Knowledge Organization and Data Modeling in the Humanities

Julia Flanders, Northeastern University

Fotis Jannidis, University of Würzburg

2015

This white paper and the workshop on which it is based were made possible with generous funding from the National Endowment for the Humanities and the Deutsche Forschungsgemeinschaft.

Preface

On March 14–16, 2012, the Women Writers Project at the Brown University and the Center for Digital Editions at the University of Würzburg organized a three-day workshop with generous sponsorship from the National Endowment for the Humanities and the Deutsche Forschungsgemeinschaft. The event brought together a group of about 35 on-site participants and a community of about 85 virtual participants who followed the live stream and the Twitter feed. Some of these virtual participants asked questions or contributed comments, and some of them helped communicate the substance of the event to a wider Twitter audience.

The design of the event was aimed at creating the conditions for a substantive, wide-ranging, interdisciplinary, expert discussion of humanities data modeling grounded in many relevant fields. To accomplish this, we invited participants to make several different kinds of contributions, including papers on theoretical approaches, case studies grounded in specific projects, focused panel discussions on theory and on pedagogy, and also more general discussion at intervals throughout the event. Slides, papers, and videos of the workshop can be found online.[1]

The following white paper tries to sum up important topics and problems which came up in the presentations and discussions and to outline some general aspects of data modeling in digital humanities. Its aim is to offer a reference point for further research and to stimulate discussion on a topic crucial to Digital Humanities.

I. Introduction

The question of what would constitute "theory" for digital humanities is contested and somewhat fraught. From the perspective of humanists for whom "theory" means the cultural and critical theory that has been naturalized in humanities departments during the past thirty years, Alan Liu's recent question, "where is cultural criticism in the digital humanities?" reflects a general sense that the digital humanities has done too little to theorize its work (Liu 2012). For others working in the field, the place to look for digital humanities "theory" would be rather in the philosophical, logical, computational, and mathematical systems that undergird the representational structures we use. Thus the statement that "the database is the theory" could be unpacked to refer to work that examines theories of database design, first-order logic, ontologies, and other fundamental systems of information structure (Bauer 2011). Both of these positions create an artificial distance: the assumption that digital humanities theory will be very close to established cultural theories ignores the fact that digital humanities is bringing together two very different fields both with their own and very different theoretical traditions and that both traditions will have to be considered and, if possible, merged in this new field.

The claim that "the database is the theory" on the other hand is in conflict with the meaning of 'theory'. Theory is usually the theory *of something*, trying to spell out the basic concepts relevant in the praxis of

[1] http://www.wwp.northeastern.edu/outreach/conference/kodm2012/index.html

doing something. An interpretation and a theory of interpretation for example are therefore very different even if the interpretation is very theory-informed — they try to achieve different goals and they typically use different means to do so. Therefore a theory of digital humanities cannot simply coincide with its praxis. It can, as a human and social activity, very probably learn a lot from older theories — the relationship of actions to power structures comes to mind — but first of all it must be founded in a very close look at the activities of digital humanists, especially if it wants to come close to the original meaning of theory. In the following paper we will discuss data modeling as an abstraction of many practices in the digital humanities and we hope this will be understood as a contribution to a theory of digital humanities.

1. Defining Data Modeling

The question of how to define data modeling is of course of central concern, not only as a question of clarification of terms, but also as a larger question of how to situate data modeling in the appropriate context. The situation is troubled at the outset by the fact that the term “data modeling” in computer science is most typically used in a fairly restrictive sense for the modeling of relational databases, while the digital humanities has a more general understanding of the term: data modeling is the modeling of some segment of the world in such a way to make some aspects computable, referring to creating database schemas, SGML DTDs, XML schemas, ontologies etc. Thus while one can learn a great deal from the extensive discussion of data modeling in computer science, the task at hand is to define and study the more general concept.

In computer science, data modeling is “a collection of conceptual tools for describing data, data relationships, data semantics, and consistency constraints” (Silberschatz et.al. 1996: 7). Another definition puts stress on the logical side: “A *data model* [...] is an abstract, self-contained, logical definition of the data structures, data operators, and so forth, that together make up the abstract machine with which users interact.” (Date 2012: 12). Data modeling is thus understood to consist of a set of steps. The first is *conceptual data modeling:* the identification and description of the entities and their relationship in the “universe of discourse” (i.e. that part of the world a modeler is modeling), and notation of the findings, for example in an entity-relationship diagram. The second is *logical data modeling*: defining the tables of a database according the underlying relational model. The third is *physical data modeling*: optimization of the database for performance, in an actual implementation. There seems to be a consensus that the third level is somewhat at the periphery of data modeling. Ideally both the conceptual and the logical model should be designed without any reference to the implementation. Thus the implementation can be optimized or even replaced at a later time. Even if the distinction between the logical and the conceptual level is a result of specific database modeling techniques, it captures an important general aspect of data modeling. The logical model provides a structure for the data which allows the user to use a set of algorithms to answer questions of interest in relation to the data. This computability is usually achieved by using a mathematical model: relation, in the case of databases, or trees/graphs in the case of XML. In these cases the logical model is a powerful formal abstraction, but it fails to represent most of the semantic information. The conceptual model addresses this lack: it captures semantic information and offers an integral and embedded view of the data, organizing the information in such a way that the logical model can either be derived automatically or is at least very easy to derive.

From the existing work in computer science and philosophy, we thus have a set of definitions and a conceptual foundation for discussing data modeling. However, when relocating or reusing these concepts to discuss data modeling in the digital humanities, some points of friction emerge and also some issues that require more detailed scrutiny. In the workshop on which this white paper reports, the definitions of data modeling that emerged from within the digital humanities frame of reference were much less clear-cut. The participants were deliberately chosen to represent fairly diverse fields, including linguistics, scholarly editing, history, visual arts, game studies, classics, philology, literary studies, and geography. For many participants, given that "data" could be considered as whatever formal information one might have in one's research landscape, and "modeling" could be considered as any act of formal structuring, "data modeling" thus could be defined in a loose sense as the set of formal representational activities through which information is constituted in digital form. In this sense, it was suggested that "data modeling" and "information modeling" might be at some level interchangeable terms.

However, the discussion also involved numerous attempts to add precision to this definition—to distinguish data modeling from other representational activities—and also to explore some related questions about how a concept like "data modeling" might function distinctively within the field of digital humanities. The effort to establish the boundaries of the concept involves thinking about how we scope the project of data modeling, the terminology we use, and the relation between data modeling and related domains. In the discussion below, we consider these questions in more detail.

The Project of Data Modeling

In the workshop discussion, it became very clear that the enterprise of data modeling itself—understood at the highest level—needs to be undertaken with an awareness of what is at stake, both for the agent doing the modeling and also for the materials being modeled. This is not simply a matter of identifying the strategic goals being served, but also of locating that enterprise within certain conceptual frameworks that help make explicit the assumptions that inform the work. Our discussion focused on three such frameworks, although these are not presented as an exhaustive list.

User requirements: curation-driven and research driven modeling. It is commonly noted in the literature on data modeling that in order to create and evaluate a model one has to have a clear understanding of the user requirements for the data model. In the workshop discussions we noted an interesting duality in this respect. On the one hand, data models 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 (archivists, libraries, others whom we might characterize as "curation-driven modelers"). On the other hand, data models also exist whose function is to express specific research ideas in cases where data is being created to support the creator's own research needs (particularly for individual scholars and projects, whom we might characterize as "research-driven modelers"). Curation-driven modelers also make assumptions about what features of the digital objects are of interest for most users and in most use cases, while research-driven modelers typically concentrate more (though not exclusively) on the needs of their own project. Thus, we have in practice two different approaches to the task of modeling. One seeks to anticipate and synthesize very different views on digital objects and aiming to establish standards, and this involves very specific processes to decide on these user needs and to connect these new models with existing traditions of modeling, for example those arising in library science. The other is interested mainly in expressing as exactly as possible the theoretical assumptions and research interests of one or more scholars.

Constraint systems: normative, descriptive, and exploratory modeling. Another framework within which we can usefully locate ourselves and the modeling enterprise is that of constraint systems and their relation to the variability of the materials being modeled. In some contexts, the role of the constraint system—the schema or set of data definitions—is to describe in formal terms a landscape of information which we seek to understand in detail. The schema in this context serves as a proxy for the source material (at least in structural terms) and the modeling process is one of elaborating the schema until it constitutes a complete inventory of the features we observe in that material. The alternative approach is one in which the constraint system represents an *a priori* set of stipulations concerning the shape of the resulting data, which dictate which features of the source material can be represented in the final digital surrogates. These stipulations may arise from functional constraints (e.g. the specific features of a specific publishing system, or the requirements of a collaborative partner or external aggregator) or from theoretical convictions about genre or method (for instance, in a linguistic modeling of a novel, page breaks would probably be omitted). In practice, the descriptive extreme is common; even a fundamentally descriptive approach (as practiced for instance by an archival project interested in representing a very broad range of document features) will necessarily be shaped by some assumptions about genre. However, in large-scale digitization projects the normative extreme is fairly common, if only because the descriptive language in play is likely to be so impoverished that it necessarily omits or elides features it cannot accommodate. A third approach is sometimes evident in research-driven modeling, where the data model may be evolving during the research process and the formal expression of the model through a schema serves as a way of exploring and expressing a set of current assumptions rather than as a way of ensuring uniformity of the data.

Products: the modeled artifact and the model itself. A third framework to consider is that of the product of the modeling process. As noted in more detail below, there are several ways to construe the term "model": as referring to the *modeled instance* (that is, a representation of an object that offers a useful simplification of it for analytical purposes), or as referring to the underlying schema or formal statement of properties that is understood as characterizing or typifying the set of instances being modeled. Both of these are useful outcomes of the modeling process, and in the digital humanities both are considered important research outcomes, but they have their value in quite different contexts. The modeled instance (especially in large collections) represents an important contribution to scholarship by providing new research artifacts to which new methods and tools can be applied. The large investment made during the past twenty years in digitizing humanities content (an investment made both by governments and by institutions) is testimony to the value of the modeled instance and the importance we attach to it in digitally mediated research. However, the model itself is arguably an equally important outcome in the sense that it offers both an expression of method (how did we approach the task of modeling?) and also an insight into the deeper patterns that inhabit the modeled instances, taken as a set. However valuable a collection of (for instance) TEI-encoded texts may be for scholars studying a particular genre or author, the TEI Guidelines themselves might be considered an equivalent achievement, as an argument about how texts behave, and this is true whether or not we agree with the TEI's specific modeling decisions.

Discipline and Terminology

Another area of additional complexity arises from the disciplinary breadth of digital humanities, and the consequent flexibility of terminology used to describe the object being modeled, with consequences for our understanding of what information is to be modeled, and what the model is for. Terms such as "text", "document", "model", "transcription" , and even "data" are used in humanities discourse in flexible ways that reflect both different conceptions of these terms (arising from different disciplinary contexts) and also in some cases long-standing debates within specific disciplines: for instance, the relationship between "text" and "document" within traditions of textual scholarship. To speak of "modeling a text" requires us to specify whether we consider a text to be a linguistic object, a material object, a conceptual object (akin to the FRBR "work"), or even in some cases a string considered apart from its possible linguistic properties. In developing concepts of data modeling in the humanities, for some purposes we thus need to move from a colloquial and implicit understanding of such terms to a more precisely and explicitly elaborated set of terms and definitions for use in contexts where definitional distinctions matter.

Data Modeling and "Just Plain Modeling"

As the workshop discussion revealed, the term "data modeling" in the humanities stands in an unclear relationship to the concept of "modeling" more generally. As Willard McCarty's exploration of the term in "Knowing … : Modeling in Literary Studies" suggests, it is possible to understand modeling as an epistemological activity:[2] "use of a likeness to gain knowledge of its original," and this sense of the term was common in the workshop discussion.

