of XML. Indeed, the semantics of the TEI offer so many mechanisms that XML can be considered as merely one possible serialisation of the definitions in the TEI Guidelines. Mathematically, there is no problem in serialising the description of a person in tables instead of in a list of elements nested in the XML/TEI element `<person>`, or to use the TEI parallel segmentation annotation for a

critical apparatus to express textual variance as a variant graph as, for instance, is done in CollateX (Dekker & Middell 2011).

The main purpose of the TEI community is to provide interchangeable semantics to the annotation expressed in the tag labels (TEI Consortium 13.2.2020). The affordance in this approach is the transportation of established concepts in the humanities into a computer-processable formal language, the ability to *speak to the computer*. For the assertive edition, this affordance seems to be highly useful, as it offers a substantial range of semantic annotations for text: names of persons (`persName`), places (`placeName`), geographic entities (`geogName`), organisations (`orgName`), physical objects (`objectName`), and structured descriptions for each of them (`person`, `place`, `org`, `object`). Bibliographic items may be identified in the text (`bibl`, `title`) and described with a variety of nested elements (e.g., `author`, `title`, `publisher`, `date`, `textLang`) or in predefined structures (`biblStruct`, `msDesc`). There is mark-up for terminological words (`term`) which can be linked to taxonomies (`taxonomy`). Index terms (`index/term`) can be associated with positions in, or ranges of, text. The editor can reuse the established concepts to identify functions of text.

Thus, the TEI offers the possibility to enrich the text with interpretations of its meaning, using terms close to natural language for the purpose. Still, affordance based on the semantics of natural language can be confusing. In court records, for example, a person could have the role of a witness; however the TEI uses `witness` to encode a textual witness in the context of critical apparatus. For the assertive edition, the main concern regarding the semantics provided by the TEI is whether it fits to the domain of interest in the particular case. The TEI semantics still offer easy-to-grasp solutions for this by providing attributes to define specialisations of existing labels (`@type`) or reference to interpretations (`@ana`).

The affordance of XML is quite different from semantic tagging of entities: XML annotations follow the paradigm of embedded mark-up in a single rooted tree. In the context of scholarly editions, the main consequence of this affordance of XML mark-up is the conceptualisation of text as a sequence of strings, separated into ranges by start and end tags. These text fragments can carry annotations expressed in the labels of the tags. XML implements the basic ideas of the OHCO (Ordered Hierarchy of Content Objects) model of text (Renear et al. 1990), i.e., the metaphor that text can be handled by super-/substructures, and by order assigned to textual fragments, as 'content', As long as you describe texts as a collection of sentences, and each sentence as a collection of words, the OHCO model fits activities necessary for scholarly editing – and, in particular, assertive editions, where sentences and words build basic entities to represent real-world phenomena. Recently, Steven DeRose has summarised the relationship between the OHCO conceptual model of text and XML and concluded "XML is particularly good for documents not because of syntax details, but because

its native constructs map readily to document models which have proven useful for serious work with non-ephemeral text documents." (DeRose 2020)

However, the metaphorical potential of the element hierarchy of XML extends the OHCO model. Jennifer Tennison describes the distinction between "containment" as a happenstance relationship between ranges of text and "dominance" as the hierarchical relationship with a meaningful semantic (Tennison 2008). Thus, the nesting of elements can either be just a mereological relationship of containers of several objects, or form a semantic context for the nested elements.

The XML definitions of the TEI semantics make use of the *canonical* order of the XML syntax for the *semantic* order of the textual objects modelled. The ranges defined by the TEI mark-up divide the text into a collection of textual fragments, e.g., sections (`div`), paragraphs (`p`), and referencing strings (`rs`, `name` etc.). These ranges can nest, i.e., one textual fragment can be part of another (containment), and this creates semantic context (dominance). For the assertive edition, semantic context is crucial: a headline to a list, for instance, adds semantics to each entry in the list (Goody 1977, Dolezalova 2009). The XML metaphor fits this need, and, in fact, in XML this context can easily be accessed from each entry in the list by an XPath pointing to the containing list and its heading (`./ancestor::list[1]/head`) or just to the first preceding heading (`./preceding::head[1]`).

Thus, we have two affordances to consider: the affordance of the TEI vocabulary, and the affordance of its serialisation in XML. The main affordance of the X-technologies stack remains the manipulation of OHCO, that is, as nested textual fragments. They can easily be addressed by expressions in XPath, which was designed to navigate the hierarchy and sequence of XML elements. Finally, embedded mark-up creates pointers to text ranges which are described in the annotation, imitating, at least partially, manual annotations in a physical text.

3 Graph Databases

The second major data modelling method in the digital humanities is grounded in graph theory. A graph model of text has been considered by digital scholarly editors (Van Zundert & Andrews 2016; Dekker and Birnbaum 2017). The application of graph models was strongly triggered by its affordances for the representation of variants. A graph model can easily represent paths of alternative readings (Schmidt & Fiormonte 2006; Schmidt and Colomb 2009; Schmidt 2010). However, this use-case does not readily apply to assertive editions.

Andreas Kuczera (2016a, 2016b) has made a case for the graph model from the perspective of an historian. He claims that the graph model makes annotation more flexible. In fact, when all textual entities become identifiable entities, they can re-

combine in multiple ways: a linguistic fragment can reference single content entities, such as named entities, it can be part of a complex graph, representing assertions, or it can reference the complex graph itself. Having single words as identifiable entities, co-references can be expanded beyond textual sequences ("dieser – jener", "the first, the second,…."): Expressions like "Count Eberhard and his son donate their property in Schmie to the monastery of Maulbronn. The latter promises to add the Wannenwald after the death of his father" (my example), can use expressions of equivalence between the references "the latter" and "his son" (both to be identified as `viaf:80363599`), or "his father" and "Count Eberhard" (both to be identified as `viaf:80337369`). These abstract identifiers can be considered implementations of the conceptual separation between signifier and signified (De Saussure 2013), or between thought, symbol, and referent (Odgen & Richards 1923); or, more simply, support the idea of the assertive edition, that information conveyed by text can be presented separately from the text itself.

The network metaphor lead to very early realisations of assertive editions: digital scholarly editions can enrich named entities in the text by pointing to identifiers from authority files (Poupeau 2006) or in structured data. The *Carl Maria von Weber-Gesamtausgabe* (Allroggen et al. 2011–2019) serves as an example of a 'linked data edition', as it provides a JSON-LD representation of its rich indices on persons, places, letters, works, documents and even commentaries. They reference the idea of the semantic web as a "giant global graph", in which the reader/user of a scholarly edition can drag information from many resources into the edition (Wettlaufer 2018). The assertive edition includes the concept of linked data editions: globally identifiable data points are a good representation of the semantic layer of a text, and the idea of contributing to the general knowledge graph is a valid metaphor for the general purpose of scholarship.

Graph theory implies an affordance which has made it highly attractive for social analysis: network structures can be visualised as graphs. Edges as links between ideas were connected to associative thinking in Deleuze and Guattari's "rhizome" image of thought (1980). One main attraction of graph technologies comes from visualisations that allow jumping from one node to the next. Force-directed graph drawing methods (Eades 1984; Fruchterman & Reingold 1991) create visual impressions of groups with high interconnection, and give an easy overview of the organisation of a graph. This kind of affordance is used when conceptualising textual relationships as nets (Andrews & Macé 2013; Andrews & Van Zundert 2012). As every social network can be expressed as a graph, the meta-model supports the historian's interest in people connected with one another. This affordance, at the very least, makes graph technologies attractive for the implementation of assertive editions.

Graph-based technologies are linked also to knowledge representation in semantic networks (Quillian 1967). James Hendler demonstrated that a substantial part of

description logic can be expressed with an RDF-based vocabulary (Hendler et al. 2005; Bechhofer et al. 2000). Based on this, the Web Ontology Language (OWL) (Lacy 2005, 134; W3C OWL Working Group 2012) was created. It is a vocabulary that realises graphs that express the logic of OWL-DL. In encoding practice, the class hierarchies, which form the entry screen to the main OWL editor Protégé, are often taken as the main affordance of OWL. In fact, the formal affordance of OWL is not really exploited in Digital Scholarly Editions, or even in Digital Humanities at all. A practical effort to define digital editing-related concepts as a formal ontology is made by the Swiss *NIE-INE* project (*NIE-INE* 2020)[2].

