Journal of the Text Encoding Initiative

Issue 8 | December 2014 - December 2015
Selected Papers from the 2013 TEI Conference

Ontologies, Data Modeling, and TEI

Øyvind Eide

Electronic version
URL: http://journals.openedition.org/jtei/1191
DOI: 10.4000/jtei.1191
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Øyvind Eide, « Ontologies, Data Modeling, and TEI », *Journal of the Text Encoding Initiative* [Online], Issue 8 | December 2014 - December 2015, Online since 09 April 2015, connection on 19 April 2019. URL : http://journals.openedition.org/jtei/1191 ; DOI : 10.4000/jtei.1191

Ontologies, Data Modeling, and TEI

Øyvind Eide

1. Introduction

1 In philosophy, *ontology* has for at least 2,500 years denoted the study of being. Computer science *ontologies*, usually in the plural, are different from the philosophical concept of ontology. Computer science ontologies refer to shared conceptualizations expressed in formal languages (Gruber 2009) and have been a topic of study for some thirty years, initially connected to the artificial intelligence community. They have not been of much importance in digital humanities until the last ten to fifteen years, but are now gaining momentum as the semantic web develops.

2 In this paper I will discuss ontologies in the context of the Text Encoding Initiative and based on the computer science tradition.[1] However, although computer science ontologies are different from philosophical ontology, the two are not totally disconnected (Zúñiga 2001), and therefore some remarks will be made on relations with philosophy as well. The focus will be on how meaning can be established in computer-based modeling. Three broad areas will be described. One is the establishment of meaning through working with and interpreting documents seen as sources for scholarly work, be they primary or secondary. Another area is the establishment of meaning through the development of models, including ontologies. A third area of meaning production

with particular reference to TEI happens through linking between TEI documents and external ontologies. How such linking can be done and what it may imply for the meaning and usability of TEI documents is an important focus of this article.

3 TEI represents a shared conceptualization of what exists in the domains relevant to text encoding. It can be expressed in formal models, but it is questionable whether TEI can be seen as an ontology in the computer science sense. According to the classification by Guarino, Oberle, and Staab (2009, 12–13), XML schemas are typically not expressive enough for the formality we need for ontologies. However, the level of language formality forms a continuum, and it is difficult to draw a strict line where the formal starts. So for instance, some parts of the TEI scheme, such as the systems used to encode dictionaries, bibliographies, and representations of persons, places, and events, may be closer to an ontology than other, less formalized parts of the standard (Ore and Eide 2009).

2. P4 to P5[2]

4 In TEI P4, as in previous versions of TEI, names of places and people, as well as other strings referring to places, people, and organizations, could be encoded using elements such as <rs> and <name>. These elements were used to encode names and other referring strings as textual features. P4 included no construct[3] to encode information about real world or fictitious entities that the strings in the text referred to:

> It should be noted however that no provision is made by the present tag set for the representation of the abstract structures, or virtual objects to which names or dates may be said to refer. In simple terms, where the core tag set allows one to represent a name, this additional tag set allows one to represent a personal name, but neither provides for the direct representation of a person.
>
> (TEI Consortium 2001, ch. 20)

5 In P5 this was changed. Real or fictitious entities pointed to by referring strings in the text were now included in the standard so that in addition to the referring strings, which could still be encoded as in P4, the external entities could be represented as well:

> This module also provides elements for the representation of information about the person, place, or organization to which a given name is understood to refer and to represent the name itself, independently of its application. In simple terms, where the core module

> allows one simply to represent that a given piece of text is a name, this module allows one further to represent a personal name, to represent the person being named, and to represent the canonical name being used. A similar range is provided for names of places and organizations. The main intended applications for this module are in biographical, historical, or geographical data systems such as gazetteers and biographical databases, where these are to be integrated with encoded texts.
>
> (TEI Consortium 2013, ch. 13)

6 This means that while P4 represented only textual occurrences, P5 goes beyond that limitation. In P5, names can be connected to a representation of the named object, which is typically placed in the header of the TEI document (Wittern 2009). The inclusion of the new elements provided one standard way to create such representations, which was needed by many text encoders. We will see later how this header information represents one way among others with which to link the referring strings as they are encoded in the text to external ontologies. As other methods do exist, the inclusion of the new elements in P5 was not strictly necessary for interoperability, but did facilitate a standard way of creating the interconnections.

3. Two Ways of Modeling

7 There are no ontologies without models — an ontology, after all, represents a model of a world or of a certain corner of it. The discussion in the paper will focus on active engagement with models, that is, on how meaning is generated and anchored when ontologies and other models are developed and used. For TEI specifically, the development of the standard was based on ontological assumptions regarding relevant textual sources (e.g., their structure and components) in the philosophical sense. Further development, such as creating the new elements described in the previous section, is based on similar considerations. Using TEI for encoding texts also involves the study of the source material and how it can be related to the standard as a whole, which has some similarities with the development of the standard itself. Here, however, I will focus on the differences between these two types of work.

