Historical Social Research
Historische Sozialforschung

Francesca Tomasi:

Modelling in the Digital Humanities: Conceptual Data Models and Knowledge Organization in the Cultural Heritage Domain

doi: 10.12759/hsr.suppl.31.2018.170-179

Published in:

Historical Social Research Supplement 31 (2018)

Cite as:

Tomasi, Francesca. 2018. Modelling in the Digital Humanities: Conceptual Data Models and Knowledge Organization in the Cultural Heritage Domain. *Historical Social Research* Supplement 31: 170-9. doi: 10.12759/hsr.suppl.31.2018.170-179.

Historical Social Research
Historische Sozialforschung

All articles published in HSR Supplement 31 (2018): Models and Modelling between Digital Humanities – A Multidisciplinary Perspective

Modelling in the Digital Humanities: Conceptual Data Models and Knowledge Organization in the Cultural Heritage Domain

Francesca Tomasi[*]

Abstract: *»Modellieren in den Digitalen Geisteswissenschaften: Konzeptuelle Datenmodelle und Wissensorganisation für das kulturelle Erbe«*. This paper explores the role of model and modelling in the field of Digital Humanities, paying special attention to the cultural heritage domain. In detail, the approach here described adopts a bi-dimensional vision: considering the model as both a process of abstraction, an interpretation from a certain point of view, and a formal language to implement this abstraction in order to create something processable by a machine. The role of conceptual models – to be converted into ontologies – as a semantic deepening of controlled vocabularies, is the translation of this vision. Ontologies are the models used in domain communities in order to share classes and predicates for conceptual interoperability. Thinking of data models as a knowledge organization system is the core of this reflection on Digital Humanities domain.

Keywords: Ontologies, knowledge, interpretation, data structures, controlled vocabularies.

1. Introduction

Model and modelling in the domain of Digital Humanities (DH) is a huge and challenging topic. It is not trivial to find a common and shareable definition, because the concept of model/modelling is related to multiple facets, integrating the humanistic point of view with the computer scientists' approach. Also, DH have their own notion of models and modelling (see in particular Orlandi 1999; Buzzetti 2002; McCarty 2004); concepts that also reflect a core method in DH in general and in my research on domain ontologies – or better on conceptual data modelling in the cultural heritage – in particular. But let us start from the beginning, from the attempt to find an appropriate definition.

We could say that the activity of modelling consists of choosing the features of the observed reality (e.g. an object in a domain) to be formally represented (the abstract model). This formalization requires the adoption of a data struc-

[*] Francesca Tomasi, Department of Classical Philology and Italian Studies, University of Bologna, via Zamboni 32, 40126 Bologna, Italy; francesca.tomasi@unibo.it.

ture related to a language useful for the description of the abstraction. Thus, a model refers to the declaration of the selected properties of an object, e.g. a plain text, to be translated into a machine-readable form by using a descriptive language as a representational method. Following this definition, a model is firstly a matter of extracting properties of an object as the result of an interpretation. And an interpretation is, naturally, the expression of a point of view.

In this sense, a model can never be exhaustive. Each point of view is only one of many ways to interpret the observed reality. The more viewpoints on the same object we have, the more models might be collected. So each abstraction is a possible, individual representation of an object in a domain, which is able to replicate the original object: "to an observer B, an object A* is a model of an object A to the extent that B can use A* to answer questions that interest him about A" (Minsky 1995).

But models have to be able to aggregate viewpoints. In fact, modelling also means to identify common features of a collection or extracting those patterns that could be recognized in similar resources. The similarity is a matter of sharing, i.e. sharing a genre, a type, a computational objective, a scope or a function. In this sense, modelling reveals a crowdsourced idea: sharing something within a community that decided to advocate a common idea.

By using this approach, we recognize two interrelated levels of modelling: on the one hand, the model as an abstraction, as the interpretation of the object through possibly shared and potentially multiple "lenses" (Peroni et al. 2014); on the other, the model as the choice of a language useful to implement this abstraction by creating something that is processable by a machine.

2. Controlled Vocabularies as an Abstraction

The representation of common features of the observed reality is then a matter of communicating a specific vision of the domain.

