Modelling in Digital Humanities: An Introduction to Methods and Practices of Knowledge Representation

Franziska Diehr (Prussian Cultural Heritage Foundation)

When talking about 'modelling,' it is not clear what exactly is being referred to. Several disciplines of the sciences, humanities and arts, and especially in the academic fields of philosophy, mathematics as well as computer and information studies are concerned with the question what 'modelling' is and if it serves a general, cross-disciplinary purpose. Models are doubtless used in all of the above fields, but to what extent? Do all pursue the same goal when they create models?

The field of digital humanities as an interdisciplinary movement is experimenting with the application of computational methods to (if not even forcing their incorporation with) the investigation of objects of humanistic inquiry. Which role do 'models' and the act of 'modelling' play in this regard?

McCarty demands that "we need to see it [modelling] as a form of craftsmanship set into the context of scholarship."[1] With this paper, I aim to contribute to the development of a practice of modelling for digital humanities while carefully considering what was discussed about the theory of models in various research areas. Although 'modelling' is often considered to be 'intuitive,' it is far from being an unformalizable process. I argue for 'modelling' to be understood as a scholarly method. Therefore, I investigate methods and practices as well as theoretical foundations of knowledge representation for application in digital humanities. In this paper, I compile practices that can be used to support the application of computational methods to objects which are inherently ambiguous, vague, and full of uncertainties.

1 McCarty, *Humanities Computing* 22.

Theories of Models as Representation of Knowledge

Following Mahr in his observation that 'models' have increasingly been the subject of investigation in various fields since the 1960s, their more general investigation (especially by Stachowiak in 1973) has also led to a broad interest in the process of 'modelling' and its applicability as an interdisciplinary method.[2]

Different formats are used to approach the questions around being a 'model' and the practical applicability of 'modelling.' In their white paper Flanders and Jannidis describe the discussions and outcomes of a workshop on knowledge organization and data modelling. In it they stress that in the context of digital humanities the understanding of 'modelling' is a more general one than data modelling in the sense of computer science: "[…] data modelling is the modelling of some segment of the world in such a way to make some aspects computable […] the task at hand is to define and study the more general concept [of modelling]."[3]

The project "Modelling between Digital and Humanities: Thinking in Practice"[4] researched the term 'modelling' and how it is used in different domains. Its main outcomes are an extensive bibliography as well as publications dedicated to classifying the term 'modelling' in the context of digital humanities.

Methods and techniques regarding the modelling of specific questions are also subject to a broad scope of research papers in the digital humanities. Special issue no. 4 of the German open access journal *ZfdG* on modelling vagueness using graph technologies[5] gives a worthwhile example of how (graph-based) modelling is used for specific use cases in humanities research.

This paper focuses on the aspects of 'model' and 'modelling' that help in their practical usage. A comprehensive overview of (all) available definitions is not given, but an assemblage of certain approaches useful for the application of the method in the context of digital humanities is provided.

2 Mahr, "Information Science" 367.
3 Flanders and Jannidis, "Conference Report" 3.
4 Ciula, "Modelling Between Digital and Humanities."
5 Kuczera et al., "Die Modellierung des Zweifels."

Approaches to Defining 'Model' and 'Modelling'

While discussing different approaches to the definition of the concept 'modelling' in his book *Humanities Computing*, McCarty also presents a more general approach to 'modelling.' He describes it as an epistemological activity in which heuristic methods are used to construct and manipulate models.[6]

As mentioned before, Stachowiak researched the theory of models in a general sense. He characterizes a 'model' with three substantial characteristics: (1) 'Abbildungsmerkmal' (the mapping characteristic), which means that "models are representations of natural or artificial originals." (2) 'Verkürzungsmerkmal' (the reduction characteristic) explains that "models do not have all the attributes of the originals they represent." (3) 'Pragmatisches Merkmal' (the pragmatic feature) lays out that a model is the outcome of an intentional process. It can only represent things that are based on and influenced by the pragmatic view of the modelling subject.[7]

In his approach, Mahr focuses on a model's purpose and determines it with the following three aspects: (1) a model is a 'thing for itself:' it is an entity on its own like a picture, a text, a set of rules. (2) A model is 'about something:' it represents characteristics and/or experiences of an object by selecting, generalizing, and correlating them in a new form. (3) A model is 'for something:' it serves a specific purpose. When put into use, 'knowledge' can be gained from the model.[8] Or as Hughes explains it: "from the behavior of the model we can draw hypothetical conclusions about the world over and above the data we started with."[9] But can we trust our models? Mahr clarifies that a model does not claim to hold any truth, "but rather forms of demonstrability, possibility, and choice."[10]