8 I will distinguish between two different, although overlapping, ways of modeling. First, one may use already existing models for data integration. An example of this is the task of integrating data from several different libraries and archives in order to create a common data warehouse in which

the detailed classifications from each of the databases are preserved. In the process, one will want to use a common ontology for the cultural heritage sector, for instance, FRBRoo (Bekiari 2013). One must develop a thorough understanding of the sources, be they TEI encoded texts or other forms, as well as of the target ontology—one will develop new knowledge.

9 The task is intellectually demanding and those engaged in it will learn new things along the way. Still, the formal specification of the corner of the world they are working towards is already defined in the standard. Only in a limited number of cases will there be a need to develop extensions to the model. Once the job is done, inferences made using the ontology-based data warehouse can be used to understand the sources and what they document even better. Yet, despite all the new knowledge acquired, the process is still mostly restricted to the use of what is already there.

10 The second way of working with models is to create an ontology or another formal model through studying a domain of interest. In this case, a group of people will analyze what exists in the domain and how one can establish classes which are related to each other. This may, for instance, be in order to understand works of fiction, as in the development of the OntoMedia ontology,[4] which is used to describe the semantic content of media expressions. It can also be based on long traditions of collection management in analog as well as digital form, as in the development of CIDOC-CRM (Crofts et al. 2011) in the museum community. Although one will often study data from existing information systems as part of the process, the main goals of such studies are not mappings in themselves, but rather to understand and learn from previous modeling exercises in the area of interest.

11 The historical and current development of TEI as a standard can be seen in this context. The domain of TEI has no clear borders, but the focus is on textual material in the humanities and cultural history. In order to develop a model of this specific corner of the world, one has to analyze what exists and how the classes of things are related to each other. This is a process in which domain specialists and people trained in the creation of data models must work together, as the history of TEI demonstrates.[5]

12 When applying either of the two methods of modeling, knowledge is gained through the process as well as in the study and use of the end products; one can learn from modeling as well as from models, from the process of creating an ontology as well as from the use of already existing ones. Getting to know a description of a modeling standard rarely gives the same level of understanding

as actively engaging with the model as a creator or a user. Reading the TEI Guidelines is a good way of getting an overview of the standard, but it is hard to understand it at a deeper level without using it in practical work, and among the best TEI experts are those who have taken part in creating the standard.

13 There is no clear line between the two approaches to modeling, and they often use similar methods in practice. They both have products as the end goal, and new knowledge is created in the process. Some of this new knowledge is expressed in the end products. For example, working to understand better what is important for a concept such as "person" in the domain under study will result in new knowledge. This knowledge will be shared by the parties involved in the modeling exercise and may be expressed in the end product. However, there is a stronger pressure towards expressing such new knowledge clearly when a data standard is created than when a mapping is created. In the latter case the boxes are already made and one has to find the right fit between one's data and those boxes, typically assuming that the ones who made them worked thoroughly on the matter. This is useful in many cases, but one is less inclined to question what is there and runs the risk of not seeing what is outside the standard (Zafrin 2007, 66). In the case of creating a standard, one has to build the boxes, establishing both the entities to be used, the relationships between them, and how to name and describe both entities, relationships, and the general principles behind the standard. Everything is open to question, and in the end a new world is built.

14 These two methods are presented as distinct categories for the sake of analysis. They are, however, prototypical, and there is no strict line between them. Middle positions can be found when one takes an existing ontology and extends it, or uses terms taken from multiple ontologies in combination. Such border cases notwithstanding, the distinction between using and creating models is clear in most cases.

15 In addition to the distinction between using models and creating models, we also have a distinction between modeling for production and modeling for understanding. Modeling for understanding is here seen as creating models with the sole or main purpose of learning new things through the modeling activity. The model in itself is not a main goal of the work. In modeling for production a version of the modeled entity is the main goal of the work. So even if the use of a model for production (for instance using TEI to encode a dictionary) is a research activity which gives new understanding, the main purpose is still the end product. The same was the case when the

dictionary module of the TEI Guidelines itself was created. It gave new understanding, but the main purpose was to create the end product—the dictionary module. A well known example of modeling for production is the project Henrik Ibsen's writings, where modeling and TEI encoding were used to create a series of printed volumes.[6] An example of the use of modeling for understanding would be be to analyze a poem through iterative rounds of TEI encoding, finding and investigating new problems along the way. In this process, one could develop an ontology based on one's growing understanding of the poem.[7] In table 1 these four types of modeling are presented as a table which expresses a model of modeling. I must repeat here that the distinctions in the table are not clear-cut, but rather an analytical tool to explore the complex and many-faceted activity of modeling.

Table 1. A model of modeling.

Why modeling?	**Using a model**	**Creating a model**
Modeling for production	Using TEI to encode and publish a document	Creating a module in the TEI Guidelines
Modeling for understanding	Using TEI encoding as a research method	Creating an ontology as part of a research effort

16 A well-established distinction in modeling theory in digital humanities and beyond is the one between "models for" and "models of." McCarty, following Geertz, distinguishes between "a denotative 'model *of*,' such as a grammar describing the features of a language, and an exemplary 'model *for*,' such as an architectural plan" (McCarty 2005, 24).[8] This analytical distinction is not clear cut, and it "also reaches its vanishing point in the convergent purposes of modeling; the model of exists to tell us what we do not know, the model for to give us what we do not yet have. Models realize" (ibid.). Although not clear-cut, the distinction still has some consequences of relevance for table 1 above. Using a model for production and creating a model for understanding both have a tendency towards creating models of, whereas creating a model for production and using a model for understanding both have a tendency towards creating models for.