Is "modeling" on its own something distinct, or is all modeling is ultimately data modeling? Two positions emerged. The first held that there is no difference between modeling and data modeling; any kind of modeling in the humanities is data modeling, since modeling is the mode through which we apprehend the perceptual world (and hence the mode through which data about that world comes into being). In this view, "data" is understood as representing a broad continuum of observations or information, not only those that are highly formalized, and "data modeling" refers simply to the inevitable perceptual activities through which such information comes into existence. The second

[2] McCarty has been one of the first to point out how important the concept of modeling is for a deeper understanding of digital humanities and that a theory of modeling should be at the core of a theoretical description of the field (McCarty 2005).

position holds that data modeling is a specialized kind of modeling, aimed specifically towards computational processes and therefore subject to more formalized constraints. In this view, the category of "data" is more narrowly defined to refer to highly structured information that has been designed for use in research, and "data modeling" refers to the work of creating that data.

In this sense, data modeling in digital humanities can be seen as a point of connection between two intellectual domains that are traditionally considered distinct: a "higher" domain in which we acknowledge an interpretative relationship with an artifact, and a "lower" domain in which there is process modeling, data structures, etc.

Data Modeling and Classification

Data modeling is — in some respects — a classification task (Sperberg-McQueen 2004): it relies on the clear definition of classes (or "entity sets" as they are termed in computer science). In Classical Theory (arising from the philosophy of cognitive science)[3] classes are defined by a list of attributes shared by all members of the class. Members of the class ("entities") are defined by these attribute values; if an entity lacks one or more attributes it is not a member of this specific class. This approach has many positive sides, because analysis of a class can thus be reduced to determining those attributes which are essential: "Most concepts (esp. lexical concepts) are structured mental representations that encode a set of necessary and sufficient conditions for their application, if possible, in sensory or perceptual terms." (Margolis and Laurence 1999: 10). However, the drawback of this approach is the fact that humans usually do not handle concepts in such a way; their everyday understanding and their usage of concepts seems to be organized around the idea of what is *typical* rather than *essential* in the class. For instance, research by Eleanor Rosch on the organisation of the concept "bird" has shown that some birds, like blackbirds or sparrows, are more typical for the concept "bird" than other birds such as ostriches or penguins. The difference can be measured: in classification tasks subjects need longer to classify the latter as birds. Hence in cognitive science, the prototype theory is well established as a way to describe the way humans handle classes: by identifying a set of features that make up the typical case, while allowing for variability in which features are actually present in any given instance.

Coming to our discussion about classification from cognitive science, one cannot help noticing that the undisputed model of a class in computer science is that of the "classical theory." The prototype theory described above is one of the most important proposals to overcome the shortcomings of the classical theory. However, according to Simsion these discussions have not found their way into data modeling in computer science (Simsion 2007: 79). Especially in cases where the main task is to model artefacts which are organized into classes by their users (an activity common in digital humanities), the application of a class concept following the requirements of the classical theory will result in a lot of unsatisfactorily borderline cases that seem to challenge the class definition by omitting one or two features.

In digital humanities, some contexts require this kind of hard-edged classification: for instance, in cases

[3] See Margolis and Laurence 1999 on the distinction between Classical Theory and Prototype Theory.

where the classification will drive retrieval or other basic functional systems (in which omission of essential attributes will result in operational failure). But in many cases a better outcome would result from an approach in which edge cases can still be considered instances of a class—for instance, letters that are missing their bylines and signatures—and our models and delivery systems thus need to include provision for graceful failure in the case of variation from the type. In digital humanities data modeling, the entity definitions we develop to describe the objects we are modeling thus are not complete and are not meant to represent a full statement about the ontology of the object. Rather, they are driven by pragmatic concerns: they contain just those features which are necessary to fulfill the user requirements the data model is designed for. In some cases these requirements may dictate that a given attribute be required, but this reflects functional considerations rather than ontological primacy. Thus the class "person" in an address database, for example, is easily described with attributes like "first name [required]" or "phone number [optional]" and avoids thus the pitfalls of attempting a full-fledged definition of personhood. For our data modeling activities to be successful, it is essential to understand clearly what kind of entity the model is supposed to represent, and what aspects of the entity are essential in the light of the user requirements.

Provisional Definitions

With these considerations in mind, we can attempt to arrive at a definition of data modeling that responds to the special circumstances of the humanities, while also taking advantage of the rigor and precision of definitions arising from computer science. A few points must be made at the outset. **First**, when we speak in the following of "data modeling", we are referring to the models we use to shape digital surrogates and born-digital objects. While data structure is the more technical term referring to the way that the data model is represented in the internal or external memory of the of the computer, 'data model' refers to the conceptual and logical view. These views can be written down in a formal notational system like an entity relationship diagram (conceptual view) or a schema (logical view) but this diagram or schema will only make sense with an accompanying prose text defining the entities and the relations.

Second, data models have three main functions, the first two of which have to do with communicating with the computer. First, the underlying logical model allows specific operations on the data; and second, data models allow us to constrain the kind of data allowed at specific points of the model thus ensuring the consistency of the data in regard to the conceptual model and in respect to the operations on the data. Third, a kind of social praxis, not just for the computer: allows us to communicate about the data (i.e. express our ideas about the structure of the data that we think is important within a specific mode of discourse being modeled). Both aspects contribute to the fact that the data model as a system of consistency constraints plays a crucial role in retaining the semantics of the data.

Third, data models describe (in a more or less formalized way) structures of data, so there is a difference between the data and this information structure. From computer science we can borrow the distinction between:

- a **modeled instance** for example the structure of a text expressed in some markup; or an address book organized as a table

- a **data model**, for example the schema the textual markup is conformant to (like TEI), or the structure of the table
- the **meta model,** for example XML, as a specific way to express information structures, or the relational model)

At least in the XML world the relationship between the modeled instance and the data model can vary considerably: the modeled instance very often instantiates just one of many very different relationships of the elements. That is to say it belongs to the class described by the data model while the class may contain many instances which can differ considerably from the structure described if one would construe a schema just based on the modeled instance.

And **fourth, we must consider ontologies.** The relation between the concepts 'data model' and 'ontology' is uneasy at best, not least because some use 'data modeling' as referring specifically to a relational database technique while some, but only very few, use it as a broader term. Looking at the following definition of ontology by Tom Gruber, who was one of the first to use the term "ontology" in the new computer science meaning [i.e. not the established meaning from philosophy, i.e. the science of being], one can easily see that it is almost identical with conceptual data modeling as described above:

> In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application. In the context of database systems, ontology can be viewed as a level of abstraction of data models, analogous to hierarchical and relational models, but intended for modeling knowledge about individuals, their attributes, and their relationships to other individuals. Ontologies are typically specified in languages that allow abstraction away from data structures and implementation strategies; in practice, the languages of ontologies are closer in expressive power to first-order logic than languages used to model databases. For this reason, ontologies are said to be at the "semantic" level, whereas database schema are models of data at the "logical" or "physical" level. (Gruber 2009)

However, there are two key differences between an ontology and a conceptual model expressed in a ER-diagram. First, as Gruber notes, "Ontologies are often equated with taxonomic hierarchies of classes, class definitions, and the subsumption relation" (Gruber 2009) while ER-diagrams express all kind of relations, for example "PROJECT-WORKER" (Chen 1976: 12). And second, ontologies are designed "for the purpose of enabling knowledge sharing and reuse"[4]—in other words, they are explicitly intended as a form of communication and interchange, rather than as an internal part of a local system.[5] As we are using the term 'data modeling' to refer to a general notion which encompasses all forms of modeling of data, the concept of ontology becomes a subclass referring to those models which have a very specific

[4] Tom Gruber is well-known for the first widely accepted definition of 'ontology' in computer science in 1992. The second point is repeatedly empasized in CS literature, see for example Spyns et al. 2002.

[5] It may be useful to consider ontologies as a pure form of what we are calling "curation-driven" (or "altruistic") data modeling.

user requirement: represent a conceptualization of a domain that is commonly agreed to by most parties and that is relatively task independent (Dillon et al. 2008). Based on these two constraints we could say that ontologies belong to the more general class of data models, but are restricted to the conceptual level: the entities are often organised in taxonomic hierarchies and the main user requirement is enabling knowledge sharing and reuse. On the other hand almost all data models do include an ontology — at least in a weaker sense — as long as there is a conceptual level and its classes can be organised in a taxonomic hierarchy and the representational primitives are not meant to be a part of a private language.[6] And for some, it is possible to see on each level a specific ontology:

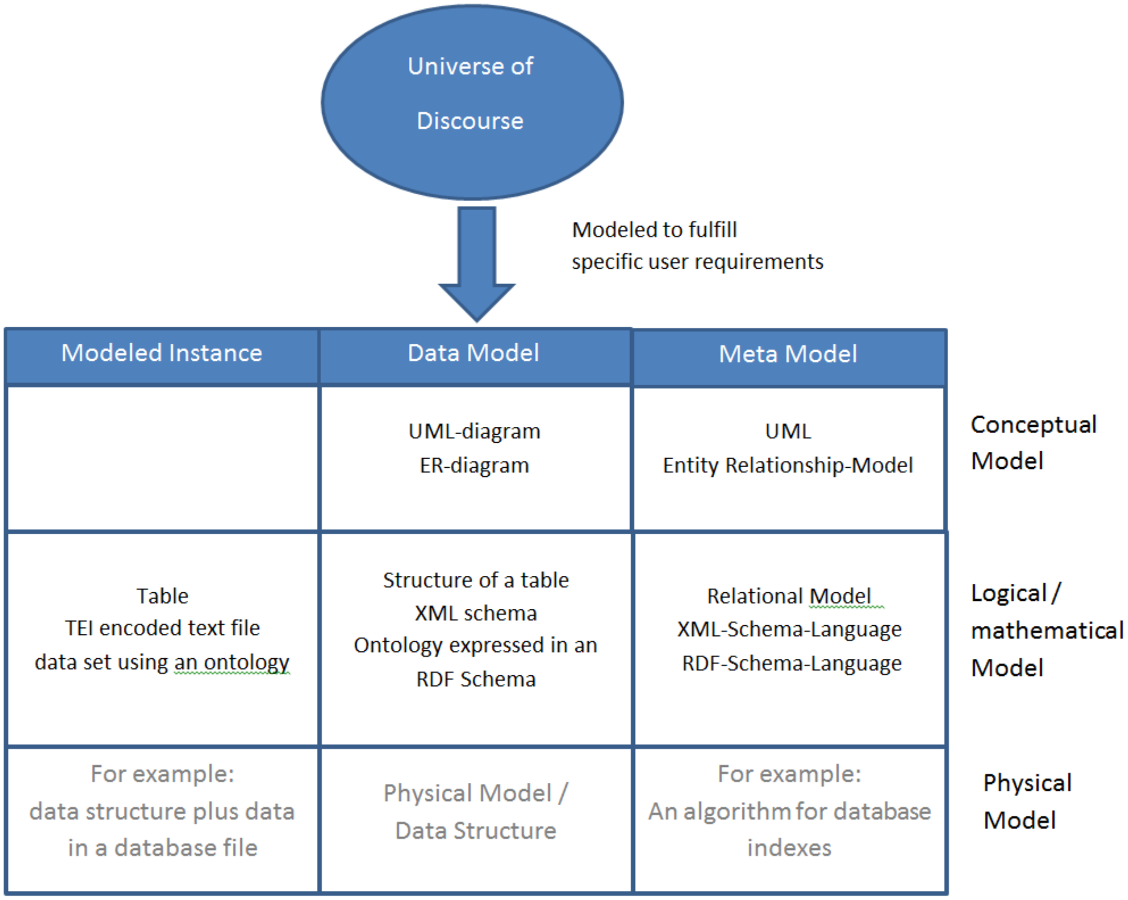
> Since an ontology is a model of a domain describing objects that inhabit it, all three types of data models can be thought of as ontologies. They range from the most expressive one that describes business concepts and processes (the conceptual model) to less expressive and progressively moving from describing business semantics to describing physical structures of the data as it is stored in the databases (the logical and physical data model). (Polikoff 2011)

Before we conclude our effort to define data modeling, a further complication needs to be considered. We have used the phrase “data modeling” to describe two closely related activities:

- the process of creating a data model.
- the process of applying a data model to data in order to create a modeled instance.