Following Geertz, who also attributes to models 'aboutness' and 'purposefulness' (a model 'of' and model 'for'),[11] McCarty also distinguishes these two aspects: "[it is] either a representation of something for the purposes of study, or [it is] a design for realizing something new."[12]

6 McCarty, *Humanities Computing* 24.
7 Stachowiak, *Allgemeine Modelltheorie* 132–133.
8 Mahr, "Das Wissen im Modell" 11–12.
9 Hughes, "Models and Representation" 331.
10 Mahr, "Information Science" 365.
11 Geertz, *The Interpretation of Cultures* 93.
12 McCarty, *Humanities Computing* 24.

'Aboutness' and 'purposefulness' together make up the model's functionality as a 'tool for knowledge creation.' Simplifying Mahr's description,[13] a model incorporates the derivation of the initial object as well as the hypothesis which premised the model's build (Stachowiak's *Pragmatisches Merkmal*). Thus informed, the model offers new modes of observation, which fundamentally differ from those which are applicable (or not applicable) to the examination of the represented object itself.[14] The information extracted from this observation can result in the extension or creation of knowledge. Or as Hughes put it "[…] if we examine a theoretical model […], we shall achieve some insight into the kind of representation that it provides."[15]

Taking those general reflections on the concept of 'model' into account, the following is focused on one particular aspect: the model's functionality of representing knowledge. Understanding what is meant by abstracting the original object and what implications Stachowiak's *Pragmatisches Merkmal* has on the model as a tool of knowledge creation is crucial when actually performing the act of 'modelling.'

Coming from the field of artificial intelligence, Davis et al. define the term 'knowledge representation' by identifying five characteristics: (1) "a knowledge representation is a surrogate, a substitute for the thing itself"—a model of something.[16] (2) A knowledge representation describes the domain the object lives in. It offers a set of "ontological commitments" explaining relevant aspects concerning the world of the object. (3) A knowledge representation is a "fragmentary theory of intelligent reasoning," meaning that the rules of logic are used to make sense of the excerpt of the world under study. (4) A knowledge representation is used for analysis with computational methods in a pragmatic way. (5) A knowledge representation is "a medium of human

13 Mahr, "Information Science" 376.

14 For example, when observing phenomena in the world of physics (e.g. quantum physics), there are no techniques applicable for the observation of the initial object because models are used to prove the existence of the object in the first place. Mahr uses the discovery of the Higgs particle to illustrate how knowledge is created by the use of models, while also explaining the interaction between the aboutness and purposefulness of a model.

15 Hughes, "Models and Representation" 329.

16 This understanding is comparable with the *Abbildungsmerkmal* defined by Stachowiak as well as the second aspect determined by Mahr—"a model of something."

expression," and in this regard also a way for human beings to express and discuss knowledge.[17]

An even more concrete approach is formulated by Silberschatz et al. when they specifically define the term 'data modelling' as applied in computer science: "A collection of conceptual tools for describing data, data relationships, data semantics, and consistency constraints."[18]

Reflecting on all presented perspectives on 'model' and 'knowledge representation,' a practical approach is needed that can be used as a guideline when creating computational models for digital humanities research. "Modelling makes the semantics of 'knowledge objects' explicit and transfers them into a data structure."[19] In this sense 'modelling' can be defined as the method used to construct a machine-readable model that represents objects in a pragmatic perspective. Constructing such a model entails being aware of this and therefore reflecting on the definitions and methods that are applied to the object and domain. Modelling is a process of building an ontological understanding. In this perspective, modelling can be seen as a hermeneutic method used to gain knowledge about something that is to be studied further using computational methods.[20]

The Practice of Modelling

Using this approach as a guideline, the following part of the paper illustrates the practical aspects of modelling.