In the context of the assertive edition, a special type of graph comes into play: the sentence. The W3C Resource Description Framework (RDF) adds this metaphor to a basic graph model: 'subject predicate object' structures translate easily into directed graphs (W3C 2014). In promoting the semantic web, this metaphor helps one to talk about the data formalised in RDF. The metaphor even leads to suggestions like that by Roland Kamzelak (2016) of taking an approach to creating RDF triples that is based more on natural language. This suggestion is close to the affordance of controlled natural languages like Attempto Controlled English (ACE) (Fuchs et al. 2006; Fuchs 2018), which are used for knowledge representation. Alexandr Ivanovs and Alexey Varfolomeyev (2014), for instance, have used ACE experimentally in a scholarly edition of charters.

The W3C has realised that this metahpor can be misleading: the label "semantic web" contributed to an unnecessary combination of the concepts of "linked data" with "knowledge representation". W3C has changed the label for its activities in the field since 2013 to "Web of Data". In fact, RDF can be used to model hierarchical knowledge systems as well: The Simple Knowledge Organisation System (SKOS) recommended by the W3C (W3C 2009) proposes a set of broader and narrower relationships which are the major type of relationship in most use cases of SKOS.

The triple-based sentence metaphor of RDF has drawbacks. Firstly, simple 'subject predicate object' sentences do not suffice to express the context of propositions such as *The people of London swear allegiance to King George I. – in 1723* (Vallance 2013) and *The city of Basel received 31 and a half pounds on wine tax – in Füllinsdorf, Lupsingen and Zieten* (Burghartz et al. 2015), which add temporal or geographical constraints to the proposition. The standard RDF solutions for these extended sentences are blank nodes (W3C Working Group 2014, #section-blank-node) and property singletons (Nguyen et al. 2014; Nguyen et al. 2015). Both methods are much less intuitive than labelled property graph technologies, e.g., Apache TinkerPop (Apache Software Foundation 2015–2019) or Neo4j (Neo4j 2010–2020), which allow the description of edges with properties.

---

[2] See also the contribution by Cools & Padlina in this volume.

Therefore, the major affordance of graph databases in an assertive edition is the use of a semantic network to connect pieces of information. The specifications of RDF add the metaphor of a 'statement' to express data as simple propositions. The W3C web-of-data stack adds to this the creation of abstract identifiers, helping to separate text and data, and to prepare the data for integration into a global network of information.

4 Stand-Off-Annotation

4.1 Layers

The assertive edition has a multi-layered approach to text. From the theoretical point of view, this need is well explained by Börje Langefors' "Infological equation" (1966): information is a function of time and data. The content of the edited text is information extracted by the editor from the base text under specific conditions. Editorial annotation and formal representation of scholarly readings of the text should be separated from the text, as Manfred Thaller has pointed out (Thaller 2012). In fact, scholarly discourse on the content of the text of an edition is very often triggered by the attempt to avoid combining scholarly interpretations with documented physical and graphical observations on the manuscripts (Zeller 1971).

There are a number of technologies dedicated to *stand-off annotation*. The separation of primary data and mark-up by semantic links dates back into the 1990s (Thompson & McKelvie 1997) and is the de facto standard in the annotation of image, audio and video material, mainly because the encoding of the annotated data is significantly different from that of the annotation (literals). Stand-off annotation has been applied to linguistic annotation of texts for a long time. Several approaches to encoding this annotation are in use: tabular lists of tokens (e.g., *TCF* Weblicht 2015), pointers to offsets in the text stream (e.g., *PAULA*, Zeldes et al. 2013), or pointers to other anchors in text (ISO-LAF 2012; Ide & Sudermann 2014) such as, for instance, embedded mark-up. The general affordance of stand-off annotation methods for assertive editions lies, on the one hand, in the effective processing of the annotations themselves, as they become separate objects with own properties, and, on the other, in the layered semantic conceptualisation of the text. Here, I want to argue that the combined affordances of XML and RDF are not sufficient for this layered approach.

Stand-off annotation is very often introduced as a solution to the problem of overlapping mark-up in XML. This opposition is informative for the perception of affordances of the technologies. Many stand-off annotation formats use XML for serialisation, and the Guidelines of the TEI have a full section describing how to create stand-off mark-up with the TEI (TEI Consortium 2020b; Cummings 2018). Still, XML is perceived as a strict hierarchy.

As explained above, the implicit semantics of XML attributes support a layered approach. Syntactically, attributes are only one type of nested node. The standard semantics assigned to this type of substructure supports this multi-layered approach: The distinction between the three types of nodes (attributes, sub-elements, and content) suggests that attributes are a separate information layer to the annotated text. In fact, attributes in the TEI can create alternative representations (e.g., `date@when`, `num/@value`, `measure/@unit|@commodity|@quantity`), and this fits in very well with the concept of the assertive edition. The generic `@ana` attribute allows these to be extended to include a layer that is completely defined by the editor.

However, the TEI does not insist on the layer metaphor for attributes: on the one hand, nesting of elements can also create layers of text, e.g., the `choice` or the `app`-element expressing alternative representations of one text. On the other hand, attributes suggest an *isomorphism* metaphor, when they encode specialisations (`@type`), or a network metaphor when they encode references (`@ana`, `@ref`, `@target`, `@spanTo`, `@facs`). This last affordance, i.e., the reference from names or reference strings to formal descriptions of the entities is, as explained above, very close to the needs of the assertive edition, but would be used rather in the context of the layer metaphor than as a network.

This is not a critique of the XML implementation of the TEI, but demonstrates that the metaphors applicable to XML's syntax might be stronger than the semantics proposed by the TEI. The expectation that nesting should be more than just an issue of serialisation, that it should have a specific meaning, drives well-known discussions of the type "why can't element X contain element Y?" In fact, the TEI goes beyond the primary affordance of XML, when it breaks the logic of embedded mark-up: semantic annotation in the TEI, for instance, uses pointers from a linguistic fragment to form a list-like data structure: `persName` points via `@ref` to `person`, for example, and `listPerson/person` describes the person as a list of properties. TEI has introduced other constructs to handle typical drawbacks of XML, like the `@part` attribute for overlapping mark-up, and a more generic method to create a sequence of XML elements beyond the sequence in the document (`@prev`, `@next`).

The distance between the technologies used to process XML and the formal openness of the TEI can be demonstrated in the handling of overlapping mark-up. TEI proposes `@part=I|M|F` or `@prev` and `@next`. Both need complex XSLT or XQuery expressions to create the merged node (listing 1a and 1b). By contrast, stand-off technologies could just use a single `name[offset-start, offset-end]` expression, if the stand-off annotation were expressed by offsets in the range of the basic text.

The same is true for stand-off annotations expressed in XML/TEI, which need to resolve the ID-references provided in `@ref` or `@spanFrom/@spanTo`.