In the following discussion, we are usually talking about “data modeling” in the first sense, of “creating a data model”, and we flag explicitly any cases where we use it in the second sense. The following diagram integrates many of the distinctions discussed above:

[6] If we use data modeling in the relational database meaning then another difference is important: databases usually make an closed world assumption, that is all rows of a database can be seen as predicates and only these are true for the universe of discourse represented while all others are false. Ontologies on the other hand usually make an open world assumption, that is knowledge not included is unknown and not false.



Modeled Instance	Data Model	Meta Model	
	UML-diagram ER-diagram	UML Entity Relationship-Model	Conceptual Model
Table TEI encoded text file data set using an ontology	Structure of a table XML schema Ontology expressed in an RDF Schema	Relational Model XML-Schema-Language RDF-Schema-Language	Logical / mathematical Model
For example: data structure plus data in a database file	Physical Model / Data Structure	For example: An algorithm for database indexes	Physical Model

The Universe of Discourse is the part of the world the model tries to represent. The representation is mainly determined by two goals: to meet the demands of the user requirements and to integrate those aspects of the world (entities, attributes, relationships etc.) that are deemed to be essential.

As explained above, we can distinguish three *levels* of data modeling. Usually nowadays there is a meta model like the XML schema language which allows users to express specific models. In addition, we have a data model representing the structures of a specific aspect of the world, but independent of a specific instance or a specific moment in time; for example, an XML schema like the TEI which describes literary and linguistic features of texts. And finally, we have a specific instance, for example a TEI-encoded text which conforms to the TEI schema.

At the same time we can distinguish three different *forms* of data modeling. The conceptual model retains a lot of semantic information about the universe of discourse but is already quite formalized, for example an Entity-Relationship diagram. The logical model is a more abstract rendering of this structure, expressed in a mathematical form which allows it to be processed by a computer: for example, the structure of a table. And finally, there is also a physical model describing the way the data is handled on a level close to the hardware. In data modeling this level is usually ignored because it belongs to an area demanding a very different kind of expertise.

In the XML world the logical model can be expressed with an XML schema, but there is no common conceptual model for an XML schema. Maler and Andaloussi propose a tree diagram which has some of the functionality of an ER-diagram (Maler and Andaloussi 1995). There have also been some attempts to extend existing technologies like UML or ORM (Object Role Modeling) but most have difficulties to cover all aspects of XML.[7] In the TEI context the ODD format[8] has been developed which contains the XML schema (as schema fragments) and the prose text explaining the tags and attributes in their context. The format lacks a graphical representation and the abstraction from the logical model, which usually can be found in the conceptual model but preserves some of the functionality of a conceptual model. As mentioned before, there is a marked difference between the logical and the conceptual model: the first is an abstraction of the latter, contains less information because this part of the conceptual model can only be expressed in natural language. For the XML world there have been attempts to overcome this limitation by describing a "formal tag-set description" which would allow inferences based on the semantics of the markup, but the problems inherent in XML and the common usage of XML are formidable (Sperberg-McQueen and Huitfeldt 2011).

Data Modeling and Metadata

When we create modeled instances (such as databases, XML-encoded texts, RDF ontologies, and the like) we are necessarily creating hybrid forms in which the apparatus of the model itself is embedded in, or overlaid upon, the data being modeled. This apparatus may be expressed through annotations that are clearly external to the modeled data, or through notations that are interwoven into the data but distinguishable from it through the markup syntax (e.g. the "tags" in XML). This special status of the modeling apparatus—separable from, yet closely related to, the data—in some ways resembles the special status of metadata and this similarity leads to some potential confusion between the two. Indeed, in introductory discussions of markup it is common to ask the question whether markup might be considered as a species of metadata. This question is understandable, particularly since in some modeling contexts there is no other formal metadata. For instance, in the world of databases, the information about records, fields, and data types that enables us to understand the semantics of the data is intelligible as a kind of meta-commentary on the structure of the database and hence as metadata. Coming at the question from a humanities perspective, the concept of metadata plays a somewhat different role, since in the humanities tradition the item under consideration is an artifact (a text, an object, an art work, etc.) which it is the purpose of metadata to describe and account for. This metadata, which may take the form of a catalogue record, a finding aid, a title page, an entry in a catalogue raisonnée, or some other similar statement, accompanies the artifact as part of its identity and the apparatus of its authentication, and hence when the artifact itself is remediated and modeled digitally, that metadata becomes by extension part of the artifactual ecology that needs to be modeled. The semantics of the modeling system itself—the field definitions or element definitions—also require documentation, but this is typically done in quite a

[7] A comprehensive overview can be found in Chen and Liao 2010.

[8] "One Document Does it all," a literate-programming format that represents the TEI schema in a conceptual, documented form from which an actual schema (in one of a number of different schema languages) can be generated.

different way, for instance through a schema (or in the case of the TEI through a meta-schema). From a digital humanities perspective, then, it is useful to retain this distinction between the metadata that documents the object itself, and the modeling apparatus that annotates and shapes the representation of that object in digital form.

2. Data Modeling in the Humanities

Identifying the relevant entities in the world of discourse, discerning their relevant features, describing their connections, all this is at the core of data modeling in general — "relevant" being always defined in relation to the user requirements. In the humanities very often these entities have been discussed at length as concepts either on the object-level or the meta-level. On the object-level humanists reconstruct the use of a term at a specific location and during a specific time; the history of ideas for example is a prominent field of research interested in this kind of reconstruction. On the meta-level humanists define and use terms to classify objects, for example when they attribute literary texts to an epoch (such as Realism or Baroque) or a genre (such as the Bildungsroman). In both cases the construction or reconstruction of these concepts is embedded in a longer history of discussing them, trying to capture their salient features and to find definitions. These definitions vary in their strictness, and thus can be formalized to a greater or lesser degree, but they almost always contain much more information about the entities than the formalization can capture. And because of the historical nature of most of the entities and because of its tendency to self-reflection researchers in digital humanities view the intellectual work on this layer as an important work on its own.[9] On the other hand, even a set of very strict definitions may be not formalized enough to be the basis for a *data* model and thus demand the extra work to formalize them. It is this embeddedness which makes brute force approaches which are often the first step in computer science so unsatisfactory for cultural objects.

Data modeling can be seen as an activity involved in many different activities in digital humanities: for example, creating databases to capture informational details of cultural objects, creating digital editions by using text markup to represent the structure of text and witness information, creating software for research purposes to work on specific data sets. All these activities are similar in that the researcher has to decide what features of the object are important enough to invest time to make them explicit, and how to describe and relate these features to each other into some general structure. The results of data modeling can be found in manuals, schemas, database designs, software designs, stylesheets and many other places; often they are not very well documented and not part of the explicit description of a project. People working in the digital humanities come across data modeling in these different ways and there is no general understanding that these activities belong to the same class and should be understood to have many commonalities our discipline has to identify, to collect, to accommodate to new ways of thinking and to teach.

In particular, we are interested in thinking about how the term "data modeling" would operate within a

[9] This can be seen in projects as different as Elaine Svenonius' description of the theoretical foundations of cataloging and the prose in the TEI Guidelines; see Svenonius 2000; TEI Guidelines.

humanities context, and whether there is anything specific about modeling in such a context: whether because of special properties of humanities research objects, humanities methods and practices, or humanistic approaches to working with data. In the workshop discussions, several important features of humanities research objects were noted:

- they are usually artifacts rather than natural objects; human agency has played a role in their making or selection;
- they are created with a purpose and an audience;
- their history and provenance is part of their identity.[10]

These three features all contribute to our understanding of such objects, but they also contribute to the field of information that we expect to model about these objects in order to create adequate representations of them for research purposes.

In addition, there seems to be a marked difference to the common understanding of data modeling in the digital humanities. In computer science, among both theorists working in the academy and those working in industry doing practical data modeling, most regard data modeling as a description of a real and objective world (which includes the possibility of assessing the correctness of data models) while only a minority views it as a design process.[11] However, in digital humanities there seems to be a general understanding that a data model, like all models, is an interpretation of an object, either in real life or in the digital realm. Michael Sperberg-McQueen in his closing keynote to the workshop stated this position clearly: “modeling is a way to make explicit our assumptions about the nature of a text/artefact.” Furthermore, most digital humanities researchers assume that data modeling is primarily a constructive and creative process and that the functions of the digital surrogate determine what aspects have to be modeled.

There is both an ontological and an epistemological question here. The ontological question concerns the presence of a common objective (or at least intersubjective) reality. Such a reality is difficult to establish even for material or natural objects, and it is even more difficult if we include all of the entities that play a larger role in data modeling, such as audience, purpose, date, etc., all of which are social constructions. The epistemological issue concerns whether it is possible to construct multiple, equally valid views of a socially constructed reality. Most modern-day researchers would probably agree that it is, because we have learned in the last hundred years how each view is entrenched in a complex set of preconditions everyone has acquired. With these two aspects we have a sound basis for a generally shared intuition that on the one hand data modeling is a way to make a specific view of the world explicit, and therefore there

[10] During the workshop Allen Renear observed that data modeling in the humanities must take account of concepts of intentionality arising from Husserl: in this view, intentionality is inherent in human engagement with objects, and hence needs to be captured in the modeling.

[11] Simsion bases this statement on the analysis of computer science literature on data modeling and a broad set of interviews and questionnaires. Simsion’s book is centered around this distinction and states: “definitions of data modeling commonly characterize it as a description, but there are dissenting views and common metaphors which align better with the design characterization.” (Simsion 2007: 32).

can be many equally valid views—but on the other hand it is also possible to speak of more or less effective or useful data models, and perhaps of useless or wrong data models as well. Some data models cover a socially constructed class in ways which seem to most observers within a research community to represent more features in a more elegant or economic way than others, yielding a consensus concerning the effectiveness or functionality of the representation rather than the object being represented. Finally, it is important to note that while the level of consensus concerning an entity to be modeled may vary considerably — for instance, the existence and nature of the "paragraph" may seem quite uncontroversial while that of "free indirect discourse" or "irony" may be much more debatable — nonetheless as we expand the boundaries of the community we inevitably find even "uncontroversial" entities being destabilized.

Based on these observations we offer our own definition, applicable to the digital humanities: Data modeling refers to the activity of designing a model of some real (or fictional) world segment to fulfill a specific set of user requirements using one or more of the metamodels available in order to make some aspects of the data computable and to enable consistency constraints. Two typical forms of user requirements can be found. One aims to define a data model for a domain in order to support data exchange and long-term use. Another typical form of user requirement is mostly interested in modeling specific research assumptions and is often only used for a short time in a specific research context. In general a data model consists of a conceptual part which defines the data semantics, the relevant entities, their features and their relations, and a logical part which expresses a subset of the conceptual model in such a way that it is an abstract, self-contained, logical definition of the data, data operators, and so forth that together make up an abstract machine which makes the data computable. For technical reasons these two parts are usually expressed as two different models in data models of today, but there is no inherent reasons that this will also be the case in the future.

II. An Anatomy of Data Modeling

1. Creating Data Models

The process of creating a data model is well known in the digital humanities as a practical activity, especially in the area of schema development.[12] In contrast with computer science, where 'data modeling' refers almost exclusively to database design and to object-oriented modeling in software systems,[13] [14] the more practically oriented discussion in digital humanities is centered around recurring problems. Some of these are fundamental problems, such as the inherent conflict between standardization and expressiveness, or the challenge of balancing functionally driven and theoretically driven modeling imperatives, or the strategic question of how to determine when a model has been sufficiently tested

[12] See for example Maler and Andaloussi.