Flanders and Jannidis observed two types of motivation for modelling in (the broader field of) digital humanities. On the one hand there is 'curation-driven modelling,' where authorities develop models and metadata schemes for the purpose of making information findable, accessible and usable.[21]

As libraries are information providers to the public and to scholars, they have pioneered curation-driven modelling. In contrast to museums and archives, libraries are far more advanced in the use of computational practices for representing knowledge about their collections. They work

17 Davis et al., "What is a Knowledge Representation?" 17.
18 Silberschatz, *Database System Concepts* 8.
19 Sowa, *Knowledge Representation* 132.
20 Diehr, "Modellierung von Entzifferungshypothesen."
21 Flanders and Jannidis, "Conference Report" 4–5.

constantly on the optimization of cataloguing practices and as well as the description of (non)bibliographical resources, such as authority files.

The curation of authority files became an especially important task which digital research relies on as well. Using persistent identifiers for the identification and disambiguation of entities like persons, organizations, places, and works has become an indispensable task for libraries. Arguing that the classification and identification of entities is also a task of knowledge representation, in this perspective curating authority files can be seen as a form of curation-driven modelling.

Another example of a library effort of curation-driven modelling is the METS/MODS standard.[22] While MODS is a schema for describing bibliographical items such as books and manuscripts, METS provides a mechanism to combine the structural features of the item with its digital facsimile. Using these standards in combination, libraries provide a useful service for research. Users do not have to visit the library in person in order to examine the contents and appearance of a book, for instance. Cross-institutional and long-term digitization efforts like the German 'VD 18'[23] made it possible for libraries to provide an enormous amount of their book collections as digital facsimiles with structured metadata.[24]

On the other hand, there is 'research-driven modelling,' where (individual) researchers develop models to study objects and domains that are subject to specific research questions. The outcome of this modelling process is either a project-specific adaptation of an existing standard, or a genuinely novel model, for instance a domain-specific ontology. As this paper focuses on research-driven modelling, the following lays out how to model research-driven knowledge representations.

As pointed out at the beginning of this paper, the act of modelling can be formalized. In doing so, a resilient description of the process is needed. In *Knowledge Organization and Data Modelling in the Humanities*, Flanders and Jannidis describe three steps of data modelling: (1) 'conceptual modelling' is concerned with "the identification and description of the entities and their

22 "MODS standard Version 3.7." https://www.loc.gov/standards/mods/; "METS standard Version 1.12." https://www.loc.gov/standards/mets/

23 *Verzeichnis Deutscher Drucke des 18. Jahrhunderts – VD18* (*Index of German prints of the 18th century*) funded by Deutsche Forschungsgemeinschaft (DFG). https://gso.gbv.de/DB=1.65/

24 An example for a digital VD 18 collection, see "VD18 digital." https://gdz.sub.uni-goettingen.de/collection/vd18.digital

relationship in the 'universe of discourse'"—attributes and relations of object and domain are to be defined. This step can be viewed as the most challenging but also the most interesting part of the whole modelling process. (2) 'Logical data modelling' refers to the process of 'translating' the conceptual model into an actual data structure, a syntax like XML or "defining the tables of a database according the underlying relational model." (3) 'Physical data modelling' is the actual implementation as a data model in a digital environment that enables functionalities for data storage and querying as well as "optimization of the database for performance."[25] This paper focuses on the first two steps.

Requirement Analysis

The conceptual modelling step should be preceded by a definition of what is required of the model. By means of the requirements analysis, information is acquired that is needed for the construction of the model. This may seem like a trivial task, but is in fact an intensive phase of revisiting the object of inquiry and the specific views of the domain it 'lives' in as well as methods, traditions, and constraints which are brought into hypothesis building.[26] In this section, practices are presented that are useful for gaining the necessary information as well as structuring and analyzing it.

Before going into detail regarding the methods, I would like to present a set of prototypical questions that can be used as a starting point to gain information about the object and its domain in each of the presented methods as well as other possible formats of information acquisition:

> Which attributes characterize the object? Which categories describe the objects regarding its theme/subject?
> Which features of the object are crucial for answering the research question?
> How can the nature of relation to other materials, actions, actors, places etc. be described? How can it be structured?
> Which sources of knowledge are consulted? Where are they derived from?
> Which methods were formerly applied to the object?
> Where does the object 'live'? Is there a specific 'ecosystem' or domain?

25 Flanders and Jannidis, "Conference Report" 3.

26 Diehr, "Modelling Vagueness."

> From which pragmatic perspective is the object described? Which (disciplinary) domain-specific perspectives are relevant?