17 To return to the examples used above: a dictionary is encoded in TEI as a model of a print dictionary. In creating it, one makes a “simple” (but usually labor-intensive) mapping. Through this process one might learn something new through making explicit certain components of the dictionary which were not apparent in the print version. In creating an ontology of a certain interpretive reading of a poem, one also creates a model of, but one of a different kind. The distance between the object being modeled and the model is much larger; one can learn more from the process.

18 Creating a TEI module as tool for encoding dictionaries is modeling for in that it is not representing a specific document but provides a toolkit one can apply to multiple texts.[9] In the case of using a model for understanding, one is aiming at discovering something new, so one is enabling a model for. Both of these are used as models for, but the former is more practically oriented, creating a tool for encoding, while the latter is more discovery oriented, using a model to elicit new knowledge.

19 As the examples show, the tendencies towards a match with the modeling of/modeling for distinction are weak and will depend on context. This may indicate that the distinction between models for and models of does not match the distinctions in table 1 very well, and that they operate at another level. A closer study of their relationship is beyond the scope of this article, but what this unclear relationship does show is the limitations of any classification such as the one proposed in table 1 and for that matter of any model.

20 Another distinction, established recently by Jannidis and Flanders, is the one between altruistic and egoistic modelers. This is a distinction which fits my classification quite well. They describe models created by altruistic modelers as the ones that “serve as an interchange format for some types of users and user communities where data is typically being created and modeled with someone else’s needs in mind” (Jannidis and Flanders 2013, 238), which corresponds nicely to my row labeled “modeling for production.” They describe models created by egoistic modelers as having the function “to express specific research ideas in cases where data is being created to support the creator’s own research needs” (ibid.), which is similar to my row labeled “modeling for understanding.”

21 According to Jannidis and Flanders, these two ways of modeling lead to very different modeling practices. If this is true—and I believe it is—it may indicate that the distinction just discussed is more important than the distinction between creating and using models. This points towards the following hypothesis: while the distinction between making or using a model is significant

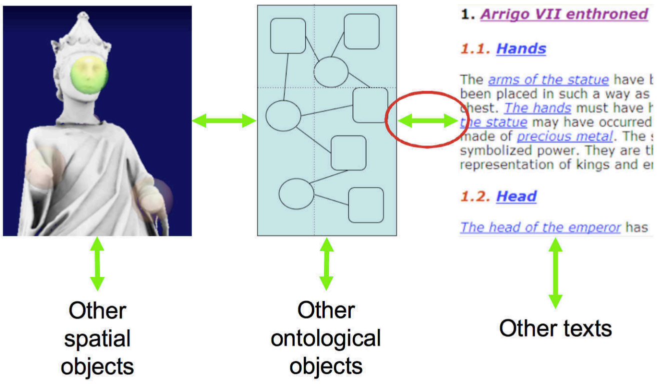
as to how deeply one engages with the material and how well one learns the model, the distinction between modeling for production and modeling for understanding has more significant consequences for how one engages with the model: the broad scope with integration of different views on the one hand, and the narrow focus of specific theoretical assumptions and research interests on the other.

22 I find it necessary here to be explicit about what was implicit in the previous paragraphs: the model I present in table 1 is tentative and should be developed further. This is an area for future research connected not only to the development and use of TEI, but also to better understanding of modeling as a digital humanities practice and how it relates to similar practices in cultural heritage (Ciula and Eide 2014). This must go hand in hand with theoretical work on modeling in digital humanities, as called for by Jannidis and Flanders (2013, 237).

4. Interconnections

23 When a TEI encoding of a text is established, it is the result of a process which includes modeling, even if the involved parties may not be explicitly aware of it. In this section I will discuss links between TEI and external ontologies used in the cultural heritage sector, using CIDOC-CRM as my main example. This will serve two purposes. First, it will give a better understanding of how the semantics of TEI documents are established and work. Are there differences between the ways one creates links between TEI and external ontologies which have consequences for the way the semantics of elements are established and used, comparable to the differences in types of modeling discussed in the previous section? Second, it is not only individual projects or specific editions created in TEI that have a need for integrating TEI documents with external ontologies. The need to integrate resources is shared by the cultural heritage and humanities community at large, where it has become a necessary addition to modeling. While integration with external standards is useful both in modeling for production and for understanding, such integration takes many different forms, as we will see.

Figure 1. Interconnected cultural heritage, showing three examples of digital cultural heritage information: 3D models of artifacts, ontologies, and TEI-encoded texts. The picture to the left is taken from the Arrigo showcase (Havemann 2009).