For example, a digital scholarly edition is a model because it represents the choices of the editor in creating the digital objects at each level of the representation: the transcription, the annotation or markup, the para-meta-intra-inter-textual elements, eventually the textual tradition, but also the interface, the criteria for browsing data and documents, etc. (Tomasi 2013). When the editors choose how to transcribe a document (e.g. in a diplomatic, interpretative or critical manner), or to define which features they want to be managed by the machine, they thereby define a model of the text, which they want to reproduce in a digital dimension. Each step of this process involves computational consequences.

In general, modelling, as the result of an interpretation, has to be in dialogue with a shared vision of the observed domain. This is the reason why each cultural heritage domain (from libraries to archives, from museums to galleries) endeavors to define strategies for a semantic dialogue within and between cultures that use different standard reference models.

The choice of the content model, i.e. a metadata vocabulary for describing a collection, is a matter of sharing. And sharing a model is what it takes to guarantee a basic semantic interoperability. Dublin Core (DC)[1] is, for example, a content model chosen for describing the reality through an abstraction: 15 categories able to collect all necessary features of the observed domain (a cultural heritage collection, a web page, an institutional repository, the papers of a journal, etc.).

The Text Encoding Initiative (TEI)[2] is a model expressed in embedded markup, i.e. a controlled vocabulary, a grammar or, better, a Schema (a set of elements, attributes, rules and constraints), for describing objects related to a domain, namely, the humanities. So, TEI is a common model for representing the observed reality (i.e. texts and documents in the humanities domain), but it also leaves the interpreter free to define his or her own model of the text(s)/document(s) by choosing the features to be in focus in the computational representation.

The Encoded Archival Description (EAD)[3] and the Encoded Archival Context – Corporate Bodies, Persons and Families (EAC-CPF)[4] are, again, firstly Document Type Definitions (DTDs) and Schemata created to describe archival finding aids and authority records. They are models in the archival domain, reference systems for the community. And they are both the result of the attempt to formalize the two related methodological standards, namely ISAD (International Standard for Archival Description)[5] for the archival description, and ISAAR-CPF (International Standard Archival Authority Records for Corporate Bodies, Persons and Families)[6], for the description of authority records.

Despite the different implementations, DC, EAD, EAC-CPF, and TEI are all examples among others of metadata element sets used in cultural heritage to resolve ambiguities by sharing a domain vocabulary. They want to present themselves as models, through elements and attributes, conventions, and declarations.

1 Dublin Core Metadata Element Set, Version 1.1: <http://dublincore.org/documents/dces/> (Accessed April 20, 2017).
2 TEI P5 Guidelines, latest version 3.1.0, 2016: <http://www.tei-c.org/Guidelines/P5/> (Accessed April 20, 2017).
3 EAD: <https://www.loc.gov/ead/> (Accessed April 20, 2017).
4 EAC-CPF: <http://eac.staatsbibliothek-berlin.de/> (Accessed April 20, 2017).
5 ISAD (2nd edition), 2011: <http://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition> (Accessed April 20, 2017).
6 ISAAR-CPF (2nd edition), 2011: <http://www.ica.org/en/isaar-cpf-international-standard-archival-authority-record-corporate-bodies-persons-and-families-2nd> (Accessed April 20, 2017).

Metadata models (and controlled vocabularies) take up the need to define a common conceptual architecture for a domain. It is worth noting that in the literature “metadata modelling” refers to a type of metamodelling used in software engineering and systems engineering for the analysis and construction of models applicable to and useful for some predefined class of problems. The activity of metadata modelling is reflected in a concept diagram. Unified Modeling Language (UML)[7] is the language used in the object-oriented paradigm to represent a model as a diagram. Concept, generalization, association, multiplicity and aggregation are all keywords for creating the model. So, a diagram is a model of the reality, able to represent objects in a context.

We move from controlled vocabularies to diagrams; and from diagrams to languages.

3. Languages, Data Structures and Data Types

It has indeed been said that modelling is also a matter of language. And a formal language, from a computational point of view, is a question of data structure and abstract data types, i.e. graph (the network), tree (a hierarchy), table (a relation), sequence (a list). They are, in fact, models. For all data to be organized in a digital environment one of these models is chosen to represent the observed reality. Some examples will help make this point clear:

- indeclarative markup languages, e.g. in XML, the model of the document is a tree-like structure. So the content (actually the structure) of the document is represented as a series of features hierarchically organized and nested;
- in database systems (DBMS), the more common model adopted is the relational one, namely the table. Objects are records (and thus data) and each value is related to one of the attributes (properties) that describe the reality of the objects;
- the network is the structure – the model – of the Web; but the network is also the hypertextual representation of documents (in the Web 1.0 environment), and now of interconnected data (from a Linked Open Data [LOD][8] perspective); from the sequence (a list of documents) to the graph (a network of data). Hypertext is then a model organizing data objects through their relationships.