The same observation can be made when expressing layered information with graph technologies: RDF offers the reification vocabulary to express that statements are made

by a person about a subject. Reified statements can express a single interpretation of a given text, and also the fact that this interpretation might be different from interpretations by other scholars. Reification proposes the creation of a statement graph (`rdf:type rdf:Statement`) that is composed of triples describing the role of entities and literals in a statement (`rdf:subject`, `rdf:predicate`, `rdf:object`). The mathematical affordance of RDF is sufficient to model the problem, but the solution does not meet the intuitive needs of scholarly editors. W3C has introduced named graphs as an alternative solution. Named graphs use an IRI to identify a full graph consisting of many triples (Carroll et al. 2005; Bizer & Cyganiak 2014). The method fits much better with the everyday experience of receiving RDF as document on the web via a specific URL.

```
<xsl:template match="*[@next]">
  <xsl:element name="name()">
    <xsl:apply-templates/>
    <xsl:apply-templates select="following::*[@xml:id=current()/@next"/>
  </xsl:element>
</xsl:template>

<xsl:template match="*[@part='I']">
  <xsl:element name="name()">
    <xsl:apply-templates/>
    <xsl:apply-templates select="intersect(current()/following::*[name()=current
        ()/name()][@part='F'][1]/preceding::*[name(), current()/name()][@part='M
        '])"/>
    <xsl:apply-templates select="current()/following::*[name()=current()/name()][
        @part='F'][1]"/>
  </xsl:element>
</xsl:template>
```

Listing 1. sample XSLT code merging TEI encoding for overlapping mark-up: a) expressed via `@next` pointer, b) expressed via `@part`

An alternative approach is to define a vocabulary with semantics dedicated to the layer metaphor: an XML element `<layer>` can create the necessary affordance, and the recently introduced `<standOff>` element in the TEI might serve this purpose. In RDF, the W3C web annotation data model (Sanderson et al. 2017) and ISO 24612:2012 (ISO-LAF 2012) describe stand-off annotations in a more generic way in RDF.

4.2 Tools

There are a range of tools supporting the layered approach. The definition of the web annotation model by the W3C is rich, but implementations in generic web annotation tools like hypothes.is (hypothes.is [2011–2020]) create only plain text annotation.

Other stand-off annotation tools focus on overlapping mark-up. Catma (Petris et al. 2008–2020), for instance, is extensively used in digital philology (Petris et al. 2008–2020, /publications). It foregrounds the basic idea of annotation by using coloured underlines linked to tag sets. The TEI export of this annotation (Petris

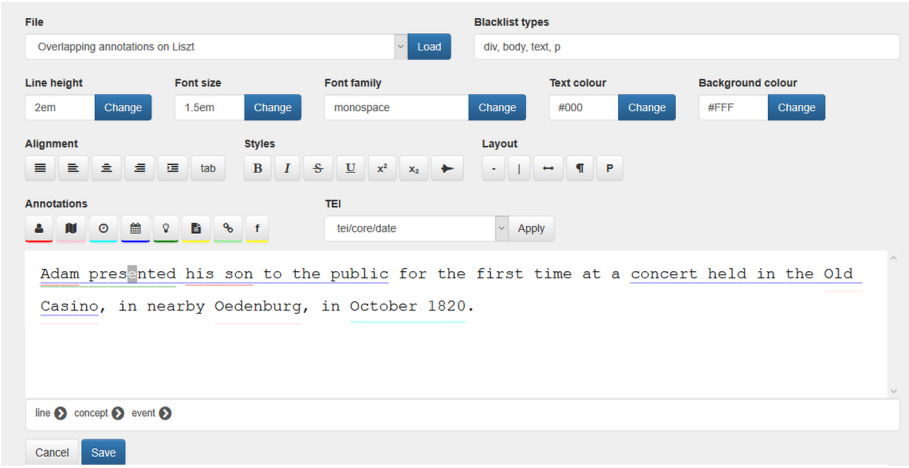

Figure 1. Screenshot from the SPEEDy annotation editor.

et al. 2008–2020, /documentation/tei-export-format/) creates a TEI structure with pointers between text, empty `<seg>` elements, and feature structure declarations (`<fs>`) referenced by the `@ana` attribute.

A similar approach is taken by the annotation tool "SPEEDy" (Figure 1) developed by Iian Neill in the context of *Codex-net* (Neill 2013–2019).[3] It creates rich stand-off annotations, which are exportable in JSON (argimenes 2018–2020). The editor is still close to the surface of the text, and it focuses on allowing overlapping mark-up (visualised as coloured underlines). However, the abstract concept of this tool is different. It considers annotations as *claims* about text (or other entities). Thus, there is the possibility of creating more complex data structures.

Still, in the context of the assertive edition, the current affordance of the 'overlapping-mark-up' approach of Catma and SPEEDy does not extend much beyond embedded mark-up: the annotation is defined by generic tag sets, not by references to individuals.

SPEEDy's 'claim' approach is similar to the factoid model proposed by John Bradley and Harold Short for prosopographical databases (2005; Pasin & Bradley 2015). It models a tripartite data structure, in which the text is only a source of information on a person. The factoid model has been applied to several applications, though typically hidden in the database model, and usually away from the user, behind graphical user interfaces, for instance in the *Personendaten-Repositorium* of the Berlin-Brandenburg

[3] See also the contribution by Neill & Schmidt in this volume.

Academy of Sciences (BBAW 2009; Neumann et al. 2011) and its *Archiveditor* (BBAW 2011).

The Archiveditor creates the semantic annotations needed in an assertive edition. Oxygen plugins such as ediarum (BBAW [2014–2019]; Dumont & Fechner 2014) offer similar affordance in the XML/TEI technology stack: users utilise graphical interface elements to link text ranges in TEI encoding to controlled vocabularies or to lists of entities. Users can switch between the XML representation of the text and separate XML documents for the annotations.

The scholarly edition of the letters of Jakob Burkhardt (Ghelardi & European Research Advanced Grant Project EUROCORR 2019) demonstrates semantic annotation using the RDF stack, which meets the needs of assertive editions. For semantic annotation, it used thepund.it, by Net7 (Net7 [2015–2020]), which creates annotations in the form of RDF-based triples (Morbidoni & Piccioli 2015). It allowed for the insertion of text fragments into triples, linking them to other text fragments or concepts defined by the user (Figure 2).

Using stand-off annotation tools to represent relational semantics is a good fit for the use case of the assertive edition: when applying automatic information extraction methods, the identification of named entities, and their position in syntactic structure, creates a different layer of information from the information usually used in databases. Only the extraction of "Who did What to Whom, and When, Where, and How?" converts the text into propositions in a formal structure, and this does not have to be close to the original text. This conversion includes phenomena like co-reference resolution and entity identification, where the linguistic surface cannot be used as identifiers in a semantic representation.

There are several tools for the annotation of semantic relations, and their affordance leads to a common meta-model beyond the representation of overlapping mark-up. The *BRAT* annotation tool (brat contributors 2010–2018), for instance, has an easy-to-use graphical user interface (see Figure 3), in which the user can identify entities of interest, and link them together. While text fragments represent the entities, label arrows indicate the relationships between the entities. The very similar design of the annotation tool in *Recogito* (Pelagios Network 2014–2020) (see Figure 4) demonstrates a recognition of the basic affordance of stand-off mark-up to be able to create links between references to named entities in a text. They both use flat text as a reference point. *Recogito* exports it to a combination of file-types: RDF files, using web annotation vocabulary; TEI files, with mark-up for the main entity types; and, as CSV lists of nodes and edges, prepared for *Gephi* to represent the relations between the entities. Brat exports as a list of numbered tokens and plain text. Even the far more feature-rich linguistic annotation platform *inception* (de Castilho et al. 2018–2020, https://inception-project.github.io/) makes use of the *BRAT* tool for its annotation.

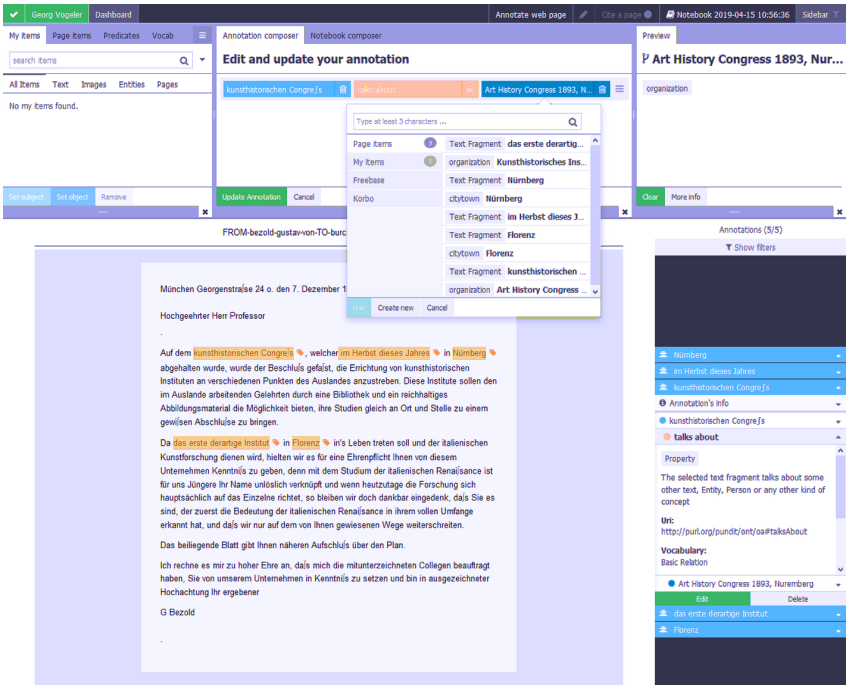

Figure 2. Triples for https://burckhardtsource.org/letter/273 as annotations with thepund.it in burckhardtsource.org/. Screenshot from 2018, in August 2021 the functionality was disabled.