1. Arrigo VII enthroned
1.1. Hands
The arms of the statue have b
been placed in such a way as
chest. The hands must have h
the statue may have occurred
made of precious metal. The s
symbolized power. They are tl
representation of kings and ei
1.2. Head
The head of the emperor has
Other spatial objects
Other ontological objects
Other texts

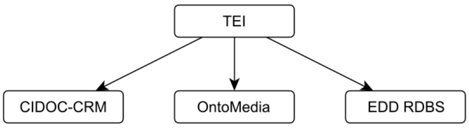
24 Whether the dream of a semantic web for cultural heritage, as expressed in a simplified way in figure 1, is ever to be established will depend on better understanding of how the integration between TEI and other standards such as CIDOC-CRM works. In this larger picture, the current paper is located as the encircled arrow shown in figure 1: in the link between texts and ontologies. Studying ways of linking them together can create practical basic understanding on which to build a theoretical framework. Such links are created when we use models, but the mechanisms for linking must be established when the models are created in the first place. This may be obvious in the creation of standards such as TEI or CIDOC-CRM, but it is also needed in ad hoc models created in research projects. In the latter case the need may not be equally obvious, but the necessity of providing links from a research argument to the sources on which it is based is crucial in the development of more open and reproducible research. Such linking is dependent on semantically well-defined points of entry in TEI for external systems to point to, and similarly on points of exit to be used to link TEI to other resources.[10]

25 How does this look from the other side? An ontology may be used to make statements of contradictory facts at different levels. This will often be related to different interpretations of source material. One example is various paper based sources for classifications of objects in a museum. What in 1880 was seen as pair of Old Norse skis could in 1980 be seen as a pair of Sami skis. The two classifications can be recorded in different protocols or index cards and may be based on different interpretations of the objects as well as of written sources about those objects. Although

one of them will be chosen as the official classification presented by the museum, both should still be recorded in the museum information system in order to track the intellectual history of the field of study.

26 While an ontology is a model of a domain, an individual mapping to an ontology will be based on specific sources. Links from ontologies to their sources are needed in order to ensure scholarly reproducibility. It is not enough that the references exist. It is also important to make the links easily accessible for human readers and computer algorithms alike. Thus, the links must be as precise as possible, point to the source at the highest possible level of detail, and must be formalized in a standardized way. This is to a large extent a matter of practical implementation, but an important one.

Figure 2. Integration seen from TEI. EDD RDBS is a local cultural heritage database system at the University of Oslo.



27 In the situation visualized in figure 2, TEI is the formalism in focus, and the rest of the world, meaning other formalisms and database systems, are seen as children that one needs to link to, based on a linking model which is coherent and usable from the TEI side. In short, it shows how the world of information integration may appear from the perspective of TEI. We will now examine how interconnections can be made in such an interlinked system. Where is the link to the external model to be found in the markup of the text? How is it formalized? It will be shown how, still from the perspective of TEI, we can provide mechanisms for well-defined links back to the encoded texts. We will also examine to what degree this will work differently for links from and to models for production as opposed to models for understanding.

28 Based on practical experience and a close study of chapter 13 of TEI P5, I have established a number of interconnection strategies for elements such as referring strings and names. These are all similar in the sense that they are using references to external targets, in line with normal linked data methods. The links back to the TEI documents are provided through `@xml:id` attributes which,

together with stable identifiers for the individual documents, also work as linked data targets. Any RDF triplet can be encoded in TEI using the `<relation>` element. Thus, the RDF namespace is not strictly needed to make such statements.

Figure 3. Method A: Elements in the body of the TEI document point to an external ontology. Reverse: Links from the external ontology to elements in the body of the TEI document.

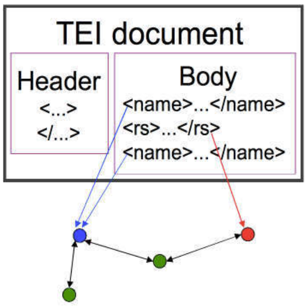


Figure 4. Method B: Elements in the body of the TEI document point to elements expressed in another namespace in the header; links from the header elements point to an external ontology. Reverse: Links from the external ontology to elements in another namespace in the header of the TEI document.

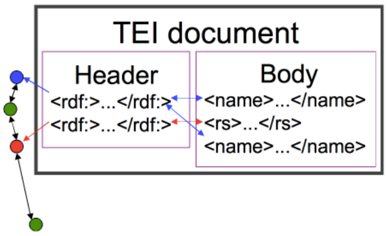


Figure 5. Method C: Elements in the body of the TEI document point to TEI elements in the header; links from the header elements point to an external ontology. Reverse: links from the external ontology to TEI elements in the header of the TEI document.

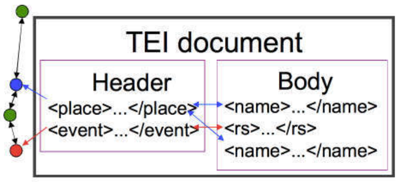


Table 2. The six ways to link between TEI and external ontologies.

	Links from TEI	**Links to TEI**
TEI in body (method A)	Elements in the body of the TEI document point to an external ontology.	Links from external ontology to elements in the body of the TEI document.
Non-TEI in header (method B)	Elements in the body of the TEI document point to elements in another namespace in the header. Links from the header elements point to an external ontology.	Links from external ontology point to elements in another namespace in the header of the TEI document.
TEI in header (method C)	Elements in the body of the TEI document point to TEI elements in the header. Links from the header elements point to an external ontology.	Links from ontology point to TEI elements in the header of the TEI document.