We have to keep this aspect in mind. The formal language is another way to conceive the concept of model. The choice of the language, and of the related

7 UML: <http://www.uml.org> (Accessed April 20, 2017).
8 W3C official page on Linked Data: <http://www.w3.org/standards/semanticweb/data> (Accessed April 20, 2017).

structures, depends on the observed domain: e.g. documents (i.e. non-structured objects) or data (i.e. structured objects). With the markup, e.g. with XML as a formal meta-language, we model documents as semi-structured objects. And the aim is to reduce the narrative, in order to model content (or, better, structure) as a collection of atomic interconnected pieces to be managed as data. We move from the property-value pairs through the tree (the declarative markup language as XML) to the graph (a model as Resource Description Framework [RDF] for creating LOD).

As it has been said: in computer science, the concept of model is related to a data structure, i.e. a possible representation of a digital content or a particular way of organizing data. In this sense, the choice of the logical model (e.g. a relational database instead of a markup language) determines the computational results or, better, the computational activities and operations on the data as based on the chosen model. Hence, models play an important role in moving from theory (the abstract model) to practice, understood as the actions that can be performed (the formal language).

3.1 A Conceptual-Oriented Position

In the document community, the markup is the model, i.e. the language to represent the structure of the reality. In the data community, the model, i.e. the traditional way to represent the content of the domain, is the database. In data modelling theory, used especially in database design (although it holds true also for other contexts), we recognize three possible models, also described as three levels of abstraction of a DBMS:

- a conceptual data model
- a logical data model
- a physical data model (or better a Schema).

We begin with the latter.

At the physical level, we deal with physical means by which data are stored, which is not our level of interest.

At the logical level, we deal with structures of models again: hierarchical, network, relational and object oriented. The importance of this level lies in the fact that each chosen data structure affects the possible computational activities: even if the model is theoretical, it involves the kind of operations that we could perform with data based on one of these abstract structures (the tree, the graph, the table, the class). So the model in this case is the content, not just the structure.

Now, let us move to the first level: the conceptual data model, i.e. the abstract conceptual representation of data. On this level, data are defined from a conceptual point of view. The meaning of data depends on the context of the interrelationships with other data.

There are several notations for data modelling. The most common model is the "Entity relationship model" (E/R), because it depicts data in terms of the entities and relationships described in the data. The E/R notation yields a model, because its aim is to represent the reality as an abstraction: "this model incorporates some of the important semantic information about the real world" (Chen 1976). The conceptual model then represents concepts (entities) and connections (relationships) between them. The notation itself is an abstraction.

3.2 Ontologies and Knowledge

The same approach is adopted by ontologies, i.e. conceptual data models translated through a formal language. Again, we range between database theory and markup languages: the data-centric approach of the DBMS, the formal declarative language (XML) and the assertion (the triple) as a graph (RDF). We could say that we are dealing with the Semantic Web approach and the LOD perspective (Bizer, Heath, and Berners-Lee 2009).

In ontology design, the model is the conceptual framework. The ontology is the conceptualization of an abstraction by identifying those features, in the form of classes and predicates, which enable us to describe a domain, observed from a specific point of view. And the aim is to move from data to information in order to extract knowledge, i.e. to reveal the latent, the yet unknown. Revealing knowledge through the analysis of, for example, the context, is necessary in order to enable inferences (Daquino and Tomasi 2015). Modelling, for instance, persons, dates, places or events is an attempt to standardize a conceptual approach through relationships (Gonano et al. 2014).

EDM[9], CIDOC CRM[10] and FRBRoo[11], SKOS[12] – just to give some heterogeneous examples (see, for example, Doerr 2009) – are nothing but points of view on reality. We could assert that ontologies are the shared ideas concerning a domain, expressed with classes and properties, relationships between concepts, rules and constraints. A domain ontology is a formal, abstract representation, useful in order to semantically describe, i.e. to model, a collection of resources and to reason on data, with an inferential aim and a problem-solving approach. Another attempt to model the reality is the translation of an XML Schema, e.g. TEI, into an ontology (see, for example, Eide 2014; Ciotti and Tomasi 2015).