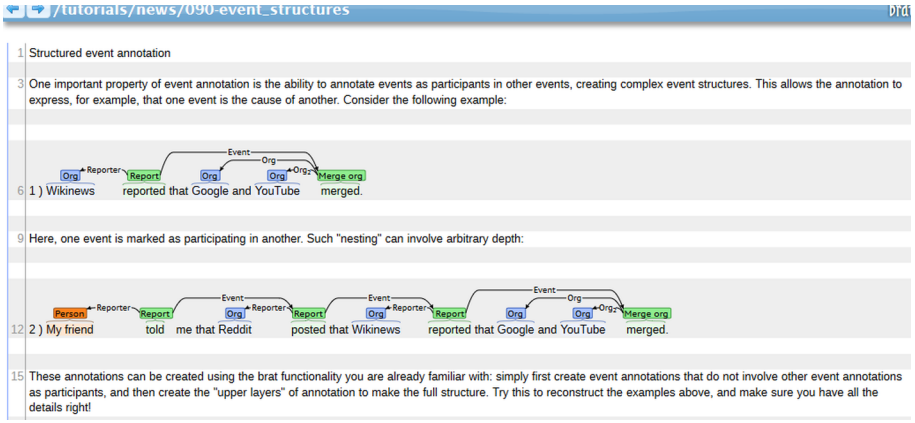

Figure 3. Stand-off annotation with *BRAT*. While *BRAT* itself is out of development, many features have been integrated into *inception*.

Figure 4. Stand-off annotation with *Recogito*.

To summarise the current state of research, the only technologies that provide explicit affordance for stand-off annotation are graphical tools. For the end user this is no problem. For technically informed users it is less comfortable: even if the design language for *BRAT* and *Recogito* is highly similar, the semantics of the stored data structures can vary significantly, as the brief overview of formats to store stand-off annotations above has shown. There is no encoding standard for stand-off mark-up that is as easy to grasp and to manipulate as pointy brackets in XML, mark-up languages that allowing overlapping mark-up like *TAGML* (Dekker et al. 2018), or RDF triples expressed in Turtle.

5 Conclusion and Future Work

In the considerations above, I have identified the *sequence*, the *hierarchy*, the *statement*, the *network*, and the *layer* as metaphors that describe the interactions of the editor with the data in the creation of an assertive edition. Of course, other edition types might prefer different metaphors. The tree is the basic metaphor of classical stemmatology. Recently, the graph metaphor has been particularly successful in the analysis of textual variants, as it can describe the complexity of diverging text sequences that do not produce a hierarchy. Genetic editions build on generative metaphors like ‘parents’, ‘derived from’, or on temporal sequences (‘protograph’, ‘apograph’). Documentary editions have a slight tendency towards a *topological* metaphor, positioning texts in a two-dimensional space, and often realised in stand-off annotations as the standard serialisations of the visual representation (as a matrix cannot easily be inserted into hierarchical models). When Scholarly editing is taken as a basis for linguistic

or philological analysis, it shares the tendency towards stand-off solutions, as this analysis is often conceptualised as a multi-layer annotation.

The affordances of the existing technologies as serialisations, as tools, as well as the conceptual meta-models do not support all of these metaphors in the same way. The editor will therefore select technologies better adapted to single tasks, and combine as many of the technologies as possible. The debate about the best technology stack should therefore move towards a debate on the best method for a given combination. To facilitate the interchange of data we should take care to avoid implicit semantics, for instance, by making relationships in TEI that result from the XML meta-model explicit. Effort should be put into the development of formal procedures for converting one serialisation into another, or into making as many data formats as possible available to a single tool, designed for a specific task, without losing the expressiveness of the original data. The DH community should indeed consider standing-off the idea that trees and graphs are fundamentally in opposition to one another. It should consider them rather as metaphors more helpful for one scholarly editing task than for others.

Bibliography

Allroggen, Gerhard, Markus Bandur, Frank Ziegler, Joachim Veit, Peter Stadler, Eveline Bartlitz, Dagmar Beck, and Solveig Schreiter (eds.), *Carl Maria von Weber-Gesamtausgabe*, 2011–2019 <https://weber-gesamtausgabe.de/>

Andrews, Tara L, and C. Macé, 'Beyond the tree of texts : Building an empirical model of scribal variation through graph analysis of texts and stemmata', *Literary and Linguistic Computing* 28,4 (2013), 504–521

Andrews, Tara L, and Joris van Zundert, *Stemmaweb: Interoperable tools for the investigation of medieval text stemmatology*, 2012 <https://stemmaweb.net/>

Apache Software Foundation, 'Apache TinkerPop', 2015–2019 <http://tinkerpop.apache.org/>

argimenes. A Standoff Properties Editor in JavaScript, 2018–2020 <https://github.com/argimenes/standoff-properties-editor>

BBAW, 'Archiv-Editor', Personendaten-Repositorium. DFG-Projekt, 2011 <http://pdr.bbaw.de/software/ae/>

———, 'ediarum - Digitale Arbeits- und Publikationsumgebung für Editionsvorhaben', 2014–2019 <http://www.bbaw.de/telota/software/ediarum>

———, 'Personendaten-Repositorium', Personendaten-Repositorium. DFG-Projekt, 2009 <https://doi.org/pdr.bbaw.de/>

Bechhofer, Sean, Jeen Broekstra, Stefan Decker, Michael Erdmann, Dieter Fensel, Carole Goble, Frank van Harmelen, et al., 'An informal description of OIL-Core and Standard OIL: a layered proposal for DAML-O', 2000 <http://web.archive.org/web/20090320081902/http://www.ontoknowledge.org:80/oil/downl/dialects.pdf>

Beretta, Francesco, 'A Challenge for Historical Research: Making Data FAIR Using a Collaborative Ontology Management Environment (OntoME)', *Semantic Web Journal*, 2020

<http://semantic-web-journal.net/content/challenge-historical-research-making-data-fair-using-collaborative-ontology-management-0>

Bizer, Chris, and Richard Cyganiak, 'RDF 1.1 TriG', W3C Recommendation, 25. Feb. 2014 <https://www.w3.org/TR/2014/REC-trig-20140225/>

Bradley, John, and Harold Short, 'Texts into Databases: The Evolving Field of New-Style Prosopography', *Literary and Linguistic Computing* 20, Suppl. (2005), 3–24 <https://doi.org/10.1093/llc/fqi022>

brat contributors, 'Brat Standoff Format', brat, 2010–2018 <http://brat.nlplab.org/standoff.html>

Burghartz, Susanna, Sonia Calvi, and Georg Vogeler, 'Urfehdebücher der Stadt Basel – digitale Edition', Digital edition. Urfehdebücher der Stadt Basel – digitale Edition (Graz: Zentrum für Informationsmodellierung 2017) <http://gams.uni-graz.at/context:ufbas>

Carroll, Jeremy J, Christian Bizer, Pat Hayes, and Patrick Stickler, 'Named Graphs, Provenance and Trust', in *Proceedings of the 14th International Conference on World Wide Web - WWW '05*, 613 (Chiba, Japan: ACM Press, 2005) <https://doi.org/10.1145/1060745.1060835>