[13] The discussion on data modeling is rather advanced in CS; one interesting area is the attempt to describe underlying patterns of data modeling, parallel to describing patterns in programming; cf. for example Silverston and Agnew 2009.

[14] There is a huge amount of literature on data modeling in the computer science meaning created by and addressing the needs of practitioners.

against the target data. Some are problems created by the evolution of the digital ecosystems the objects live in. For example, the existence of multple linguistic corpora, each using a slightly different approach to modeling linguistic features like part of speech, created the need for new models allowing retrieval of the same kind of object independent of its project-specific label. Others are problems arising in connection with specific tools or activities, such as the fact that the use of XML entails the development of workarounds for the problem of overlapping hierarchies, or the question of how to provide for durable annotation of dynamically changing data. In digital humanities, we might say that data modeling is not primarily understood as an activity that takes place prior to building systems, but rather is understood as being deeply involved with a variety of processes and tools that are implicated throughout the life cycle of data. The recent growth in emphasis on data curation has given even greater visibility than before to the idea that data models are created and reworked in the process of tool design, data capture, documentation, interface design, publication, archiving, and reuse.

To advance our understanding of data modeling in digital humanities, then, we should take a closer look at these practical contexts and at the ways in which they reveal opportunities for theorizing these modeling practices more rigorously and explicitly. In the following sections we consider the context of data modeling, the practice of data modeling, forms of notation and documentation and the role they play in shaping the semantics of data models, and finally tools for data modeling.

The Context of Data Modeling

Data modeling in the humanities usually means to work on a type of entity or object that has a history, often a very complex one. The class of documents we term "letters," for example, has a long and manifold history which over time has changed in many aspects, both in the nature of the material object and in the structure of the text. Furthermore, while we can consider a "letter" as a text (in the linguistic sense) and model it as such, there are aspects of layout, use of the writing space, and the arrangement of words on the page that are so important to our understanding of the letter's meaning that we may find it valuable to model it as a graphical object. Data modeling in the humanities is always happening in the context of former attempts to model either these specific classes of objects (for example letters) or the more generic class (texts, graphical objects). Therefore the history of attempts to describe these classes should be known and understood in the community, and indeed constitutes an important strand of expertise for those undertaking the modeling. And this history predates the introduction of digital tools: for example the reflection of philologists about their objects and the establishment of practices to describe them goes back into antiquity. There is thus an important historiographical aspect to modeling, considered as a record of intellectual practice; the discussion of the shortcomings of traditional or digital models (as exemplified, in the workshop, by Allen Renear's discussion of the logical problems of FRBROO) is an important way to contribute to data modeling and digital humanities in general.

All this knowledge about these historical artefacts constitutes the context of all data modeling of these objects. It is one of the main tasks of the humanities to construct, collect and organize this knowledge, and the humanities has amassed a huge amount of knowledge about its objects, and has made systematic attempts at formalization and definition of terms. Sometimes, as in the case of library science, this has taken place in a very formal way with rather exact concepts and the use of controlled vocabularies. In other cases, as for example with the tradition of scholarly editions, this formalization has taken place in a looser sense, operating more to define concepts and practices than to define specific terms with complete precision. At the moment, very little of this knowledge is expressed formally enough to be processed computationally. This is understandable; a formalized knowledge organisation is far from easy to achieve, as we can see if we look at the discussions on FRBR and FRBROO (Functional Requirements For Bibliographic Records 1998) or other proposals to describe systematically even very local areas of this vast landscape under a specific perspective of usage. However, if the digital humanities is to realize the potential of this kind of historical and contextual knowledge as part of our data modeling practice, such formalizations will be increasingly necessary. And any data model in the digital humanities ignoring this knowledge risks overlooking important aspects of the object. This is especially true for what we have called "curation-driven models," because in these cases older models can be regarded as descriptions of user requirements.

On the other hand there is a noticeable gap between the older, traditional, analog-world models and data models of the same objects. One reason is, as we already mentioned, the need for a stricter formalization in order to avoid the ambiguity and polysemy of natural language, which people can resolve so easily in most cases but computer still cannot. Another reason for the gap are the new requirements created by the new abilities and features of digital objects like complex searching or advanced forms of visualization. And last not least relying too much on older descriptions of cultural objects and their features "can perpetuate existing views of data" (Simsion 2007: 71) instead of highlighting new features visible now in the context of the digital media.

The Practice of Data Modeling

The concrete work practices of data modeling received less attention overall in the data modeling workshop. In our summarizing discussion, we noted the complexity of the human and work processes involved in creating the data model, and also the importance of understanding the relationship between the complexity of the data and the process of developing data models, both of which are areas that deserve further attention. But despite the inclusion of case studies (intended to represent the dimension of work practices) these processes were almost invisible in the presentations at the workshop event. This invisibility may have been an effect of the comparative brevity of each presentation: participants may naturally have felt that a detailed account of the work practices and interactions through which data modeling is accomplished in specific project contexts would be out of place or would take too long to narrate. We need to think about how to elicit this information, and also about where it lives (in many cases, it is not even visible in project documentation).[15]

On the other hand data modeling has been a key area for IT professionals in recent decades and a number of books have been published by computer science researchers—and even more by practitioners—which made this an important source of concepts and ideas to all related fields.

[15] There are useful precedents in the field of information science and in particular the study of the development of information systems, e.g. researchers like Bruno Latour or Cathy Marshall.

The practical steps of data modeling have been described repeatedly, usually not in a generic way but in respect to a specific metamodel, as for example in Maler and Andaloussi (1996) for textual models expressed in SGML or Oppel 2009 for database design.[16] Some concepts and ideas seem to be widely shared: for example, that the first step in data modeling is to write up a clear statement of the goals of your project: "The first step in designing a bibliographic system is to state its objectives." (Svenonius 2000: 15) Every aspect of the data model should be related to this set of objectives and this relation works in both ways: Are all objectives covered by the model and are all features motivated by one or more objective? It is a commonly noted problem that scholarly users often don't have clear understanding of their requirements (or have trouble articulating those requirements in terms that map clearly onto the work of data modeling or information design), and the challenge is compounded by the rapidly evolving potential of the digital medium. This is especially true in the humanities where professional socialisation is still strongly shaped by traditions of analogue media.

The following diagram (by a specialist on relational modeling) shows a summary of the data modeling process, as envisioned in the computer science research literature and the literature written by practitioners:[17]

[16] For example Oppel 2009. The example is rather arbitrary.

[17] Taken from Simsion 2007: 35.

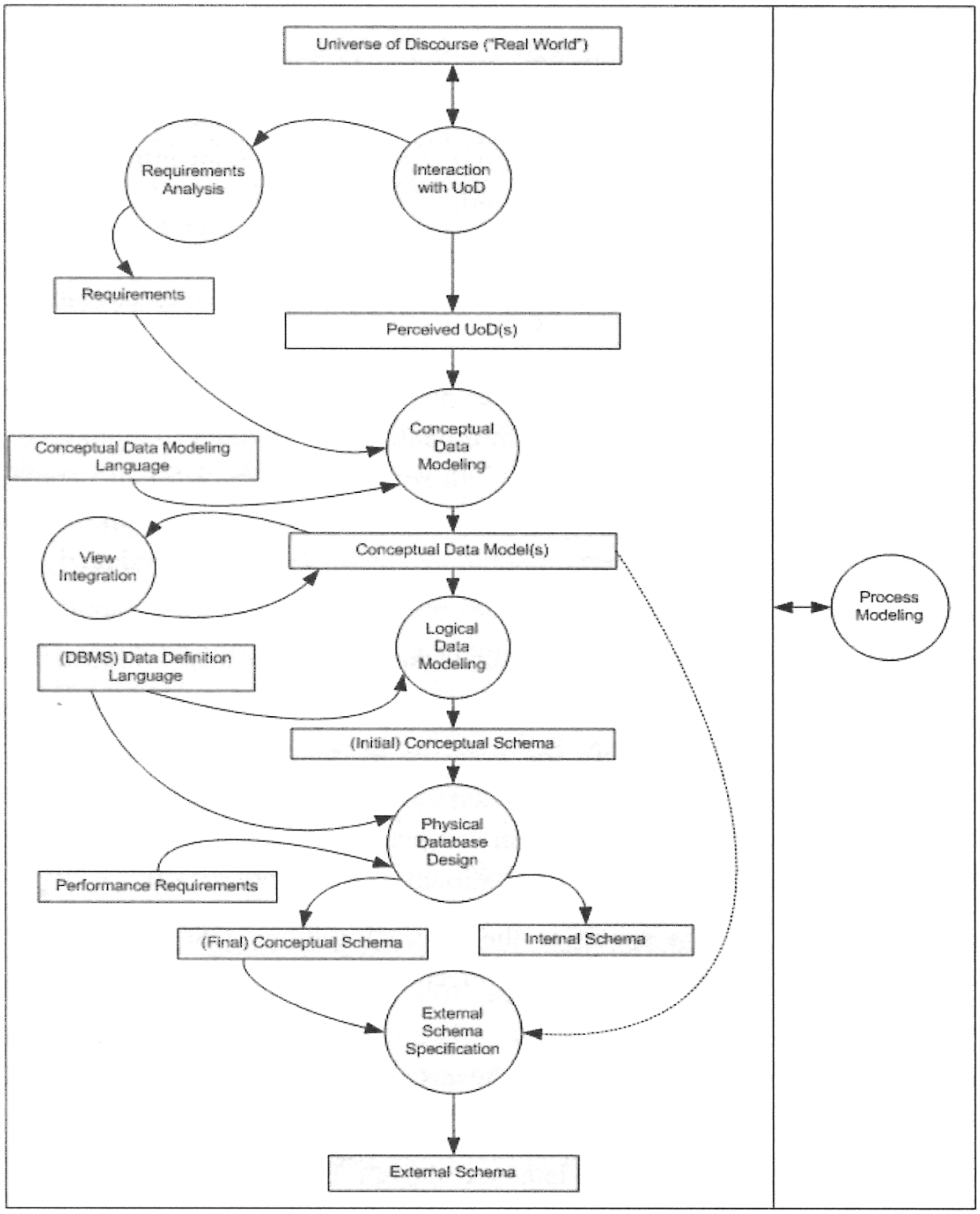


Figure 3-1: Stages in database design – a generic framework

"Universe of Discourse" refers to that part of reality which the data model tries to describe or which it constructs. The different views of this UoD will be expressed in conceptual data models using a conceptual data modeling language, and in the next step these views will be integrated into one conceptual data model. Based on a specific database data definition language this model will then be expressed as a logical model, which in turn will be the basis of the physical database design process resulting in the external schema. Data modeling during all these stages is informed by the process modeling. It is interesting to note that, according to Simsion's overview of the data modeling literature, practitioners use the term 'Conceptual Data Model' to refer to the scoping model, obviously not yet expressed in a data modeling language which mirrors the experience in the digital humanities, that many concepts which one wants to express in a formal data modeling language are so complex that an informal description is the first step to model them.