9 Europeana Data Model: <http://pro.europeana.eu/page/edm-documentation> (Accessed April 20, 2017).
10 CIDOC Conceptual Reference Model: <http://www.cidoc-crm.org/> (Accessed April 20, 2017).
11 FRBRoo: <http://www.cidoc-crm.org/frbroo/home-0> (Accessed April 20, 2017).
12 SKOS: <https://www.w3.org/2004/02/skos/> (Accessed April 20, 2017).

So, ontologies are models, and I think that conceptualization is the core of modelling, with reference to the issue of knowledge organization.[13] In fact, ontologies are both a way to express the semantics of a domain and a method to organize knowledge through concepts. I personally believe that in the DH domain, ontology engineering is the most effective and persuasive modelling strategy: it is a method enabling us to reproduce the brain's reasoning, i.e., the humanistic approach to interpretation.

The act of moving from controlled vocabularies to ontologies reflects the need to express the semantics that are hidden because of the absence of a content model. The creation of an interconnection through typed links is the key to solve relationships between entities in order to reveal real knowledge.

4. Final Remarks

Another definition, from the Linked Open Data perspective, is the concept of model as a conversion method:

Linked Data modelling involves data going from one model to another. For example, modelling may involve converting a tabular representation of the data to a graph-based representation. Often extracts from relational databases are modeled and converted to Linked Data to more rapidly integrate datasets from different authorities or with other open source datasets.[14]

So, the act of converting data into a different format, or using another data structure, is again a practice. The model gives the theoretical basis for a practical activity.

Finally, a model is also a question of interface. The template for a web page, for instance, is a model. The design of a page in a Content Management System (CMS) is a model. The architecture of information, understood as the position of logical components of a page, is a model. The iconic symbols are models of the reality. So, when we model a web resource, we chose a way to represent information in the visual interface: we define spaces for components and we use icons as an abstraction of an idea, we adopt glyphs as a representation of graphemes.

In conclusion, models are a guideline, models are shared by a community, models are the representation of a domain, models refer to languages and data

[13] An interesting event related to these themes is the "three-day workshop held at Brown University on data modelling in the humanities, sponsored by the NEH and the DFG, and co-organized by Fotis Jannidis and Julia Flanders". *Knowledge Organization and Data Modeling in the Humanities: An ongoing conversation*, 2012 <https://datasymposium.wordpress.com> (Accessed April 20, 2017).

[14] Best Practices for Publishing Linked Data. W3C Working Group Note 09 January 2014: <http://www.w3.org/TR/ld- bp/#MODEL> (Accessed April 20, 2017).

structures, models are a visual and iconic abstraction. Ontologies are models. Modelling is my favorite job.

5. Discussion

Günther Görz's questions (Q) and my answers (A)

Q1. I plead for a more restricted and terminological use of the term “model”. As Nelson Goodman already wrote in “Languages of Art”: “Few terms are used in popular and scientific discourse more promiscuously than ‘model’” (171).

A1. The scope of this paper is to reflect on the concept of model by using multiple perspectives. So, yes, the term is used here in order to refer to different levels, but this is exactly what I would like to get across: the ambiguity, the multiplicity and the polysemy of the word “model”.

Q2. It is true that modelling in DH is a challenging topic, but I can't see that DH already has its own notion of models and modelling compared with other interdisciplinary enterprises with computer science such as the social sciences, (cognitive) psychology, or economics.

A2. The literature in DH regarding the concept of model and modelling is so vast that I could assert that DH is elaborating its own definition.

Q3. For the formal language, the distinction between abstract data types, (concrete) data structures and their implementation should be noticed. Nevertheless, e.g. in the mentioned case of digital scholarly editions, we should distinguish between a model (the concepts, properties, constraints, structures, rules, etc.) and a particular result.

A3. In digital scholarly editing, the concept of model refers to the choice of the features to be formalized at each level of the scholarly activity. In this sense, the edition is a model: it represents the interpretative act of the editor.

Q4. I see a similar problem in calling TEI a model. In my view, TEI is first of all a formal language with an informal semantics. This view imposes several severe constraints, e.g. a fundamental tree structure due to its commitment to XML. So, I still see a deficit on the theoretical side; for me, TEI is yet more a representational framework than a model.

A4. From the formal point of view, TEI is not a model. It lacks semantics. But, from the point of view of models as a shared definition of elements and attributes related to the classification of hermeneutic aspects of a domain, TEI is a model.