Castilho, Richard Eckart de, Iryna Gurevych, Anna-Felicitas Hausmann, Jan-Christoph Klie, and Ute Winchenbach, 'INCEpTION', Annotation platform. INCEpTION, 2018–2020 <https://inception-project.github.io/>

Ciula, Arianna, and Øyvind Eide, 'Modelling in Digital Humanities: Signs in Context', *Digital Scholarship in the Humanities* 32, Suppl._1 (2017), i33–46 <https://doi.org/10.1093/llc/fqw045>

Ciula, Arianna, Øyvind Eide, Cristina Marras, and Patrick Sahle, 'Modelling: Thinking in Practice. An Introduction', *Historical Social Research, Supplement* 31 (2018), 7–29 <https://doi.org/10.12759/hsr.suppl.31.2018.7–29>

Cummings, James, 'A World of Difference: Myths and Misconceptions about the TEI', *Digital Scholarship in the Humanities*, 24, suppl_1 (2019), i58–i79 <https://doi.org/10.1093/llc/fqy071>

De Saussure, Ferdinand, *Cours de linguistique générale: zweisprachige Ausgabe französisch-deutsch mit Einleitung, Anmerkungen und Kommentar von Peter Wunderli* (Tübingen: Narr, 2013)

Dekker, Ronald, and David J. Birnbaum, 'It's more than just overlap: Text As Graph', Washington, DC, 2017 <https://doi.org/10.4242/BalisageVol19.Dekker01>

Dekker, Ronald, Elli Bleeker, Bram Buitendijk, Astrid Kulsdom, and David J. Birnbaum, 'TAGML: A Markup Language of Many Dimensions', in *Proceedings of Balisage: The Markup Conference 2018. Balisage Series on Markup Technologies*, vol. 21 (2018) <https://doi.org/10.4242/BalisageVol21.HaentjensDekker01>

Dekker, Ronald, and Gregor Middell, 'Computer-Supported Collation with CollateX', in *Supporting Digital Humanities 2011: Answering the unaskable, Kopenhagen 17–18 November 2011, proceedings ed. by B. Maegaard* (2011)

Deleuze, Gilles, and Félix Guattari, *A Thousand Plateaus. Trans. Brian Massumi. London and New York: Continuum, 2004. Vol. 2 of Capitalism and Schizophrenia. 2 vols. 1972–1980. Trans. of Mille Plateaux* (Paris: Les Editions de Minuit, 1980)

DeRose, Steven, 'What is a diagram, really?', in *Balisage: The Markup Conference 2020, Washington, DC*, 2020 <https://doi.org/10.4242/BalisageVol25.DeRose01>

DFG, 'Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft', November 2015 <https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/foerderkriterien_editionen_literaturwissenschaft.pdf>

Di Iorio, Angelo, Silvio Peroni, and Fabio Vitali, 'Towards Markup Support for Full GODDAGs and beyond: The EARMARK Approach', in *Balisage: The Markup Conference 2009, Montréal, Canada*, 2009 <https://doi.org/10.4242/BalisageVol3.Peroni01>

Digitale Akademie der Mainzer Akademie der Wissenschaften und der Literatur, 'eXGraphs', LOD.Academy, 2019 <https://lod.academy/site/tools/digicademy/exgraphs>

———, 'XTriples', Webservice. XTriples, 2012–2019 <https://xtriples.lod.academy/index.html>

Dolezalova, Lucie, ed., *The Charm of a List: From the Sumerians to Computerised Data Processing* (Newcastle upon Tyne: Cambridge Scholars Publishing, 2009)

Dumont, Stefan, and Martin Fechner, 'Bridging the Gap: Greater Usability for TEI Encoding', *Journal of the Text Encoding Initiative*, Issue 8 (2014) <https://doi.org/10.4000/jtei.1242>

Eades, Peter, 'A Heuristic for Graph Drawing', *Congressus Numerantium* 42,11 (1984), 149–160

Fruchterman, Thomas M.J, and Edward M. Reingold, 'Graph drawing by force-directed placement', *Software: Practice and Experience* 21,11 (1991), 129–1164 <https://doi.org/10.1002/spe.4380211102>

Fuchs, Norbert E, 'Understanding Texts in Attempto Controlled English', in *Proceedings of the 6th International Workshop on Controlled Natural Language (CNL 2018), Maynooth, Ireland*, 2018 <http://attempto.ifi.uzh.ch/site/pubs/papers/cnl2018main_fuchs.pdf>

Fuchs, Norbert E, Kaarel Kaljurand, and Gerold Schneider, 'Attempto Controlled English Meets the Challenges of Knowledge Representation, Reasoning, Interoperability and User Interfaces', in *Proceedings of 19th International Florida Artificial Intelligence Research Society Conference*, 2006 <https://doi.org/10.1.1.541.9039>

Ghelardi, Maurizio, and European Research Advanced Grant Project EUROCORR, 'J.Burckhardt', Digital scholarly edition. J.Burckhardt, 2019 <https://burckhardtsource.org/>

Gibson, James J., 'The Theory of Affordances', in *Perceiving, Acting, and Knowing. Towards an Ecological Psycology, ed. by Robert Shaw and John Bransford* (Hillsdale, NJ u.a, 1977), 67–82 <https://monoskop.org/images/c/c6/Gibson_James_J_1977_1979_The_Theory_of_Affordances.pdf>

Goody, Jack, 'What's in a List?', in *The Domestication of the Savage Mind*, (Cambridge: Cambridge University Press, 1977), 74–111

Hendler, James, Deborah McGuiness, Richard Fikes, and Lynn Andrea Stein, 'DAML-ONT: An Ontology Language for the Semantic Web', in *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*, ed. by Dieter Fensel, Jim Hendler, Henry Lieberman, and Wolfgang Wahlster, 65–93, 2005 <http://www.ksl.stanford.edu/people/dlm/papers/daml-ont-semantic-web.htm>

Knublauch, Holger, and Dimitris Kontokostas, 'Shapes Constraint Language (SHACL)', W3C Recommendation, 20 July 2017 <https://www.w3.org/TR/shacl/>

Huitfeldt, Claus, and C. Michael Sperberg-McQueen, 'TexMECS', 2001 <http://mlcd.blackmesatech.com/mlcd/2003/Papers/texmecs.htm>

'Hypothes.is', 2011–2020 <https://web.hypothes.is/>

Ide, Nancy, and Keith Suderman, 'The Linguistic Annotation Framework: A Standard for Annotation Interchange and Merging', in *Language Resources and Evaluation* 48,3 (2014), 395–418 <https://doi.org/10.1007/s10579–014–9268–1>

'ISO-LAF: Language resource management – Linguistic annotation framework (LAF). ISO 24612:2012' (ISO, 2012)

Ivanovs, Aleksandrs, and Aleksey Varfolomeyev, 'Some Approaches to the Semantic Publication of Charter Corpora. The Case of the Diplomatic Edition of Old Russian Charters', *Digital Diplomatics. The Computer as a Tool for the Diplomatist?*, ed. by Antonella Ambrosio, Sébastien Barret, and Georg Vogeler (Köln: Böhlau, 2014), 149–68

Kamzelak, Roland, 'Digitale Editionen im semantic web. Chancen und Grenzen von Normdaten, FRBR und RDF', in *'Ei, dem alten Herrn zoll' ich Achtung gern', FS Joachim Veit zum 60. Geburtstag*, ed. by Kristina Richts and Peter Stadler (München: Allitera Verlag 2016), 423–35

Kuczera, Andreas, 'Digital Editions beyond XML – Graph-Based Digital Editions', in *HistoInformatics 2016. Poceedings of the 3rd HistoInformatics Workshop on Computational History (HistoInformatics 2016)*, ed. by M. Düring, A. Jatowt, J. Preiser-Kapeller, A. van Den Bosch, CEUR Workshop Proceedings 1632 (Budapest: CEU 2016a), 37–46