Though most of the experience reflected in this diagram has been gathered in designing data models for enterprises and businesses, there can be no doubt that digital humanists can learn a lot from this expertise and insight, even if they have to accommodate the principles and ideas to their area of research and modeling.[18] A closer collaboration and a organized transfer of knowledge would probably minimize the need for unnecessary double research and reinventing the wheel. But a closer look into this literature also conveys a warning. In a context where modeling decisions are driven by practical necessity and functional outcomes, the limits of the data modeling language and its notational system may place limits on what such systems can treat as knowable, and these limits may become reified as if they had some epistemological reality. Humanists have a more fluid and much more embedded understanding of their objects, and their research context constitutes the horizon within which knowability must be determined, even where that exceeds the representational capacity of current modeling and notation systems. Even if computer systems cannot currently handle this kind of information structure gracefully, it should be kept in mind that these are the requirements we ultimately want to satisfy.[19]

Notation and Documentation or the Semantics of Data Models

As already mentioned above the formalization in data modeling only affects some aspects of the modeled concepts. Looking at the most common data models in Digital Humanities (databases and XML-encoded texts) it is easy to see that there are three main types of information in the model:

1. Features. In XML this kind of information is usually encoded in elements and attributes, which identify and name features in the data. In a logical database model it is often encoded in the headings of the columns.
2. Relations. In XML this kind of information is usually encoded in the schema (if present), which describes where each element is allowed and which elements it can contain. In an ER diagram, relations are explicitly marked by arrows and their semantics are described using ordinary language. Most of this semantic information is lost in the logical model, which simply expresses the relational structure using primary and foreign keys without saying anything about the meaning of the relationship.
3. Data types. Data types contribute to the semantics by limiting the kind of data which can be input at specific points of the data model. In XML, this information is represented in the schema, which may constrain the data types of both elements and attributes. Data types

[18] A similar opinion is expressed by Ronald J. Murray in his discussion of the shortcomings of the FRBR model (Murray 2008).

[19] Workshop presentations of particular relevance to this section include Muñoz, Stäcker, Kraus, and Piez.

> allow us to understand how to interpret a given piece of data (e.g. as a text string, an integer, a postcode, a call number, etc.) and hence also contribute to more knowledgeable processing.

Usually the largest part of the semantics is stored in the natural language headings, descriptions and definitions coming with the data model. For example it is the main goal of the ODD file format, developed by the TEI, to formalize this relationship between these parts of the data model. Another large part of the semantics can be found in the processing model, especially as it is described by the processing software, for example a set of stylesheets or a database application. (As Stephen Ramsay argued in his presentation at the workshop, "Every data model is asymptotically approaching a processing model.")[20]

Perhaps computer processing will be more intelligent in some future, but as long as each data model is basically an island and the context giving it its semantics is brought in by the humans, data models have to be addressed to two recipients at the same time: humans and machines. This affects its notation: it belongs to the basic principles of XML to use tag names which can be understood by humans and the same is true for the components of database systems. However, this human intelligibility does not currently translate into semantically-aware processing; any awareness of semantics needs to be built explicitly into our processing as a separate matter. In the digital humanities it is an unsolved problem how to integrate the different information about the semantics of a data model, which may be stored in the different places such as encoding, metadata, processing instructions, stylesheets, documentation, and so forth, into one larger survey. The problem is compounded by the fact that XML does not offer anything similar to an ER model through which these relations can be made somewhat explicit.

The situation in digital humanities is essentially the same. Both computer science and digital humanities have to work with notation and constraint, and both domains hit the wall at the same place: everyone has the same problem with semantics. However, the digital humanities feels it more because we deal in semantics above all: because it matters more to us, we feel the problem with greater philosophical urgency. Interestingly, in some areas of humanities (for instance, linguistics, narratology, rhetorical analysis), a structural view to some extent tries to de-emphasize semantics and emphasize structure instead.[21]

Tools

The tools we use for creating and processing data are another place where data models are created and instantiated, and this fact is a revealing source of conflict. Humanists feel intuitively on the one hand that tools ought not to influence intellectual processes (we are not slaves to our tools; our ideas are independent of their material instantiation) and also, on the other hand, that the material circumstances of work (scholarly work, artistic work) do constitute a set of constraints that influence the final product. Digital humanists, and particularly those working in the tradition of text markup and open standards, feel strongly that data should be tool-independent, and that the significant information carried in a

[20] Salminen describes the relationship between the data model and the process model for his use case in Salminen 2003. He models the roles, that is the most essential document users, separately.

[21] Workshop presentations of particular relevance to this section include Stefan Gradmann, Stephen Ramsay, and Trevor Muñoz.

specific data format should be convertible without loss into other formats and should be portable from tool to tool. In the XML world particularly, where concepts like “archival formats,” “transformation,” and “single-source publication” are so deeply entrenched, tools are understood as convenient but temporary carriers whose specific requirements are accommodated by temporary, reversible transformations; adapting the data to the tool (or allowing the tool to exercise a shaping influence on the data) is generally considered an unfortunate and short-sighted compromise. Along similar lines, the deep-seated premise in the XML world that markup should describe information constructs rather than prescribe processing (the “descriptive” vs. “procedural” debate [see Renear 2004]) aligns closely with this preference for tool-independence in our data.

Another useful tool to consider is documentation; although ordinarily considered a process that comes after the fact, documentation (especially in a context where it is tightly integrated with the model as in the case of the TEI ODD) can also be treated in a formative light. In this sense one might speak of “formative documentation” in the same spirit as one speaks of “formative evaluation”); documentation is part of an iterative modeling process that also includes the development of test data and the refinement of models through usage. We can understand the role of documentation as the place where semantics is added to, or anchored securely in, the data model.

2. Evaluating Data Models

Evaluating data models is from one perspective a highly practical matter. In this view data models have to serve functions specified by the user requirements, and the key issue for their evaluation is how well they serve these functions. Success or failure in this case will be closely linked with the effectiveness of communication between a data modeler and the domain specialists who are supposed to work with the model at the end. However, with “curation-driven” or “archival” data models (which are developed with deliberate deferral or generalization of specific user requirements) we face a more complex situation: the dilemma between standardization and expressiveness, or (put another way) the fact that the better a model suits one specific case the worse it will fare in the general case, and vice versa. Given the prevalence of these more “archival” data modeling efforts in the digital humanities, we need to consider how to evaluate such data models in other less practically driven ways.

Data models become more robust the greater the diversity of user requirements being considered in their design; in these cases data models will cover more use cases and will be applicable to more situations. But making provisions for a broad range of specific user needs usually increases either the complexity of the model (the TEI Guidelines offer a striking illustration) or the level of generality at which the model operates (as in the case of a standard like Dublin Core). Though a more complex model will be more likely to cover more of the user requirements of any given project in its domain, there is still a theoretical limit to that likelihood; no model can cover all conceivable needs simply through added complexity. However, resorting to generalization carries its own risks. Very often the more general a model is in its coverage, the higher the probability that users will find it lacking in semantic specificity needed for meaningful communication: the model fails to express differences that are considered essential to the research domain. Interestingly, neither of these limitations prevents users from working with an ill-fitting model: instead, they will try to find workarounds for their problems, either by using a category manifestly meant for a different purpose (“tag abuse”) or will compensate for overly-general semantics through overlaid usage conventions and documentation. These behaviors can themselves be used as a kind of evaluative index to assess the fitness of a model.

While fulfilling user requirements seems to be the most important evaluative criterion, there is also another which we would call tentatively the problem of ‘truthfulness’ or ‘adequacy’ of a model. Most participants of the workshop agreed on the role of social construction, convention, and interpretation in modeling: data modeling involves making explicit an interpretation of an object. The discussion came repeatedly back to questions about levels and types of controversiality in models, and to consideration of domains in which interpretations are more or less widely shared. As already noted, there was a continuum with a radical position on one end (everything is interpretation) and a perhaps equally radical position on the other end (certain things really do exist) and in the middle a position framed around social consensus (interpretation plays an important role but consensus can be achieved on some uncontroversial things). There was significant discussion concerning where things get problematic or controversial (the post-structuralist controversy). On the other hand there seems to be an understanding that in curation-driven data modeling the modeling of an object has to conform to the social construction of this object, and often a fruitful way to access this social construction is a closer look into older codifications of descriptions, such as standards of book cataloguing. In this perspective we evaluate the model not by its truth-value but with respect to how well it captures a shared understanding of the some aspect of the world — independent of more ambitious theories of truth.

But maybe one of the more important insights is the distinction between what we called above “curation-driven” and “research-driven” modeling. These two groups usually follow rather distinct approaches how to determine the user requirements for a data model, how to evaluate the models against these requirements and how they implement procedures to improve the data models. In the case of a researcher just working on his own data, this procedure of improvement may consist of tweaking a schema based on test runs of an analysis, while on the other hand one may have a rather formal way like the Unicode group or the Special Interest Groups of the TEI.

A third important dimension of evaluation for data models can be described as inherent: we are talking about aspects which make a data model more robust in its use over time and more robust in the context of an application. Although in practice a data model will be dependent on aspects of the application which process the data, as already mentioned above a data model has to be designed independent from the application. Sometimes it is difficult to distinguish between the requirements of the users and the requirements of the application which is built turn to satisfy the user requirements. But we need to be especially careful to distinguish information which is needed for the management of a workflow or the data life cycle (i.e. technical metadata) from other aspects of the data model. Thus the processing model and the data model have to be designed separately and the needs of the first should not determine the latter (with the exception of technical metadata).

But the robustness of a data model depends not only on independence of the application and the processing model: like all products of the digital world data models are also always threatened by the rapid evolutions and revolutions of the field. An obvious source of trouble are changes on the level of the operating system (for example the character encoding of the system) or applications on which the data model is dependent, or changes to the metamodel with which the data model is expressed.

On the first glance choosing a data model may look like a primarily technical problem, but it soon becomes evident that it is embedded in social practices and relations. By using the framework of some larger standard like TEI or CIDOC one chooses a specific way of looking at digital objects, a way to discuss and even to evaluate strategies, and also a community of practice for those activities, even if one is not immediately aware of the full implications of that choice. And the reverse is also true: the decision to develop a project-specific modeling approach is also a decision concerning one's relationship to standards and carries with it certain practical and social consequences concerning data longevity, shareability, and so forth. How then can we evaluate modeling approaches when the issues that distinguish them are potentially so slippery and situational? To unpack this point in more detail, as we discovered during the workshop discussion, we need to consider the following questions:

- How do we measure the value of a model: by reference to functionality, intellectual adequacy, conformance to community standards, level of adoption, other factors?
- By what metrics do we measure adherence or conformance to a model? (As an example, the TEI provides definitions of terms like "conformant" or "conformable" which have some clear-cut components and others that are more impressionistic.)
- What kinds of evaluation methods are appropriate for the humanities context? (For instance, what can we learn from the problem of OAIS, where over-complexity and success in one metric meant failure in another metric?)
- How can we perform effective formative assessment as part of our project development, to inform the development and improvement of our models?

3. Using and Applying Data Models

The workshop discussion also touched in significant ways on how data models are used, and what we can learn from that usage. Several topics in particular received special emphasis. The first was the role that data modeling plays as a challenge for the researcher's assumptions: how does the process of data modeling change, reveal, or consolidate one's relation to one's research materials or methods? There are various ways in which this might happen: by making assumptions more explicit; by changing our understanding of language or what aspects of language are perceptible to us; and by changing our understanding of "text" (from product to process). In the discussion, we noted that it would be useful to look at areas where data modeling has actually brought about a change in disciplinary practice, compared with areas where data modeling has heightened our self-consciousness about our practices and made them "more like what they were already" by bringing assumptions to the level of perceptibility that were previously implicit. We also considered whether this might constitute an interesting historiographic point: does such a shift in perceptibility mark an important turning point in the development of a discipline? Furthermore, to the extent that data models permit externalization and examination and discussion of a set of assumptions, it was suggested that these might strengthen transactions between disciplines, and at a higher level between the “arts” and “sciences.” Finally, we also raised the issue of whether data models have an isolating effect, creating a separation—perhaps artificially—between those who should ideally be able to collaborate, by virtue of reifying differences of approach that (if not instantiated formally in a data model) would not really constitute difference at all. If so, perhaps there’s a need to look for ways to move models closer together or avoid unnecessary distance.