Q5. Another issue is the depth of semantic modelling. In this respect, EDM, CIDOC CRM + FRBR and SKOS are not on the same level. I think we are in substantial agreement on what is said about formal ontologies: the question of semantics is tightly connected to a well-defined inference relation. Taking up TEI again, marking up named entities such as place names and representing places in a formal ontology such as CIDOC CRM are on reasonably different abstraction levels. The anything but simple question is then, how the relationship between TEI elements and CRM concepts can be formally recorded and mapped into a (partial) semantic and interoperable representation in terms of CRM, expressible in RDF/LOD. In the actually used formal systems, the most advanced of which are Descriptions logics (cf. OWL), we can deal with under specification, but not with vagueness. This is one of the very big challenges of the humanities and science.

A5. EDM, CIDOC CRM, FRBR, and SKOS are not on the same level, I agree. The semantic depth is surely different. But they are all models, i.e. point of views: how to integrate metadata vocabularies (EDM), how to use an event-centric approach in the cultural heritage (CIDOC CRM), how to document the stratification of object descriptions (FRBR), how to express structured subjects in a domain (SKOS). So, again, they are not all models from the viewpoint of formal languages to describe concepts, i.e. ontology, but because of their attempt to define a shared conceptualization. Translating the TEI Schema into an ontology (e.g. an OWL representation), or thinking on TEI as a CRM, is a challenging issue (see, for example, Eide 2014; Ciotti and Tomasi 2015).

Q6. Finally, reasoning with formal ontologies is, up to now, deductive reasoning. But for reasoning in the humanities and in science other forms are also needed, something that Leibniz called "ars inventoria".

A6. Yes, formal reasoning is the final aim. And the role of ontologies is to enable inferences through description logic formalism. But this is just one of the various ways to interpret the concept of model.

References

Bizer, Christian, Tom Heath, and Tim Berners-Lee. 2009. Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems* 5 (3): 1-22.

Buzzetti, Dino. 2002. Digital Representation and the Text Model. *New Literary History* 33 (1): 61-88.

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

Ciotti, Fabio, and Francesca Tomasi. 2016. Formal ontologies, Linked Data and TEI Semantics. *Journal of the Text Encoding Initiative* 9 <http://jtei.revues.org/1480> (Accessed April 21, 2016).

Daquino, Marilena, and Francesca Tomasi. 2015. Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects. In *Metadata and Semantics Research*, 424-36. Berlin: Springer.

Doerr, Martin. 2009. Ontologies for cultural heritage. In *Handbook on Ontologies*, 463-86. Berlin: Springer.

Eide, Øyvind. 2014. Ontologies, Data Modeling, and TEI. *Journal of the Text Encoding Initiative* 8 <http://jtei.revues.org/1191> (Accessed April 21, 2016).

Gonano, Ciro M., Francesca Mambelli, Silvio Peroni, Francesca Tomasi, and Fabio Vitali. 2014. Zeri e LODE. Extracting the Zeri photo archive to Linked Open Data: formalizing the conceptual model. In *Digital Libraries (JCDL)*, 289-98. London: IEEE.
McCarty, Willard. 2004. Modeling: A Study in Words and Meanings. In *Companion to Digital Humanities*. Oxford: Blackwell <http://www.digitalhumanities.org/companion/> (Accessed April 21, 2016).
Minsky, Marvin L. 1995. *Matter, Mind and Models* <http://groups.csail.mit.edu/medg/people/doyle/gallery/minsky/mmm.html> (Accessed April 21, 2016). (Rev. version of the essay in *Semantic Information Processing*, ed. Marvin Minsky. Cambridge, MA: MIT Press, 1968).
Orlandi, Tito. 1999. Linguistica, sistemi, e modelli. In *Il ruolo del modello nella scienza e nel sapere* (Roma, 27-28 ottobre 1998) (= Contributi del Centro Linceo Interdisciplinare, n. 100), Roma: Accademia dei Lincei, 73-90.
Peroni, Silvio, Francesca Tomasi, Fabio Vitali, and Jacopo Zingoni. 2014. Semantic lenses as exploration method for scholarly article. In *Bridging between Cultural Heritage Institutions*, 118-29. Berlin: Springer.
Tomasi, Francesca. 2013. *Vespasiano da Bisticci, Lettere. A digital edition*. Bologna: University of Bologna <http://www.vespasianodabisticciletters.unibo.it> (Accessed April 21, 2016).