———, 'Graphbasierte digitale Editionen', *Mittelalter: Opuscula* (blog), 19. August 2016b <https://mittelalter.hypotheses.org/7994>

Lacy, Lee W., *OWL: Representing Information Using the Web Ontology Language* (Victoria, BC: Trafford Publishing, 2005)

Langefors, Börje, *Theoretical Analysis of Information Systems* (Lund, 1966)

Morbidoni, Christian, and Alessio Piccioli, 'Curating a Document Collection via Crowdsourcing with Pundit 2.0', in *The Semantic Web: ESWC 2015 Satellite Events*, ed. by Fabien Gandon, Christophe Guéret, Serena Villata, John Breslin, Catherine Faron-Zucker, and Antoine Zimmermann, Lecture Notes in Computer Science 9341 (Cham: Springer International Publishing, 2015), 102–6 <https://doi.org/10.1007/978–3-319–25639–9_20>

Neill, Ian, 'The Codex. An Atlas of History', the-codex.net, 2013–2019 <http://web.archive.org/web/20190506014532/http:/the-codex.net/>

Neo4j, 2010–2020 <https://neo4j.com/>

Net7, *Pundit - Pin the Web*, 2015–2020 <http://thepund.it/>

Neumann, Gerald, Fabian Körner, Torsten Roeder, and Niels-Oliver Walkowski, 'Personendaten-Repositorium', *Berlin-Brandenburgische Akademie der Wissenschaften. Jahrbuch* 2010 (2011), 320–26

Nguyen, Vinh, Olivier Bodenreider, and Amit Sheth, 'Don't Like RDF Reification? Making Statements about Statements Using Singleton Property', in *Proceedings of the International World-Wide Web Conference. International WWW Conference*, 2014, 759–70 <https://doi.org/10.1145/2566486.2567973>

Nguyen, Vinh, Olivier Bodenreider, Krishnaprasad Thirunarayan, Gang Fu, Evan Bolton, Núria Queralt Rosinach, Laura I. Furlong, Michel Dumontier, and Amit Sheth, 'On Reasoning

with RDF Statements about Statements using Singleton Property Triples', *arXiv:1509.04513 [cs]*, 2015 <http://arxiv.org/abs/1509.04513>

NIE-INE, 'Machine Interpretable and Interoperable Semantics for Humanities', e-editiones, 2020 <http://e-editiones.ch/>

Norman, Donald A., *The Design of Everyday Things* (New York: Doubleday, 1988)

Odgen, Charles K., and Ivor Armstrong Richards, *The Meaning Of Meaning : A Study of the Influence of Language upon Thought and of the Science of Symbolism* (London and New York, 1923)

Pasin, Michele, and John Bradley, 'Factoid-Based Prosopography and Computer Ontologies: Towards an Integrated Approach', *Literary and Linguistic Computing* 30,1 (1. April 2015), 86–97 <https://doi.org/10.1093/llc/fqt037>

Pelagios Network, Exeter University, Humboldt Institute for Internet and Society, The Open University, and University of London, 'Recogito', Software. Recogito, 2014–2020 <https://recogito.pelagios.org/>

Peroni, Silvio, *LODE - Live Owl Documentation Environment.* Webservice, 2012–2020 <https://essepuntato.it/lode/>

Petris, Marco, Jan Christoph Meister, Jan Horstmann, Janina Jacke, Christian Bruck, Mareike Schumacher, and others, *CATMA. Computer Assisted Text Markup and Analysis*, 2008–2020 <https://catma.de/>

Pichler, Alois and Zöllner-Weber, Amélie, 'Sharing and Debating Wittgenstein by Using an Ontology', *Literary and Linguistic Computing*, 28.4 (2013), 700–707 <https://doi.org/10.1093/llc/fqt049>

Pichler, Alois (ed.), *Wittgenstein Ontology*, The Wittgenstein Archives at the University of Bergen (WAB), 2020 <http://wab.uib.no/cost-a32_philospace/wittgenstein.owl>

Pollin, Christopher, and Georg Vogeler, 'Semantically Enriched Historical Data : Drawing on the Example of the Digital Edition of the ‚Urfehdebucher der Stadt Basel'', 2014, in *WHiSe 2017. Workshop on Humanities in the Semantic Web, Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II) co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22, 2017*, ed. by Alessandro Adamou, Enrico Daga, and Leif Isaksen, CEUR Workshop Proceedings 2014. (Budapest: CEUR, 2017), 27–32

Poupeau, Gautier, 'De l'index nominum à l'ontologie. Comment mettre en lumière les réseaux sociaux dans les corpus historiques numériques?', in *Digital Humanities 2006. The First ADHO International Conference: Conference Abstracts. Université Paris-Sorbonne Paris 2006* (s.l.: ADHO , 2006), 161–64

Quillian, M. Ross, 'Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities', *Behavioral Science* 12,5 (1967), 410–30 <https://doi.org/10.1002/bs.3830120511>

Renear, Allen H, Steven J. DeRose, David G. Durand, and Elli Mylonas, 'What is text, really?', *Journal of Computing in Higher Education* 1,2 (1990), 3–26

Rennau, Hans-Jürgen, 'RDFe – expression-based mapping of XML documents to RDF triples', in *XML Prague 2019. Conference Proceedings*, Prag 2019, 381–404 <http://archive.xmlprague.cz/2019/files/xmlprague-2019-proceedings.pdf#page=393>

Burghartz, Susanna, Jonas Sagelsdorff, and Sonia Calvi, 'Jahrrechnung Stadt Basel 1545/1546',

in *Jahrrechnungen der Stadt Basel 1535 bis 1610 – digital*, ed. by Susanna Burghartz, 2015. <http://gams.uni-graz.at/o:srbas.1545>

Sahle, Patrick, *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung.* Schriften des Instituts für Dokumentologie und Editorik 9 (Norderstedt: Books on Demand, 2013)

Sanderson, Robert, Paolo Ciccarese, and Benjamin Young, 'Web Annotation Data Model', W3C Recommendation, 23 February 2017 <https://www.w3.org/TR/annotation-model/>

Schmidt, Desmond, and Domenico Fiormonte, 'A Fresh Approach to Textual Variation', *Digital Humanities 2006. Conference Abstracts* (s.l.: ADHO, 2006), 193–96

Schmidt, Desmond, 'The Inadequacy of Embedded Markup for Cultural Heritage Texts', *Literary and Linguistic Computing* 25,3 (2010), 37–56 <https://doi.org/10.1093/llc/fqq007>

Schmidt, Desmond, and Robert Colomb, 'A data structure for representing multi-version texts online', *International Journal of Human-Computer Studies* 67,6 (1 June 2009), 497–514 <https://doi.org/10.1016/j.ijhcs.2009.02.001>

Schrade, Torsten, *digicademy/xtriples* (Version 1.4.0). Zenodo, 2019 <http://doi.org/10.5281/zenodo.2604986>

simogih.org. *Le projet symogih.org : un système modulaire de gestion de l'information historique. Platform to store primary data*, 2012–2020 <http://symogih.org/>

Sperberg-McQueen, C. Michael, and Claus Huitfeldt, 'GODDAG: A Data Structure for Overlapping Hierarchies', in *Digital Documents: Systems and Principles, 8th International Conference on Digital on Documents and Electronic Publishing, DDEP 2000, 5th International Workshop on the Principles of Digital Document Processing, PODDP 2000, Munich, Germany, September 13–15, 2000, Revised Papers*, ed. P. King and E.V. Munson, Lecture Notes in Computer Science 2023 (Berlin: Springer, 2000), 139–60 <https://doi.org/10.1007/978–3-540–39916–2_12>

TCF Weblicht, 'The TCF Format', WebLichtWiki, 12 February 2015 <https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format>

TEI Consortium, '<witness>', P5: Guidelines for Electronic Text Encoding and Interchange. TEI <Text Encoding Initiative>, 2020a <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-witness.html>

———, 'Non-Hierarchical Structures', P5: Guidelines for Electronic Text Encoding and Interchange, 2020b <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html>