A second topic of interest was the question of how we can make effective, useful connections between "adjacent" modeling systems: e.g. TEI, RDF, CIDOC. What are the appropriate levels of interoperability for such systems, and what do we mean by interoperability in this context? It is important to note here that the question of “appropriateness” is specific to the domain of modeling activity: “curation-driven modelers” will have different criteria for assessing appropriate or necessary levels of interoperability from modelers who are driven by more local research goals. One possible position would suggest that to achieve true interoperability we need to make clear the relations between each component in the data model and the functions associated with them. Although we can't do this for the many subjective tags used in the humanities, we can improve interoperability by separating markup from the text and interoperating with each component separately.

There is also a larger question of whether and to what extent we can/should model for the general case. To what extent can we predict future processing environments? To what extent can we anticipate future use cases? These questions require us to pay closer attention to the philosophical differences between the modeling approaches adopted by these user communities (e.g. the “curation-driven” and “research-driven”) to understand whether their differences in themselves pose a structural challenge for humanities data (interoperation, interchange, reuse) in the long term. In other words, if “research-driven users” need to produce what looks like “over-described” data (modeled at too project-specific a level) in order to support their research, what aspects of this data can be usefully repurposed? If “curation-driven users” need to produce what looks like uselessly “under-described” data (so as to support a maximum level of open-ended reuse) what further processes are needed when using this data in real research contexts? In the discussion, one proposed solution to this conundrum (articulated by Desmond Schmidt) was to avoid using markup altogether and focus instead on producing essentially unmodeled data. Others argued that in many domains unmodeled data is a practical impossibility, since even a plain-text transcription must by necessity represent some editorial choices (for example in cases of revision, textual variation, or transcriptional uncertainty). Finally, from a practical perspective, as Elke Teich noted in her presentation, processing pipelines and work processes may produce or require different models of the same object, and it falls to the researcher to handle the different formats and information and to integrate them into one model.

4. Typical Problems of Data Modeling in Digital Humanities

One particularly fertile area of discussion which we deliberately sought out during the workshop was the distinctive problems and challenges for data modeling in digital humanities. A few in particular merit special attention here. The first of these is the limitations imposed by the nature of specific meta-modeling standards: that is, the standards through which we create our models. How do our meta-models affect both the kinds of data processing we are able to do, and also the larger strategic contexts in which that data is used? The case that received the most attention is also the most familiar and long-standing, namely that of XML and its enforcement of a tree structure. As the extensive literature in digital humanities on XML and overlapping hierarchies reveals, this case has special resonance for humanists because it raises larger philosophical questions about the nature of textuality that resonate with other significant debates in a number of humanities domains, notably editorial theory, manuscript studies, and the history of the book. During the workshop, Wendell Piez, Gregor Middell, and Desmond Schmidt spent time in their presentations and comments considering cases where alternative forms (concurrent hierarchies, graph models) could provide a more appropriate and illuminating way of modeling their data.

Another issue of particular interest is the question of how to arrive at a data model to represent objects which are valued precisely for their uniqueness. Creating a data model usually implies that there is a list of attributes common to all or many specific objects which are part of the data set which the model is supposed to describe. On the other hand there is a general feeling in the humanities that many categories of research object—including works of art, archival materials, historical artefacts and the like—are most importantly intelligible as individuals in the sense that there is a unique set of attributes structured by a unique set of relations necessary to describe them adequately. Ontologies may provide a more flexible way to express these attributes and relations in a manner that better reflects how humanists think about such objects.

The question of how humanists think about their objects of study, and how the digital humanities may change that relationship, was also of importance to the discussion. Scholars have been making modeling decisions for a very long time as part of their representation and shaping of research materials. However, in more traditional modes scholars have been somewhat insulated from the contexts in which the operational consequences of those decisions could fully play out: for instance, the domains of library and information science and publishing, where the information originated by scholars is put into circulation in various concrete forms. In the digital humanities (and to a lesser extent in the humanities as now practiced in the digital era), scholars are now in direct contact with these operational domains and in some cases are responsible for overseeing their successful execution. As already noted, the pressure from operational and functional settings upon data modeling decisions is considerable and also a point of potential vulnerability, since this pressure constitutes a motivation that may run counter to the long-range intellectual interests of the scholarly undertaking. From a data modeling perspective, scholars in the digital humanities need training in how to oversee the development of data models (and modeled data) from their most intellectually expressive (“scholarly”) forms through to their most functional and process-oriented forms. The question of how processing models should relate to data models (and the related question of how a priori and ex posto models, top-down and bottom-up models, interrelate) was strongly debated, particularly in the context of the presentations by Wendell Piez and Stephen Ramsay. These issues also impinge upon questions of project development and workflow, inasmuch as they affect how we derive and refine our schemas (from instances? from theories?) and our encoding specifications (from schemas? from disciplinary models? from observed documents?).

5. Teaching Data Modeling

The workshop touched repeatedly on the topic of pedagogy and the issue of how to teach data modeling (or incorporate it usefully into other teaching contexts), devoting one entire panel discussion to it and revisiting it at other points as well. There was agreement that data modeling represents a core competency in digital humanities—albeit so pervasive as to be at times invisible or cloaked under other activities such as scholarly editing, qualitative analysis, manuscript description—and that it merited a larger and more explicit role in digital humanities pedagogy. For one thing, it provides a set of insights that transcend disciplinary boundaries, making it a valuable way of conceptualizing digital pedagogies that need to work outside of traditional disciplinary structures. It also provides a way of looking intelligently and critically at a variety of digital tools and techniques, framing a practically grounded course in a way that also engages scholarly questions. Indeed, framing a "tools" course in this way makes it possible to turn attention usefully away from the specificity of individual tools (an expertise which may date or become obsolete) and towards genres of tools in a way that conveys a higher-level form of mastery. At the same time, because data modeling practices are necessarily situated within specific disciplinary contexts, any discussion of data modeling (particularly in a classroom with multiple disciplines represented) will inevitably yield critical perspective on discipline formation and individual interpretive frameworks and encourage students to formalize their subject knowledge. Training in data modeling also contributes to a number of important professional directions for students, including digital humanities librarianship, consultancy, administration and strategic planning for digital projects, digital humanities pedagogy, and other roles in which a high-level understanding of data structures and strategies for data management is important.

Several points of debate are worth exploring in more detail as the field matures. First, what kinds of practical applications of data modeling best lend themselves to a pedagogical context, to help situate and ground more general theoretical questions? How can students practice data modeling in a classroom setting and how would such activities best be assessed? Second, given that many disciplinary practices are already "doing data modeling" even though they may not be imagined as such, how can we make connections between these less visible forms of modeling and the explicit questions we want to teach? Do any of these disciplinary practices lend themselves to broader use (outside of the specific domain in which they are naturalized)? Finally, given the demand for digital humanities courses that can draw in students from a variety of disciplines, is it in fact feasible to teach data modeling in a way that rises above or abstracts away from discipline-specific knowledge?

6. Contextualizing Data Models

Even if a data model is first of all the technical expression of a set of concepts and their relations, these concepts are usually embedded in larger intellectual contexts, in disciplines and professions. And as we pointed out above, often there is also a community using a specific data model and this community has build up a specific expertise in handling a set of data types and genres. Not the least data models are dependent on tools to express them, visualize them, apply them to data etc. The history of digital humanities is full of stories how the advent of specific tools boosted the use of some data models, for example the publication of the freeware SGML-viewer SoftQuad Panorama arranged by Yuri Rubinsky. And finally data models are created and used in the context of specific institutions, for example research projects, libraries, etc. So even if the most immediate context of an data model usually is a processing model, all the other aspects are importants contexts to data models and data modeling.

Another way to contextualize our data models is to consider their consequences, the impact they have in the world. The sphere of these consequences may in some cases be quite local: for instance, our decision to model our data at all (and if so at what level of detail or rigor) and our choice of models may chiefly affect ourselves and our personal research; it may result in a greater level of self-consciousness concerning how our research affects our representation of research artifacts such as texts. Some of these local consequences may however produce longer-term effects that have a broader impact: for instance, our choice of specific modeling approaches (XML, database, etc.) may have downstream effects on how our data is disseminated, as a result of the kinds of tools we are able to use, which in turn may affect how the data is exposed to others. In the increasingly interdependent ecology of linked open data, the means by which we expose our data (and the predisposition of our data to exposure) may turn out to be among the most consequential choices we make, projected out over time. Similarly, given the rising emphasis on digital repositories and institutional commitments to preserve scholarly research data (on the assumption that it will be reused), what look at first like personal modeling choices may have a profound impact on the overall scholarly ecology as they accumulate. Digital repositories will carry the cumulative effects of individual data modeling practices into the future. They will also carry the presence (or absence) of explicit accounts of those practices, in the form of documentation and schemas of greater or lesser thoughtfulness, into a world where those accounts are the only way we can make sense of past modeling efforts.

Data modeling in the digital humanities can be understood as a process of technical decision-making concerning how best to represent scholarly data to enable specific operations and analysis. However, this process does not happen in a social vacuum, but is embedded in social practice. The political aspects of data modeling are of increasing interest in digital humanities and were an important dimension of the workshop discussion.

A well established data model like the TEI becomes a kind of institution in itself: even if not a formal standard, it functions socially to establish and formalize consensus. This function encompasses both the data model itself—in the case of the TEI a set of schemas plus the guidelines how to use them—and also the supporting organisation, which manages the curation and the development of the data model, and finally a research community of users and programmers using the data model, who are involved in an ongoing discussion process how to bridge the gap between the general concepts found in the guidelines and the specific needs of a project (Jannidis 2009).

The organisation responsible for the development of a data model has a more or less formalized structure and more or less formalized ways to integrate new technical concepts or new user requirements: for example, the ISO working groups or the TEI special interest groups. If a data model has been widely adopted, those adopters may have a vested interest in slowing down the rate of change, while newcomers to the field may be highly motivated to realize their innovative ideas. These institutionalized ways of enabling change must find a safe middle road between two risks: 1) the risk of discouraging newcomers with new ideas and requirements who find the process of change too slow and cumbersome and will turn somewhere else, and 2) the risk of putting off those who have already applied the model to larger amounts of data and see goals like long-term sustainability in danger. So the process of determining which scholarly requirements and technical concepts will find their way into a data model can also be understood as a power struggle which will become fiercer and more complicated the more people adopt it, the more programs support it, and the larger the market around this standard becomes.

Especially for curation-driven modelers, the choice of a data model is very often a challenge of finding the right trade-off between finding a model supporting their needs and using a model which is used by many others. In this context a data model is also the promise to enable some operations not only now but in the foreseeable future, and the fulfilment of this promise becomes more probable the more people use the model. This can be seen as another instance of the "network effect": in networks, an additional user increases the usefulness and value of the network to all who are part of the network (a classic example is the telephone). If more adopt a data model as a standard, then the data model really becomes a standard – which always includes a reduced probability to choose something else. This becomes even more important in communicating with people who usually lack the technical or scholarly understanding to judge the data model but who are responsible for other important goods (such as funding or professional recognition).

Working on the (further) development of data models, especially curation-driven ones, is a new and challenging task for scholars. It demands the ability to understand very clearly not only the needs of your own project but also those of others, and also the ability to formalize and generalize this understanding into a usable data model. Nonetheless this task has yet to be recognized as a contribution similar to writing an essay or producing other traditional forms of scholarly work. In this case the politics of the academic world still have trouble adjusting themselves to the new ways scholars can contribute in essential ways to their discipline.