The 'interdisciplinary workshop' is a practice that is especially useful if many different perspectives on the object of inquiry are needed, for instance as a kick-off for a research project or as an initial step into a new field of interest. In this scenario, a variety of experts from different domains share insights on how the object of inquiry is regarded in their field and how research challenges are approached with respect to the field's tradition and current practice. As a creative input session or trigger for a fruitful discussion, this practice also risks information overload. If resources are available to manage the post-processing workload (such as extracting key issues from a video recording), modelling will hugely benefit from this approach.

A common method of information acquisition described by Pickard[27] and Reinhold is the 'expert interview.' The person (or team) in charge of the actual modelling interviews (several) domain experts using a prepared questionnaire. As well as the previous described practice, "expert interviews can be used to gain a high degree of context knowledge, which is of crucial importance when modelling domain-specific knowledge."[28] But the challenge with this method is that the interviewer already has to have deep insights into the domain in order to be able to prepare and ask relevant questions.[29] Thus this practice can be recommended if the modelling person (or team) and the interviewees are both equally familiar with the field of research. Also, it is a suitable information acquisition method if a focus on the research questions is already set and other perspectives are ruled out.

This paper presents a new method that partially integrates procedures from the expert interview. In comparison to the previously described methods, this method aims at much more detailed results. It is highly integrated and also functions in dialogue with software development practices.

As the names suggests, the 'iterative questionnaire' integrates the questioning of experts in an ongoing process. On the basis of an initial examination of the object and its domain a draft is conceived. This is then presented to a fixed group of domain experts, who work together in the upcoming phase of questioning rounds. The first evaluation will show where weaknesses in the questionnaire lie: How can object attributes be described even more clearly?

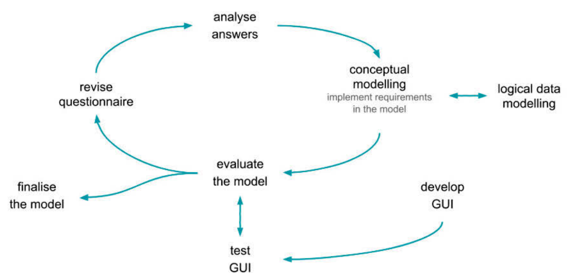
27 Pickard, *Research Methods in Information*.
28 Reinhold, "Das Experteninterview als zentrale Methode" 330.
29 Ibid.

Which relations to adjacent topics come to light? This is followed by a specification of the question catalogue, in which certain questions will result in more refined answers, while others will allow associative thoughts.

While this phase is key to conceptual modelling, ideally the advanced questionnaire is also used to specify the functionalities of the software components: e.g. which field type is required to describe a specific object attribute, how to establish relations between objects, and so on.

This illustrates that requirement analysis is not an isolated task; it is accompanied by the actual modelling process (Figure 1). After an initial phase of information aggregation, an iterative process of questioning, modelling, and evaluating will be initiated. The model will be finalized when evaluation finds no more irregularities. Because it is hard to communicate the progress of the modelling process to those who are not modelling themselves, the parallel development of a graphical user interface (GUI) is useful. Even a prototyped GUI with reduced functionality and a sparse design communicates the underlying model much better than any diagram of the model itself could (in most cases). This makes it much easier to display the modelling progress and make the results of the modelling comprehensible.

Figure 1: The iterative process of modelling.



The 'iterative questionnaire' aims at gradually asking more precise and specific questions so that a high degree of accuracy is achieved at the end of the process. Compared with the previous methods, this one is more time consuming, but it also has key benefits: a high level of precision and accuracy is reached and the developed model will be most fitting to the specific research needs. This method also contributes significantly to the hermeneutic process: Because experts work collaboratively on the questionnaire and discuss the subject regularly, knowledge transfer processes are activated. That helps to reflect on common practices and traditions of disciplinary views. In this perspective the method may lead to new insights on the research object, and it also may contribute to novel approaches to methods for studying it.[30]

Conceptual Modelling

The act of conceptual modelling can be described as a heuristic performance: from a pragmatic perspective, assumptions are made about things and how they work. But how were those assumptions made in the first place and how are they to be formalized into a model? In this section, I approach the methodology of conceptual modelling as a heuristic practice that is used to represent knowledge.