29 As figures 3, 4, and 5 indicate,[11] there are six different ways identified for links between TEI and external ontologies, shown together in table 2; three types of links from TEI to external formalisms and three types of return links.[12] They are all based on a common methodology and on the same basic linked-data concepts. However, the differences we have identified have certain implications for how the links are used as part of modeling and how integration works in practice. The natural center of gravity in the linked system will reside in different places. These differences have significant theoretical consequences.

30 Method C uses TEI elements. Thus, there is one and only one element available to represent a person, namely the <person> element. The same goes for other elements such as <place> and <event>. This means that the use of this method ensures that anyone aiming to harvest or link to such elements in TEI headers[13] will have one set of elements with well-defined semantics to connect to.

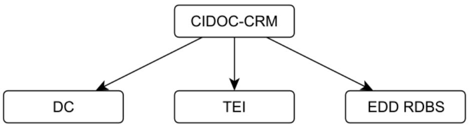
31 The same can to a certain degree be said of method A: as no information in the TEI header is used to represent the entities that the referring strings refer to, the linking from the external system must be done directly to the <rs> element and to the various elements used for names in the body of the TEI document. Thus, the elements are standardized and have well-defined semantics. The problem with this method, compared to method C, is the potentially high number of link targets in the TEI document because links must be made to every occurrence of a reference in the text, rather than to every external entity referred to by one or more referring strings. Links must be established to each particular reference rather than to each particular referred object.

32 Method B is different from the other two in that it just states that the representation in the TEI header is expressed in an external namespace. This can be any namespace in which entities such as persons, places, or events are modeled. A number of potential namespaces and different elements exist, each with potentially different meanings. Examples include the CIDOC-CRM as it is expressed in the Erlangen ecrm namespace,[14] FOAF,[15] and TEI. Semantic interoperability must therefore be ensured on a case-by-case basis. For instance, CIDOC-CRM includes only real-world persons in the definition of a person, whereas other models, such as TEI and FOAF, include fictitious and mythical persons. In FOAF all persons are agents, which is a restriction not found in TEI. Thus, all three have something they call "person" but the extensions[16] are not identical.

33 The semantic openness of the latter method puts different demands on models for production than on models for understanding. In a model for production, the openness should as far as possible be restricted through the establishment of well-defined comparisons between the meanings, at least stating the differences. In a model for understanding, on the other hand, this semantic openness may be used as an important part of the modeling research, making clearer to the researcher any differences detected in attempts to formalize the source material.

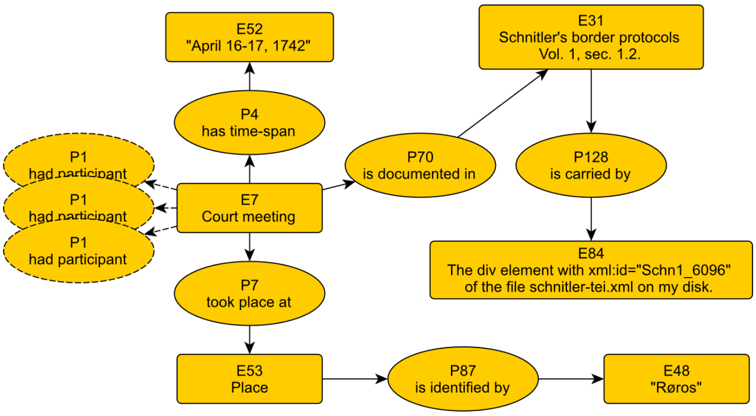
34 This semantic openness is complex, however. When a group of TEI documents is encoded, method C will ensure the same semantic coverage of the elements for all documents. The openness is in the link to the external ontology. In method A, no semantic commitment is made in the TEI documents. Method B injects the semantic openness into the TEI header, so to speak. This may be desirable in certain kinds of modeling for understanding. It is a question of where to establish the semantic openness: in the TEI header (B), in the links between TEI and other formalisms (C), or in the other formalisms (A). In modeling for production the choice should be informed by the question of where in the system at large semantic mapping makes the most sense. In modeling for understanding it should be informed by the choice of the core area of the experimental modeling, if such a core exists. What is the most elegant solution will depend on the aims of the encoder—simply speaking, on whether TEI or another formalism is in focus. The world will indeed look different depending on which formalism is in focus.

Figure 6. Integration seen from CIDOC-CRM. DC is Dublin Core, EDD RDBS is a local cultural heritage database system at the University of Oslo.



35 How would this look if we saw it from the perspective of CIDOC-CRM instead, as in figure 6? A TEI document, or a fragment thereof, can be modeled as a CIDOC-CRM conceptual object. This can be used to explicitly document elements of the CIDOC-CRM model, as we see in figure 7. In this example, an event is documented in CIDOC-CRM and the link is made directly to a TEI `<div>` element without using any of the TEI semantic elements.

Figure 7. A CIDOC-CRM model with a link to a specific element in a TEI document as an E84 Information Carrier entity.



36 Another solution would be to use a referring string typed as representing an event: `<rs type="event">`. But seen from an external formalism, this does not really matter. In many cases it is even good to be able to link to a document without changing its markup, so if references to events are not marked up in the text, linking to another element, even if it lacks the relevant semantic baggage, may be a good choice. Using mechanisms such as XPointer, one can even link to details within the text of an encoded element.