We also need to consider the ways in which our models reach out (or fail to do so) beyond the academy. To what extent are they visible in non-academic contexts, and to what extent does the non-academic community contribute to or affect modeling practices and decisions? One interesting example is the use of crowdsourcing, which is now common both as a way of accomplishing the primary modeling of source material (as for instance in the well-known cases of the New York Public Library menu transcription project and the Transcribe Bentham project[22]) and as a way of creating or enhancing metadata through annotation, keywording, and other practices. As the two projects just mention demonstrate, crowdsourcing clearly can operate in both a broad-based way that welcomes very wide participation or in a manner that requires more intensive training and support; these approaches naturally entail quite different approaches to the transcription process and to quality assurance. But in both cases, the involvement of the public means exposing a set of modeling assumptions to public view and also testing them against cultural expectations: the NYPL menu project depends on shared familiarity with the ways food is described and consumed, and also historicizes those practices by exposing transcribers to menus from time periods and cultural contexts beyond their personal experience.

Another relevant example is the ways in which metadata developed in an expert context (such as a museum or library setting) now serves as a highly visible point of public access to collections that were formerly accessible chiefly to experts. As materials from museums and archives are digitized and made accessible, the terminology and organizational systems through which those materials were originally modeled are coming under a different kind of scrutiny, and are being asked to perform in functional contexts quite different from those of their original creation. Organizations like the DPLA or Europeana offer a context for rethinking the ways in which metadata performs as a system of cultural meaning and as a mode of access to cultural capital.

7. Research Outlook

This workshop on Knowledge Organization and Data Modeling in Digital Humanities was a first step towards bring data modeling into more explicit view as a topic of digital humanities research. The presentations and discussion showed a substantial amount of existing work in this area, but more importantly identified a number of directions for further research. This is not an exhaustive list, but it is already clear that the digital humanities community would benefit from work in the following areas.

1. **A history of data modeling.** Because of the rapid growth of the digital humanities as a field—and ironically, given that the field now has a history going back at least fifty years—there is generally too little awareness of the history of current debates and research questions, and of the separate disciplinary strands that contributed to the early shaping of the field. Historicizing digital humanities methods is a crucial priority, and a history of data modeling in the fields where its key terms and processes originate would be a significant contribution to this work.

2. **An exploration of uncertainty and how to model it.** This is an area of acknowledged importance. Some of the major data representation systems (such as the TEI Guidelines) include provision for modeling uncertainty in specific domains, notably the representation of transcriptional and editorial uncertainty, but most modeling systems at their core require uncertainty to be represented in the content (e.g. through qualifying notations) rather than in the modeling itself. We also need further exploration of how to use information about uncertainty in the context of

[22] See http://menus.nypl.org and http://blogs.ucl.ac.uk/transcribe-bentham/, respectively.

retrieval, display, and analysis.

3. **A general theory of data modeling**, which treats modeling in the digital realm as a set of activities which share common features, which are embedded into cultural contexts in a similar way and which can be evaluated in similar terms even though the models (relational databases, XML) are markedly different and are based on different mathematical concepts.

4. **The relation between data models and process models** is crucial for the work on the data model and its evaluation — this is a well-known fact in the literature on data modeling and was also discussed during the workshop (for example by Ramsay and Piez). But it is unclear whether this relation can be analyzed on a more detailed level beyond this very general acknowledgement of its importance.

5. **The role of process models vs. the role of data models in digital humanities**. In DH the discussion seems to focus on data models (TEI being an obvious example) while process models seem to be more difficult to generalize. For example, a lot of work has been done on how to model a text (for example Coombs et.al. 1987, Buzzetti 2002, Sperberg-McQueen and Huitfeldt 2011). Other disciplines, for example from the social sciences or natural sciences, seem more interested in modeling processes: for example psychologists modeling the reading process, or physicists modeling processes like the fall of objects. If one looks into books on mathematical modeling most use cases seem to be processes like growth or change (Mooney and Swift 1999). Process modeling in this sense is an attempt to model laws and behaviors as a general case and then derive insight from the ways those models work. However, for the most part humanists don’t believe that such laws can be identified and modeled for the material in the center of our disciplines; to the extent that we look at processes, we view them as individual and specific, and we seek to model them for interpretive rather than predictive purposes. This is obviously related to the more general question of how the humanities constitute their research objects in contrast to those processes in the natural sciences, a difference which has been described by the opposition of nomothetic vs. idiographic (Windelband) approaches. But how does this opposition work in times when the humanities are integrating more empirical methods (for example Moretti 2005; Jockers 2013; Erlin and Tatlock 2014) and at least some of the social and natural sciences are establishing their field of studies beyond a naive realism? In other words, how is data modeling influenced by these more general questions in philosophy of science and epistemology?

6. **Models and notation.** In a time when we know more about the subtle relationship between our thoughts and the tools to express them, it is natural to ask about the relation between models and the notation systems used to represent them. Interestingly enough, in some areas graphical notations have been used widely and successfully, for example crow’s foot notation in relational database design, while in others, for example XML schemas, none of the offerings has found a widespread acceptance, which seems to point to marked differences in the way the models are constructed and communicated.

7. **Modeling tools and humanities research practice.** The case studies presented at the workshop gave a detailed view of data modeling tools and their various functions within digital humanities research practice. But a fuller study of such tools could yield important historiographic and theoretical insight. How have the tools we use for data modeling evolved over time? How do they express existing assumptions and practices? And how do they affect the way we approach the modeling process?

Bibliography

Works Cited

Bauer, Jean. Who You Calling Untheoretical? In: *Journal of Digital Humanities* 1,1 (2011). <http://journalofdigitalhumanities.org/1-1/who-you-calling-untheoretical-by-jean-bauer/>

Burnard, Lou and Syd Bauman (eds.). *TEI Guidelines P 5*. 2007 ff. <http://www.tei-c.org/Guidelines/P5/>

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

Chen, Haitao and Husheng Liao. A Survey to Conceptual Modeling for XML. In: *ICCSIT* 8 (2010), p. 473–477. <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5564759>

Chen, Peter. "The Entity-Relationship Model — Toward a Unified View of Data". In: *ACM Transactions on Database Systems* 1,1 (1976): p. 9–36. <http://dx.doi.org/10.1145%2F320434.320440>.

Coombs, James H., Allen H. Renear and Steven J. DeRose. Markup Systems and the Future of Scholarly Text Processing. In: *Communications of the ACM* 30,11 (1987), p. 933–947.

Date, Christopher J. *SQL and Relational Theory*. Sebastopol: O'Reilly 2012 (2nd edition).

Dillon, Tharam et al. Differentiating Conceptual Modelling From Data Modelling, Knowledge Modeling and Ontology Modeling and a Notation for Ontology Modeling. In: *APCCM '08 Proceedings of the fifth Asia-Pacific conference on Conceptual Modelling*. 79 (2008), p. 7–17. <http://dl.acm.org/citation.cfm?id=1379433>

Erlin, Matt and Lynne Tatlock (eds.). *Distant Readings. Topologies of German Culture in the Long Nineteenth Century.* Rochester: Camden House 2014.

Functional Requirements for Bibliographic Records. Final report. Eds. IFLA Study Group on the Functional Requirements for Bibliographic Records. München: K.G. Saur 1998. <http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf>

Gruber, Tom. A Translation Approach to Portable Ontologies. In: *Knowledge Acquisition* 5.2 (1993), p. 199–220. <http://tomgruber.org/writing/ontolingua-kaj-1993.htm>

Gruber, Tom. Ontology. In: Liu, Ling and M. Tamer Özsu (eds.): *Encyclopedia of Database Systems*. New York: Springer-Verlag 2009. <http://tomgruber.org/writing/ontology-definition-2007.htm>

Hay, David C. Different Kinds of Data Models: History and a Suggestion. In: *The Data Administration Newsletter*. Blog post 1.10.2010. <http://www.tdan.com/view-articles/14400>

Jannidis, Fotis. TEI in a Crystal Ball. In: *Literary and Linguistic Computing* 24.3 (2009), p. 253–265.

Jockers, Matthew L. Macroanalysis. *Digital Methods & Literary History*. Urbana etc.: University of Illinois Press 2013.

Liu, Alan. Where is Cultural Criticism in the Digital Humanities? In: Matthew K. Gold (ed.): *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press, 2012. <http://dhdebates.gc.cuny.edu/debates/text/20>

Maler, Eva and Jeanne El Andaloussi. *Developing SGML DTDs: From Text to Model to Markup*. Upper Saddle River: Prentice Hall 1995. <http://www.xmlgrrl.com/publications/DSDTD/>

Margolis, Eric and Stephen Laurence (eds.). *Concepts. Core Readings*. Cambridge, Mass.: MIT Press 1999.

McCarty, Willard. Modelling. In: W.M.: *Humanities Computing*. Basingstoke: Palgrave Macmillan 2005, p. 20–72.

Mooney, Douglas and Randall Swift. *A Course in Mathematical Modeling*. Washington, DC: The Mathematical Association of America 1999.

Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for a Literary History.* London etc.: Verso 2005.

Murray, Ronald J. The FRBR-Theoretic Library: The Role of Conceptual Data Modeling in Cultural Heritage Information System Design. In: *Proceedings of The Fifth International Conference on Preservation of Digital Objects. iPRES* 2008. p. 172–177. <http://www.bl.uk/ipres2008/ipres2008-proceedings.pdf#page=172>

Oppel, Andy. *Data Modeling. A Beginner's Guide.* New York: McGraw-Hill 2009.

Polikoff, Irene. Ontologies and Data Models - Are They the Same? In: *Voyages of the Semantic Enterprise*. Blog post 30.9.2011. <http://topquadrantblog.blogspot.de/2011/09/ontologies-and-data-models-are-they.html>

Ramsay, Stephen. Databases. In: Susan Schreibman, Ray Siemens and John Unsworth (eds.): *A Companion to Digital Humanities*. Malden / Oxford / Carlton: Blackwell 2004, p. 177–197. <http://www.digitalhumanities.org/companion/>

Salminen, Airi. Document Analysis Methods. In: Miriam A. Drake (ed.): *Encyclopedia of Library and Information Science*. New York / Basel: Marcel Dekker 2003 (2nd edition), p. 916–927.

Schmidt, Benjamin M. Theory First. In: *Journal of Digital Humanities* 1.1 (2011). <http://journalofdigitalhumanities.org/1-1/theory-first-by-ben-schmidt/>

Silberschatz, Abraham, Henry F. Korth and S. Sudarshan (eds.). *Database System Concepts.* New York: McGraw-Hill 1996 (3rd edition).

Silverston, Len and Paul Agnew. *The Data Model Resource Book Vol. 3. Universal Patterns for Data Modeling*.

Indianapolis: John Wiley 2009.

Simsion, Graeme. *Data Modeling: Theory and Practice*. Bradley Beach: Technics Publications 2007.

Sperberg-McQueen, Michael. Classification and its Structures. In: Susan Schreibman, Ray Siemens and John Unsworth (eds.): *A Companion to Digital Humanities*. Malden / Oxford / Carlton: Blackwell 2004, p. 161–176. <http://www.digitalhumanities.org/companion/>

Sperberg-McQueen, Michael and Claus Huitfeldt. Ten Problems in the Interpretation of XML-Documents. In: A. Mehler et.al.: *Modeling, Learning, and Processing of Text-Technological Data Structures*. Berlin: Springer 2011, p. 157–174.

Spyns, Peter, Robert Meersman and Mustafa Jarrar. Data Modelling versus Ontology Engineering. In: *ACM Sigmod Record* 31.4 (2002), p. 12–17. <http://www.sigmod.org/publications/sigmod-record/0212/SPECIAL/2.Meersman.pdf>

Svenonius, Elaine. *The Intellectual Foundations of Information Organization*. Cambridge: MIT Press 2000.

Other titles of interest

Audenaert, Neal and Richard Furuta. Annotated Facsimile Editions: Defining Macro-level Structure for Image-based Electronic Editions. In: *Literary and Linguistic Computing* 24.2 (2009), p. 143–151.