———, 'TEI P5: Guidelines for Electronic Text Encoding and Interchange', 13 February 2020 <https://web.archive.org/web/20200508073512/https:/www.tei-c.org/release/doc/tei-p5-doc/en/html/>

Tennison, Jeni, 'Overlap, Containment and Dominance', 6 December 2008 <https://www.jenitennison.com/2008/12/06/overlap-containment-and-dominance.html>

Tennison, Jeni, and Wendell Piez, 'The Layered Markup and Annotation Language (LMNL)', Montreal, Canada, 2002 <http://xml.coverpages.org/LMNL-Abstract.html>

Thaller, Manfred, 'kleio', kleio - on the web, 2003–2009 <http://web.archive.org/web/20090509075526/http:/www.hki.uni-koeln.de:80/kleio/>

———, 'What is a text within the Digital Humanities, or some of them, at least?', in *DH2012 -*

Conference Abstracts. Hamburg, 2012 <http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/beyond-embedded-markup>

Thompson, Henry, and Dave McKelvie, 'Hyperlink se-mantics for standoff markup of read-only documents', in *Proceedings of SGML Europe'97* (1997) <http://www.ltg.ed.ac.uk/~ht/sgmleu97.html>

Vallance, Edward, 'The 1723 Oath Rolls in England: An Electronic Finding List', *History Working Papers Project* (2013) <http://www.historyworkingpapers.org/?page_id=373>

Van Zundert, Joris, and Tara L. Andrews, 'Apparatus vs. Graph: New Models and Interfaces for Text', in *Interface Critique*, ed. by F. Hadler and J. Haupt, Kaleidogramme 139 (Berlin: Kulturverlag Kadmos, 2016), 183–205

Verborgh, Ruben, and Jos De Roo, 'Drawing Conclusions from Linked Data on the Web: The EYE Reasoner', *IEEE Software* 32,3 (2015), 23–27 <https://doi.org/10.1109/MS.2015.63>

Vogeler, Georg, 'The 'Assertive Edition', *International Journal of Digital Humanities* 1,2 (1. July 2019), 309–22 <https://doi.org/10.1007/s42803–019–00025–5>

Vogeler, Georg, and Christopher Pollin, 'DEPCHA - Alpha Version', DEPCHA - Digital Edition Publishing Cooperative for Historical Accounts. Alpha-Version, 2018–2020 <http://gams.uni-graz.at/context:depcha>

W3C, 'SKOS Simple Knowledge Organization System Reference. W3C Recommendation', Herausgegeben von Alistair Miles und Sean Bechhofer, 18th August 2009 <https://www.w3.org/TR/2009/REC-skos-reference-20090818/>

W3C OWL Working Group, 'OWL 2 Web Ontology Language Document Overview (Second Edition)', W3C Recommendation, 2012 <https://www.w3.org/TR/owl-overview/>

W3C, 'RDF 1.1 Primer. W3C Working Group Note', W3C Working Group Note, 2014 <https://www.w3.org/TR/rdf11-primer/#section-blank-node>

Wettlaufer, Jörg, 'Der nächste Schritt? : Semantic Web und digitale Editionen', in *Digitale Metamorphose: Digital Humanities und Editionswissenschaft*, ed. by Roland S. Kamzelak and Timo Steyer, Zeitschrift für digitale Geisteswissenschaften Sonderband 2, 2018 <http://zfdg.de/sb002_007>

Zeldes, Amir, Florian Zipser, and Arne Neumann, 'PAULA XML Documentation - Format Version 1.1', 21 January 2013 <http://www.sfb632.uni-potsdam.de/paula.html>

Zeller, Hans, 'Befund und Deutung. Interpretation und Dokumentation als Ziel und Methode der Edition', *Texte und Varianten. Probleme ihrer Edition Und Interpretation*, ed. by Gunter Martens and Hans Zeller (München: Beck, 1971), 45–90

Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, *GAMS - Geisteswissenschaftliches Asset Management System. Publication plattform*, 2014–2020 <http://gams.uni-graz.at/>

Formal Models

Projects and Editions

Appendices

Biographical Notes

Thomas Ahrend (University of Basel, Switzerland – thomas.ahrend@unibas.ch) studied Musicology, Philosophy and Literary Studies in Frankfurt a. M. and Berlin. He received his MA 1996, and his PhD 2005 at Technische Universität Berlin with a dissertation on the instrumental music of Hanns Eisler. 1997–2010 member of the editorial staff of the Hanns Eisler Gesamtausgabe in Berlin. Since September 2010, member of the editorial staff of the Anton Webern Gesamtausgabe at Musikwissenschaftliches Seminar at University of Basel.

Peter Boot (Huygens ING, The Netherlands – peter.boot@huygens.knaw.nl) studied mathematics and Dutch language and literature; he wrote his PhD thesis about annotation in scholarly digital editions and its implications for humanities scholarship. He oversaw the creation of the digital edition of the letters of Vincent van Gogh. He is employed as a senior researcher at the Huygens Institute for the History of the Netherlands where he works, among other things, as a consultant in several edition projects.

Manuel Burghardt (University of Leipzig, Germany – burghardt@informatik.uni-leipzig.de) is head of the Computational Humanities Group at Leipzig University. He is interested in the use of digital tools and computational techniques to explore new modes of doing research in the humanities. His most recent areas of research are Sentiment Analysis in the Humanities, Drametrics, Computational Intertextuality, Computational Analysis of Movies and Series and Music Information Retrieval.

Toby Burrows (University of Oxford, United Kingdom – toby.burrows@oerc.ox.ac.uk) is a Senior Researcher in the Oxford e-Research Centre at the University of Oxford, and a Senior Honorary Research Fellow in the School of Humanities at the University of Western Australia.

Hugh Cayless (Duke University, USA - hugh.cayless@duke.edu) is Senior Digital Humanities Developer at the Duke Collaboratory for Classics Computing. Hugh has over a decade of software engineering expertise in both academic and industrial settings. He also holds a Ph.D. in Classics and a Master's in Information Science. He is one of the founders of the EpiDoc collaborative and currently serves on the Technical Council of the Text Encoding Initiative.

Hans Cools (University of Basel, Switzerland – 1961-2021) had a master degree in medicine and a specialization in orthopaedic surgery and traumatology (Universities of Ghent and Antwerp, Belgium, 1997), a bachelor's degree in physical

therapy, and a standalone degree in informatics (1999). Through various research and project management positions, in both companies and academic institutions, he gained expertise in different aspects of the Semantic Web technologies, focusing particularly on formal data modeling and machine reasoning. Those positions were in internationally collaborative research projects in a biomedical setting, mainly of the 5-7th EU Framework Program. Foremost in these projects were semantic interoperability and reusability of data. Since 2016, he worked in the humanities, as knowledge engineer, ontologist, and Semantic Web technology expert, at the University of Basel, as part of the NIE-INE project, which highlights scholarly editing. He (co-)published several articles, and gave workshops on the implementation of Semantic Web technologies in biomedicine and the humanities. He passed away in April 2021.

Francesca Giovannetti (University of Bologna, Italy – francesc.giovannett6@unibo.it) is a second-year PhD student in Digital Humanities at the Department of Classical Philology and Italian Studies, University of Bologna. She received an MA in Digital Humanities from King's College London and a second cycle degree in Digital Humanities and Digital Knowledge from the University of Bologna. She is interested in combining digital scholarly editing with semantic web technologies and in the use of digital technologies in education.

Matthew Holford (University of Oxford, United Kingdom – matthew.holford@bodleian.ox.ac.uk) is Tolkien Curator of Medieval Manuscripts at the Bodleian Library, University of Oxford.

Marijn Koolen (Royal Netherlands Academy of Arts and Sciences - Humanities Cluster, The Netherlands – marijn.koolen@gmail.coml) studied artificial intelligence and wrote his PhD thesis on using hyperlinks in information retrieval algorithms. He has worked on scholarly annotation for digital humanities research and on annotation-related information behaviour and information systems. He works as a researcher and developer at the Humanities Cluster of the Royal Netherlands Academy of Arts and Sciences, where he leads a project on developing annotation support within the *CLARIAH research infrastructure* project.