37 Another significant point is that seen from the perspective of an external formalism, linking to a representation of the event itself in the TEI header (an `<event>` element) may not be a good solution, because the relevant textual documentation may be one specific string, not any string in the TEI document referring to the event. The link goes to a particular use of a referring string (an `<rs type="event">` in the body of the TEI document) rather than to a representation of the entity referred to, and it does so on purpose. This may, for example, be necessary in order to identify an occurrence of a name in the dedication of a poem or in the list of witnesses at the end of a legal act.

38 So in this interconnected world everyone can be at the center of their own universe, and where you stand with respect to which linking method is best depends on where you sit; that is, which linking method you prefer depends on the specificities of your task at hand and your relationship to other projects or standards. I see two important ways to cope with this potentially complex situation from the perspective of the TEI community. One is to establish canonical examples of

linking out from and into TEI documents. The use of the `<equiv>` element as a tool for specifying semantic similarity is an important part of this, making it easier to establish common structures for linking between TEI and other formalisms, leading to further interoperability. The other is to make sure that TEI documents have sockets for linking to. While XPointers can be used for advanced references to any TEI document, an easier and more robust method would be to have `@xml:id` attributes available on all elements of published TEI documents, along with stable identifiers for the documents themselves. This gives well-defined linking targets, in line with best practice in museum documentation.[17] These practices may be more crucial in modeling for production than in modeling for understanding. However, keeping the semantic room open, as is often a goal in modeling for understanding, should not lead to ill-defined linking mechanisms.

5. How Wide Do You Want Your Straitjacket?

39 Standards for digital humanities and cultural heritage information will continue to be developed, and even if we develop common concepts for e.g., people and places, those concepts will still be developed in the contexts of those separate standards. A person referred to in a work of fiction is significantly different from a person referred to in a museum documentation system, and the only way of coping with such differences is to establish mappings. Mappings do not necessarily have to be done at a document-by-document level, but there are limits to how general this level can be. While each separate TEI document does not have to be mapped to CIDOC-CRM, one general mapping for all TEI documents will not work either. The solution lies somewhere in between.

40 Such mapping can be used to close or at least reduce the semantic openness discussed above. While such openness is generally accepted as inherent in humanities and culture heritage data, we see a clear difference between modeling for production and modeling for understanding. In modeling for production, one would typically want to clarify the relationships as far as possible, trying to reduce openness in order to make more coherent models. In modeling for understanding, such openness may be an important element of the research: far from being seen as a problem to be solved, it may be central to the experimental modeling. Computer science tends to be solution-oriented in its approach, which is in line with what Mahr (2009) calls the leading question of computer science.[18] Thus, modeling for production is closer to traditional computer science, and modeling for understanding closer to traditional research in the humanities.

41 TEI is a system commonly used for text encoding, where documents are directly created in TEI XML. This is, however, not its only possible use. TEI is also used as an export format for data produced in other systems, such as databases, which can be serialized as TEI XML documents. Dictionaries are a typical example. Encoding can be a step on the way towards typesetting the dictionary, but also works well as a solution to semantic problems connected to long-term preservation. Whether TEI was the production format or not, other systems will need access to TEI-encoded data, such as TEI header data for importing into library catalogues. The data may be semantic information from the header, like the types discussed in this article, or even full TEI documents. This is not limited to the present: if TEI is used as a long-term preservation format, then it will be necessary to import TEI-encoded data into other formalisms in the future, possibly including future non-XML versions of TEI. TEI P5 will not be the encoding system of choice in 2067.

42 TEI is a formalism, and like any formalism, it limits what can be said. The scope of TEI is wide, but not limitless. One can claim, and I have even done so quite recently (Eide 2014), that TEI, being based on XML, can become too much of a straitjacket. Sometimes even oral language is felt as too restrictive for the communication one needs; for example, “I could not speak, I just had to give her a hug.” No formalism works in all situations. Dubin, Senseney, and Jett (2013) point out that this important truth is often hidden in the standards themselves: “As a result of this trade-off one sees two complementary costs in standards adoption: that of relinquishing full control over stipulations to a broader community vs. the complexity necessary for generality of scope and flexibility of application. But the language of information standards doesn’t lay out this complementarity to their audience.”

43 Different formalisms, however, work at their best in particular applications and for specific purposes. The difference between modeling for production and modeling for understanding is such a difference. We saw above that Jannidis and Flanders’s distinction between altruistic and egoistic modeling fits the distinction between modeling for production and modeling for understanding. Their claim that the differences between altruistic and egoistic modeling lead to different modeling practices could be followed by a claim that they will also lead to different linking practices. Such differences are indicated in this article, but a deeper study using real life examples is clearly an important area for further research.

44 The products of any kind of intellectual and creative work, be they written or not, include the potential for future communication. If the work is scholarly, future integration with other resources should also be considered. Even when I sit privately in my study using modeling to better understand the text I am working on, I have an aim of communication, even if it may seem to be in a very distant future. Whether I model for production or for understanding, I need to be able to communicate not only my results, but also the source material behind them. In such a situation also the lack of formalism can be a hindrance to communication. So not only can overly-strict formalisms prevent communication, overly-loose formalisms can have a similar effect. We need our straitjackets to be just tight enough.