Barnard, David, Ron Hayter, Maria Karababa, George Logan, and John McFadden. SGML-based markup for literary texts: Two problems and some solutions. In: *Computers and the Humanities* 22 (1988), p. 265–276. <http://link.springer.com/article/10.1007%2FBF00118602>

Bauman, Bruce Todd. "Prying Apart Semantics and Implementation: Generating XML Schemata directly from ontologically sound conceptual models." Presented at Balisage: The Markup Conference 2009, Montréal, Canada, August 11–14, 2009. In *Proceedings of Balisage: The Markup Conference 2009*. Balisage Series on Markup Technologies, 3 (2009). doi:10.4242/BalisageVol3.Bauman01. <http://www.balisage.net/Proceedings/vol3/html/Bauman01/BalisageVol3-Bauman01.html>

Bernauer, Martin, Gerti Kappel and Gerhard Kramler. *Representing XML Schema in UML – A Comparison of Approaches*. In: Nora Koch et.al. (eds.): Web Engineering. Berlin, Heidelberg: Springer 2004, p. 440–444. <http://link.springer.com/chapter/10.1007%2F978-3-540-27834-4_54#>

Beynon, Meurig, Steve Russ and Willard McCarty. Human Computing—Modelling with Meaning. In: *Literary and Linguistic Computing* 21,2 (2006), p. 141–157.

Birnbaum, David. *TEI Tables*. Paper presented at TEI 2007 (University of Maryland). Abstract at <http://www.tei-c.org/Vault/MembersMeetings/2007/program.html>.

Birnbaum, David. *In Defense of Invalid SGML*. Paper presented at ACH-ALLC 1997 (Queen's University, Ontario). Abstract available online at the Cover Pages: <http://xml.coverpages.org/birnbaumACH97.html>.

Bradley, John. Documents and Data: Modeling Materials for Humanities Research in XML and Relational

Databases. In: *Literary and Linguistic Computing* 20.1 (2005), p. 133–151.

Breiman, Leo. Statistical Modeling: The Two Cultures. In: *Statistical Science* 16.3 (2001), p. 199–231. <http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.ss/1009213726>

Brüggemann-Klein, Anne, Tamer Demirel, Dennis Pagano and Andreas Tai. "Reverse Modeling for Domain-Driven Engineering of Publishing Technology." Presented at Balisage: The Markup Conference 2010, Montréal, Canada, August 3–6, 2010. In: *Proceedings of Balisage: The Markup Conference 2010*. Balisage Series on Markup Technologies 5 (2010). doi:10.4242/BalisageVol5.Bruggemann-Klein01. <http://www.balisage.net/Proceedings/vol5/html/Bruggemann-Klein01/BalisageVol5-Bruggemann-Klein01.html>

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

Chiarcos, Christian. An ontology of linguistic annotations. In: *LDV-Forum* 23.1 (2008), p. 1–16.

Chiarcos, Christian, Philipp Cimiano, Thierry Declerck and John P. McCrae. Linguistic Linked Open Data (LLOD) — Introduction and Overview. In: Christian Chiarcos, Philipp Cimiano, Thierry Declerck and John P. McCrae (eds.): *2nd Workshop on Linked Data in Linguistics*. Pisa: CEURS 2013, p. i–xi. <http://www.dfki.de/web/forschung/iwi/publikationen/renameFileForDownload?filename=proceedings_LDL.pdf&file_id=uploads_2076>

Coombs, James H., Allen H. Renear and Steven J. DeRose. Markup Systems and the Future of Scholarly Text Processing. In: *Communications of the ACM* 30.11 (1987), p. 933–947.

Conrad, R., D. Scheffner and J.-C. Freytag. XML Conceptual Modeling Using UML. In: A.H.F. Laender, S.W. Liddle and V.C. Storey (eds.): *International Conference on Conceptual Modeling* (ER 2000). Berlin etc.: Springer-Verlag 2000, p. 558–571.

Durusau, Patrick. "How and Why to Formalize Your Markup". In: John Unsworth, Katherine O'Brien O'Keeffe and Lou Burnard (eds.): *Electronic Textual Editing*. New York: MLA 2006. <http://www.tei-c.org/About/Archive_new/ETE/Preview/durusau.xml>

Eide, Øyvind. The Exhibition Problem. A Real Life Example with a Suggested Solution. In: *Literary and Linguistic Computing* 23,1 (2008), p. 27–37.

Finney, Timothy J. Manuscript Markup. In: Larry W. Hurtado (ed.): *The Freer Biblical Manuscripts. Fresh Studies of an American Treasure Trove*. Leiden: Brill 2006, p. 263–287.

Flanders, Julia. Data and Wisdom: Electronic Editing and the Quantification of Knowledge. In: *Literary and Linguistic Computing* 24,1 (2009), p. 53–62.

Forbus, K.D. Qualitative Modeling. In: F. van Harmelen, V. Lifschitz and B. Porter, eds. Handbook of Knowledge Representation. Foundations of Artificial Intelligence. Amsterdam / Boston / Heidelberg: Elsevier 2008, p. 361–394. <http://www.dai.fmph.uniba.sk/~sefranek/kri/handbook/chapter09.pdf>

Freedman, David A. *Statistical Models. Theory and Practice*. Revided Edition. Cambridge: Cambridge

University Press 2009.

Halpin, Terry. *Information Modeling and Relational Databases*. Amsterdam: Elsevier/ Morgan Kaufmann Publ. 2008.

Hooland, Seth van and Ruben Verborgh. Modeling. In: Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish Your Metadata. London: Facet Publishing 2014, p. 11–70.

Hillesund, Terje. Digital Text Cycles. From Medieval Manuscripts to Modern Markup. In: *Journal of Digital Information* 6.1 (2005) <http://journals.tdl.org/jodi/index.php/jodi/article/view/62/65>

Huitfeldt, Claus and C. M. Sperberg-McQueen. What is transcription? In: *Literary and Linguistic Computing* 23.3 (2008), p. 295–310.

Ide, Nancy M. and C.M. Sperberg-McQueen: *Toward a Unified Docuverse: Standardizing Document Markup and Access without Procrustean Bargains* (Submitted to ASIS 97). <http://www.cs.vassar.edu/~ide/papers/asisfmt.html>

Johnsen, Lars G. and Claus Huitfeldt. “TagAl: A Tag Algebra for Document Markup.” Presented at Balisage: The Markup Conference 2011, Montréal, Canada, August 2–5, 2011. In *Proceedings of Balisage: The Markup Conference 2011*. Balisage Series on Markup Technologies 7 (2011). doi:10.4242/BalisageVol7.Johnsen01. Available online at <http://balisage.net/Proceedings/vol7/html/Johnsen01/BalisageVol7-Johnsen01.html>

Lauritsen, Marc and Gordon, Thomas F. Toward a general theory of document modeling. In: *International Conference on Artificial Intelligence and Law archive. Proceedings of the 12th International Conference on Artificial Intelligence and Law table of contents*. Barcelona 2009, p. 202–211. <http://portal.acm.org/citation.cfm?id=1568234.1568257>

Maler, Eve and Jeanne El Andaloussi. *Developing SGML DTDs: From Text to Model to Markup*. London: Prentice-Hall International (UK) 1996. <http://www.xmlgrrl.com/publications/DSDTD/>

Megginson, David. *Structuring XML Documents*. Upper Saddle River: Prentice Hall, 1998.

Mehler, A. et. al. *Modeling, Learning, and Processing of Text-Technological Data Structures*. Berlin: Springer 2011. (Introduction, p. 1–11)

Metzing, D. and Andreas Witt (ed.). *Linguistic Modelling of Information and Markup Languages*. Berlin: Springer 2010.

Miller, John H. and Scott E. Page. *Complex Adaptive Systems. An Introduction to Computational Models of Social Life*. Princeton and Oxford: Princeton University Press 2007.

Murray, Ronald J. The FRBR-Theoretic Library: The Role of Conceptual Data Modeling in Cultural Heritage Information System Design. In: *Proceedings of The Fifth International Conference on Preservation of Digital Objects. iPRES* 2008. p. 172–177. <http://www.bl.uk/ipres2008/ipres2008-proceedings.pdf#page=172>

Ore, Christian-Emil and Øyvind Eide. TEI and Cultural Heritage Ontologies: Exchange of information?

In: *Literary and Linguistic Computing* 24.2 (2009), p. 161–172.

Pierazzo, Elena. A Rationale of Digital Documentary Editions. In: *Literary and Linguistic Computing* 26.4 (2011), p. 463–477. <http://llc.oxfordjournals.org/content/26/4/463.full.pdf+html?sid=d84290fb-7aad-4216-b1e0-5c66957d32c6>

Raymond, Darrell and Frank Tompa. From Data Representation to Data Model: Meta-Semantic Issues in the Evolution of SGML. In: *Computer Standards and Interfaces* 1995 <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=AFE1393A076E2D12C9821749EC3EEFD6?doi=10.1.1.165.568&rep=rep1&type=pdf>

Renear, Allen, David Dubin, C.M. Sperberg-McQueen and Claus Huitfeldt. Towards a Semantics for XML Markup. In: *Proceedings of the 2002 ACM Symposium on Document Engineering*. New York: ACM Press 2002, p. 119–125.

Salminen, Airi, K. Kauppinen and M. Lehtovaara. Towards a methodology for document analysis. In: *Journal of the American Society for Information Science* 48.7 (1997), p. 644–655. <http://www3.interscience.wiley.com/cgi-bin/fulltext/39721/PDFSTART>

Salminen, Airi, Lyytikäinen, V. and Tiitinen, P. Putting Documents into their Work Context in Document Analysis. In: *Information Processing & Management* 36.4 (2000), p. 623–641. <http://dx.doi.org/10.1016/S0306-4573%2899%2900070-9>

Salminen, Airi. Document Analysis Methods. In: Miriam A. Drake (ed.): *Encyclopedia of Library and Information Science*. New York, Basel: Marcel Dekker 2003 (2nd edition), p. 916–927.

Schmidt, Ingrid. Modellierung von Metadaten. In: Henning Lobin and Lothar Lemnitzer (ed.): *Texttechnologie. Perspektiven und Anwendungen*. Tübingen: Stauffenburg 2004, ISBN 3-86057-287-3, p. 143–164. (in German)

Sperberg-McQueen, C. M. and Claus Huitfeldt. GODDAG: A Data Structure for Overlapping Hierarchies. In: King, Peter and Ethan Munson (eds.): Digital Documents: Systems and Principles. Berlin, Heidelberg: Springer 2004, p. 139–160. <http://link.springer.com/chapter/10.1007%2F978-3-540-39916-2_12#>

Sperberg-McQueen, C. M. and Claus Huitfeldt. Ten Problems in the Interpretation of XML Documents. In: A. Mehler et.al (eds): *Modeling, Learning, and Processing of Text-Technological Data Structures*. Heidelberg: Springer 2011, p. 157–174.

Sperberg-McQueen, C. M., Claus Huitfeldt and Allen Renear. Meaning and Interpretation of Markup. In: *Markup Languages: Theory & Practice* 2.3 (2000), p. 215–234. <http://cmsmcq.com/2000/mim.html>

Stachowiak, H. *Allgemeine Modelltheorie [General Model Theory]*. München: Fink 1973. [Basic concepts of the books are presented in this talk: Gelbmann, Gerhard (2002) *An Outline of Pragmatologic Model-Theory (sec. Stachowiak). Semiotic Subjectivity II (lecture).* http://sammelpunkt.philo.at:8080/565/]

How to cite this item:

Julia Flanders and Fotis Jannidis, “Knowledge Organization and Data Modeling in the Humanities.” 2015.

http://www.wwp.northeastern.edu/outreach/conference/kodm2012/flanders_jannidis_datamodeling.pdf

Pid: urn:nbn:de:bvb:20-opus-111270 – Resolver: http://nbn-resolving.de/urn:nbn:de:bvb:20-opus-111270