David Lewis (University of Oxford, United Kingdom – david.lewis@oerc.ox.ac.uk) is a Research Associate in the Oxford e-Research Centre at the University of Oxford.

Andrew Morrison (University of Oxford, United Kingdom – andrew.morrison@bodleian.ox.ac.uk) is a Software Engineer in the Bodleian Digital Library Systems and Services, Bodleian Library, University of Oxford.

Stefan Münnich (University of Basel, Switzerland – stefan.muennich@unibas.ch) studied musicology and communication science at the Technische Universität Berlin, MA 2011 with a thesis on cantional setting in Heinrich Schütz's Becker-Psalter. 2012 research assistant, 2013–2015 research associate of the Felix Mendelssohn Bartholdy. Sämtliche Briefe edition at University of Leipzig (co-editor of vols. 9 & 12). Since October 2015 research associate of the Anton Webern Gesamtausgabe, Basel; received his Doctorate degree in 2020 at the department of musicology at the University of Basel with a dissertation about music notation and its codes.

Iian Neill (Digital Academy of the Academy of Sciences and Literature, University of Mainz - Iian.Neill@adwmainz.de) is a visiting researcher at the Digital Academy of the Academy of Sciences and Literature Department at the University of Mainz, Germany. He is the creator of Codex, a text annotation environment which uses standoff property annotation to generate entities in a graph meta-model. Codex is currently being used to produce a digital edition of the epistles of Hildegard von Bingen at the Digital Academy in Mainz.

Roberta Padlina (University of Basel, Switzerland – roberta.padlina@unibas.ch) studied medieval philosophy at the University of Fribourg, Switzerland, obtaining a doctoral degree in June 2020. She has twelve years of professional experience in the field of Digital Humanities, thanks to which she has been able to work closely with different actors involved in the online publication of open access research. Roberta has worked for several years for e-codices –Virtual Library of Manuscripts in Switzerland and currently coordinates the National Infrastructure for Editions (NIE-INE) project. Roberta's main focus is on the opportunities and challenges that the digital shift poses for traditional education and research institutions, including developing semantic web strategies for scholarly publications and cultural goods.

Kevin Page (University of Oxford, United Kingdom – kevin.page@oerc.ox.ac.uk) is a Senior Researcher in the Oxford e-Research Centre and Associate Member of Faculty in the Department of Engineering in the University of Oxford.

Miller C. Prosser (University of Chicago, USA – m-prosser@uchicago.edu) earned his Ph.D. in Northwest Semitic Philology from the University of Chicago. His academic interests include the social and economic structure of Late Bronze Age Ras Shamra-Ugarit and the use of computational methods for philological and archaeological research. Miller is the Associate Director of the Digital Studies MA program at the University of Chicago where he teaches courses on Data Management and Data Publication for the Humanities. He also works as a

researcher at the OCHRE Data Service of the Oriental Institute of the University of Chicago where he consults with and supports research projects using the Online Cultural and Historical Research Environment (OCHRE). He has also worked as a tablet photographer for the Mission de Ras Shamra (Ugarit) and the Persepolis Fortification Archive Project, employing advanced digital photographic methods such as reflectance transformation imaging, photogrammetry, and high-resolution digital scanning.

Matteo Romanello (Université de Lausanne, Switzerland - matteo.romanello@unil.ch) is Ambizione SNF Lecturer at the University of Lausanne, where he conducts a project on the commentary tradition of Sophocles' Ajax. Matteo is a Classicist and a Digital Humanities specialist with expertise in various areas of the Humanities, including archaeology and history. After obtaining his PhD from King's College London, he worked as a research scientist at EPFL's DHLAB on the Linked Books and Impresso projects, before moving to his current position. He was also teaching fellow at the University of Rostock, researcher at the German Archaeological Institute, and visiting research scholar at Tufts University.

Sandra Schloen (University of Chicago, USA – sschloen@uchicago.edu) is the Manager of the OCHRE Data Service at the Oriental Institute of the University of Chicago, and is the co-designer and developer of the Online Cultural and Historical Research Environment (OCHRE). Trained in computer science and mathematics (B.Sc. University of Toronto; M.Ed. Harvard University), Sandra has spent over 30 years working with technology as a systems analyst, technical trainer, and software developer. A long association with colleagues in the academic community has enabled her to develop a specialty in solving problems in the Digital Humanities where challenges of data capture, data representation and data management abound. Specifically, she has served extensively as a database manager for several archaeological projects in Israel and Turkey, and supports a wide range of research projects at the Oriental Institute and at other universities.

Desmond Schmidt (University of Bologna - desmond.allan.schmidt@gmail.com) has a background in classical Greek philology, information security and eResearch. He has worked on several scholarly edition projects, including the Vienna Wittgenstein Edition (1990–2001), Digital Variants (2004–2008), the Australian Electronic Scholarly Editions project (2012–2013), the Charles Harpur Critical Archive (2014-) and a pilot edition of Gianfrano Leopardi's Idilli (2018-). He currently works on developing practical web-based tools for making, visualising and publishing digital scholarly editions.

Colin Sippl (University of Regensburg, Germany – colin.sippl@ur.de) is currently a project employee at the University Library of Regensburg. Since 2017, he has been working on extending the open access services of the Electronic Journals Library (EZB). More recently, he has started developing and setting up a digital repository for literature, artefacts and experiments relating to the early life sciences based on the Invenio framework. He specialised in textual data mining and the development of media services in the institutional domain.

Elena Spadini (University of Lausanne - elena.spadini@unil.ch) is a postdoctoral researcher at the University of Lausanne. She holds a Ph.D. in Romance Philology from the University of Rome Sapienza (2016) and a M.A. in Digital Humanities from the École nationale des chartes (2014). She was a Marie Curie fellow in the IT Network DiXiT and co-directed the related volume Advances in Digital Scholarly Editing (Sidestone Press, 2017). She published in international journals and taught specialized courses in various European countries in the field of Digital Philology.

Francesca Tomasi (University of Bologna - francesca.tomasi@unibo.it) is associate professor in Archival Science, Bibliography and Librarianship at the University of Bologna (Italy). Her research is mostly devoted to digital cultural heritage, with a special attention to documentary digital edition, and a focus on knowledge organization methods in archives and libraries. She is member of different scientific committees of both associations and journals. In particular, she is President of the Library of the School of Humanities in the University of Bologna (BDU - Biblioteca di Discipline Umanistiche), Director of the international second cycle degree in Digital Humanities and Digital Knowledge (DHDK), President of the Italian Association of Digital Humanities (AIUCD – Associazione per l'Informatica Umanistica e la Cultura Digitale), and co-head of the Digital Humanities Advanced Research Center (/DH.ARC). She wrote about 100 papers and 4 monographs related to DH topics. She is editor and scientific director of several digital scholarly environments.

Athanasios Velios (University of the Arts London, United Kingdom – a.velios@arts.ac.uk) is Reader in Documentation at the University of the Arts London.

Georg Vogeler (University of Graz - georg.vogeler@uni-graz.at) is professor for Digital Humanities at the University of Graz and scientific director of the Austrian Center for Digital Humanities and Cultural Heritage at the Austrian Academy of Sciences. He is a trained historian (Historical Auxiliary Sciences). He spent several years in Italy (Lecce, Venice). In 2011, he became member of faculty at the Centre for Information Modelling at Graz University, where he was nominated

full professor for Digital Humanities in 2016 and head of department in 2019. His research interests lie in late medieval and early modern administrative records, diplomatics (digital and non digital), digital scholarly editing and the history of Frederic II of Hohenstaufen (1194–1250). He was and is part in several national and international research projects related to his research interests.

Christian Wolff (University of Regensburg, Germany – christian.wolff@ur.de) has been Professor of Media Informatics at the Institute for Information and Media, Language and Culture at the University of Regensburg since 2003. He holds a PhD in information science and is a habilitated computer scientist. His research interests include: human-computer interaction, multimedia and web-based information systems, (multimedia) software engineering and information retrieval (in particular information literacy and social media).