BIBLIOGRAPHY

Bekiari, Chryssoula, Martin Doerr, Patrick Le Bœuf, and Pat Riva, eds. 2013. *FRBR: Object-oriented Definition and Mapping from FRBRER, FRAD and FRSAD (version 2.0).* [Heraklion]: International Working Group on FRBR and CIDOC CRM Harmonisation. Accessed February 22, 2015. http://www.cidoc-crm.org/docs/frbr_oo//frbr_docs/FRBRoo_V2.0_draft_2013May.pdf.

Burnard, Lou. 2013. "The Evolution of the Text Encoding Initiative: From Research Project to Research Infrastructure." *Journal of the Text Encoding Initiative* 5. http://jtei.revues.org/811; doi:10.4000/jtei.811.

Ciula, Arianna, Paul Spence, and José Miguel Vieira. 2008. "Expressing Complex Associations in Medieval Historical Documents: The Henry III Fine Rolls Project." *Literary and Linguistic Computing* 23, no. 3: 311–25.

Ciula, Arianna, and Øyvind Eide. 2014. "Reflections on Cultural Heritage and Digital Humanities: Modelling in Practice and Theory." In *DATeCH '14: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* (May 19–20, Madrid, Spain), 35–41. New York: ACM. doi:10.1145/2595188.2595207.

Crofts, Nick, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff, eds. 2011. *Definition of the CIDOC Conceptual Reference Model. Version 5.0.4.* [Heraklion]: ICOM/CIDOC CRM Special Interest Group. http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf.

Dubin, David, Megan Senseney, and Jacob Jett. 2013. "What It Is vs. How We Shall: Complementary Agendas for Data Models and Architectures." In *Proceedings of Balisage: The Markup Conference 2013.* Balisage Series on Markup Technologies 10. doi:10.4242/BalisageVol10.Dubin01.

Eide, Øyvind, and Christian-Emil Ore. 2007. “Mapping from TEI to CIDOC-CRM: Will the New TEI Elements Make Any Difference?” Paper presented at TEI@20: 20 Years of Supporting the Digital Humanities. The 20th Anniversary Text Encoding Initiative Consortium Members’ Meeting, University of Maryland, College Park, November 1–3. http://www.tei-c.org/Vault/MembersMeetings/2007/.

Eide, Øyvind. 2007. “The Perspective of the Text Encoding Initiative.” In *Ontology-Driven Interoperability for Cultural Heritage Objects: Working Notes. DELOS–MultiMatch Workshop, Tirrenia, Italy, 15 February 2007*, edited by Vittore Casarosa and Carol Peters, 29–31. N.p.: DELOS/MultiMatch. http://www.delos.info/files/pdf/DELOS%20Multimatch%202007/papersdelostirrenia.pdf.

———. 2014. “Sequence, Tree and Graph at the Tip of Your Java Classes.” In “Digital Humanities 2014: Conference Abstracts,” Lausanne. http://dharchive.org/paper/DH2014/Paper-639.xml.

Gruber, Thomas Robert. 2009. “Ontology.” In *Encyclopedia of Database Systems*, edited by Ling Liu and M. Tamer Özsu, 1963–65. Boston, MA: Springer US.

Guarino, Nicola, Daniel Oberle, and Steffen Staab. 2009. “What Is an Ontology?” In *Handbook on Ontologies*, edited by Steffen Staab and Rudi Studer, 1–17. Berlin: Springer.

Havemann, Sven, Volker Settgast, René Berndt, Øyvind Eide, and Dieter W. Fellner. 2009. “The Arrigo Showcase Reloaded—Towards a Sustainable Link between 3D and Semantics.” *ACM Journal on Computing and Cultural Heritage* 2, no. 1. doi:10.1145/1551676.1551680.

Jannidis, Fotis, and Julia Flanders. 2013. “A Concept of Data Modeling for the Humanities.” In *Digital Humanities 2013: Conference Abstracts*, 237–39. Lincoln: Center for Digital Research in the Humanities. Available at http://dh2013.unl.edu/abstracts/ab-313.html.

Mahr, Bernd. 2009. “Information Science and the Logic of Models.” In *Software & Systems Modeling* 8, no. 3: 365–83. doi:10.1007/s10270-009-0119-2.

McCarty, Willard. 2005. *Humanities Computing*. Basingstoke: Palgrave Macmillan.

Ore, Christian-Emil, and Øyvind Eide. 2009. “TEI and Cultural Heritage Ontologies: Exchange of Information?” *Literary & Linguistic Computing* 24, no. 2: 161–72. doi:10.1093/llc/fqp010.

TEI Consortium. 2001. *TEI P4: Guidelines for Electronic Text Encoding and Interchange: XML-Compatible Edition*, edited by C. M. Sperberg-McQueen and Lou Burnard. N.p.: TEI Consortium. http://www.tei-c.org/release/doc/tei-p4-doc/html/.

———. 2013. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.5.0. Last updated July 26. N.p.: TEI Consortium. http://www.tei-c.org/Vault/P5/2.5.0/doc/tei-p5-doc/en/html/.

Wittern, Christian, Arianna Ciula, and Conal Tuohy. 2009. “The Making of Tei P5.” *Literary and Linguistic Computing* 24, no. 3: 281–96. doi: 10.1093/llc/fqp017.

Zafrin, Vika. 2007. “RolandHT, a Hypertext and Corpus Study.” PhD diss., Brown University. http://rolandht.org/.

Zúñiga, Gloria L. 2001. "Ontology: Its Transformation from Philosophy to Information Systems." In *FOIS '01: Proceedings of the International Conference on Formal Ontology in Information Systems—Volume 2001*, edited by Nicola Guarino, Barry Smith, and Christopher Welty, 187–97. Ogunquit, Maine: ACM. doi:10.1145/505168.505187.

NOTES

[1] This paper springs out of work in the TEI Ontologies SIG since 2004. It is also based on long discussions with colleagues at the Unit for Digital Documentation at the University of Oslo and beyond. Two anonymous reviewers of this article and the editors of the journal contributed significantly to its final form. However, the interpretation and formulation, and responsibility for its errors and omissions, is mine alone.

[2] This section is based on Eide and Ore (2007) and Eide (2007).

[3] The guidelines pointed to "Simple Analytic Mechanisms" and "Feature Structures" as "[a]ppropriate mechanisms for the encoding of such interpretative gestures" (TEI Consortium 2001, ch. 20) but these mechanisms were never in common use, and P4 did not include any standard way to represent specific external entities such as places or persons, just mechanisms that could be used to model them.

[4] K. Faith Lawrence, Michael Jewell, and Mischa Tuffield, "The OntoMedia Model," accessed January 29, 2014, http://www.contextus.net/ontomedia/model.

[5] Burnard (2013). In addition to separate persons with separate skill sets working together we also find several individuals who are acting as both modeling experts and domain experts at the same time.

[6] This was not the only output of the project, and other results were closer to modeling for understanding. See the webpage for details, accessed February 22, 2015, http://ibsen.uio.no/OmUtgaven.xhtml.

[7] Such model development will happen through iterative cycles, and the connections to hermeneutics are interesting. They are, however, beyond the scope of this paper.

[8] This is in line with the distinction between a data model as an interpretation of a domain (a representational agenda) and as a plan of action (a cohortative agenda) (Dubin, Senseney, and Jett 2013).

9 A model can always be seen both as a model of and a model for. In this specific example the TEI module can also be seen as a model of an abstract structure, that is the class of all dictionaries. Model of and model for represent different aspects of modeling and the “choice of perspective on the model will of course determine what will be brought into the foreground.” (Mahr 2009, 372).

10 This is not at all a new understanding. Some earlier examples include Ciula, Spence, and Vieira (2008), Haveman et al. (2009), as well as the harvesting of TEI data into a CIDOC-CRM–based common system in the CLAROS project, http://www.clarosnet.org/ (accessed January 29, 2014).

11 These three figures were first presented in a paper at the seminar The Message of the Old Book in the New Environment, L’Institut Finlandais en France, Paris, March 18–19, 2011, in the context of the Paris Book Fair. The colors of the dots indicate different classes in the ontology.

12 The TEI header and body in these examples do not have to be in the same document, but the situation with a header in a different document would not change the essence of the argument.

13 These elements do not need to be in the TEI header, they can, for instance, be used in authority lists in the body of a separate TEI document. The argument about the semantics remains the same, however.

14 “The Erlangen CRM / OWL,” accessed February 22, 2015, http://erlangen-crm.org.

15 “FOAF Vocabulary Specification 0.99. Namespace Document 14 January 2014 - Paddington Edition,” accessed February 22, 2015, http://xmlns.com/foaf/spec/.

16 The extension of a word is what it refers to, that is a set of things in any real or non-real world that are labelled by that word.

17 See CIDOC’s “Statement on Linked Data identifiers for museum objects,” accessed May 22, 2014, http://network.icom.museum/fileadmin/user_upload/minisites/cidoc/PDF/StatementOnLinkedDataIdentifiersForMuseumObjects.pdf.

18 Note that Mahr (2009) is a translation of a German article using the term *Informatik*, which in Germany denotes what in English is usually called computer science. In the English version of the article, however, the term was translated to *information science*.

ABSTRACT

This paper discusses the relationships between TEI and ontologies from the perspective of computer-based modeling, understood here as a way to establish meaning. The distinctions between creation and use of models as well as between modeling for production and modeling for understanding are presented and compared with other categorizations or models of modeling. One method of establishing meaning in TEI documents is achieved via linking mechanisms between TEI and external ontologies. How such linking can be done and what it may imply for the semantic openness and usability of TEI documents is the practical focus of this article.

INDEX

Keywords: ontologies, conceptual modeling, interchange, CIDOC-CRM, linked data

AUTHOR

ØYVIND EIDE

Øyvind Eide holds a PhD in Digital Humanities from King's College London (2013) and is one of the two founding conveners of the TEI Ontologies SIG. His research interests are focused on the modeling of cultural heritage information, especially as a tool for critical engagement with the relationships between texts and maps as media of communication. He is a lecturer and research associate at the Chair of Digital Humanities, University of Passau, Germany.