Conceptual modelling aims at developing an ontological understanding through the explicit description and definition of objects, their relationship to each other, and to their domain. By acquiring information on object and domain through the requirement analysis, a knowledge base was gained. In order to formulate a conceptual model that formalizes this knowledge into a workable and computable model, that which Sowa calls "ontological categories"[31] must firstly be defined. Defining those categories means performing an "intentional activity that has scope, granularity, and (perhaps multiple) levels of abstractions" as well as providing "explicit semantics."[32] This definition is particularly difficult when it comes to vague and uncertain information, regarding which Sowa points out: "any incompleteness, distortions, or restrictions in the framework of categories must inevitably limit the generality of every program and database that uses those categories."[33] That is why the process of conceptual modelling should be undertaken with an acute awareness of the pragmatic perspectives that are applied to the model.

In their presentation of *A Reference Framework for Conceptual Modelling*, Delcambre et al. refine the act of conceptual modelling even further, distinguishing between a 'conceptual model' as a product of modelling and 'conceptual model language' as the grammar used to formulate the model. In their argument a conceptual model refers to a domain-specific model or ontology and

30 Diehr, "Modelling Vagueness."
31 Sowa, *Knowledge Representation* 51.
32 Delcambre et al., "A Reference Framework" 30.
33 Sowa, *Knowledge Representation* 51.

the conceptual model language to a foundational ontology, a reference model used to specify more or less generic concepts.[34]

Conceptual modelling is a process of identification and categorization. Attributes of objects and relations between them and their domain are identified and formally described and categorized. In their paper *Toward a Methodology for Building Ontologies*, Uschold and King reflect on the modelling process—and specifically upon the act of categorization—when formulating a conceptual model (in this case an ontology).[35] Therefore they consult Lakoff's theory of categorization, in which he outlines three aspects that are used when categorizing reality:

> "(1) 'Basic-level categorization:' The idea that categories are not merely organized in a hierarchy from the most general to the most specific but are organized so that the categories that are cognitively basic are in the middle of a general-to-specific hierarchy. Generalization proceeds upward from the basic level and specialization proceeds downward.
> (2) 'Basic-level primacy:' The idea that basic-level categories are functionally and epistemologically primary with respect to the following factors: gestalt perception image formation, motor movement, knowledge organization, ease of cognitive processing (learning, recognition, memory etc.) and ease of linguistic expression.
> (3) 'Reference-point, or "metonymic" reasoning:' The idea that a part of a category (that is, a member or subcategory) can stand for the whole category in a certain reasoning process."[36]

It was Rosch who first introduced the concept of "basic-level categories or prototype theory."[37] As it is not pertinent to the case being made here, I do not want to go into much more detail; it should be sufficient to point out that a basic-level category can be understood as a prototype of a real thing, the first 'mental image' or concept that comes to mind when observing something. Lakoff illustrates this as follows: [38]

The attributes applied to the superordinate will be inherited by all following sub-categories and therefore be true for them. An attribute of a subordi-

34 Delcambre et al., "A Reference Framework" 31.
35 Uschold and King, *Towards a Methodology for Building Ontologies*.
36 Lakoff, *Women, Fire and Dangerous Things* 13.
37 Rosch et al., "Basic Objects in Natural Categories."
38 Lakoff, *Women, Fire and Dangerous Things* 46.

SUPERORDINATE	ANIMAL	FURNITURE
BASIC LEVEL	DOG	CHAIR
SUBORDINATE	RETRIEVER	ROCKER

nate is a specified attribute of a superordinate (like a chair for 'seating' as a specification of furniture which 'support human activities') or a novel feature that is not applicable to any of its superordinates (a rocking chair that 'swings back and forth').

For the practice of conceptual modelling this can be used as a blueprint: The subordinate could be used to describe an instance of a category, e.g. 'Ludwig Wittgenstein.' Depending on the domain (which specifies what is of main interest), the basic-level category could be 'philosopher' or 'logician,' and the superordinate 'person.' The next step is to assign attributes and relations to the basic level and superordinate, using inheritance to build an ontological understanding that makes more and more specific statements the further down the hierarchy it proceeds. In this regard, 'instances' can be understood as the data onto which the model is to be applied.

When designing a model that comprises such an ontological understanding of the domain, there is always a danger of incorporating too many attributes and/or categories into the model. As the model should be designed to answer hypotheses which were made prior to the modelling, it should be limited to "minimal ontological commitment;" Gruber explains further that

> "[...] an ontology should make as few claims as possible about the world being modeled, allowing the parties committed to the ontology freedom to specialize and instantiate the ontology as needed. [...And it should define] only those terms that are essential to the communication of knowledge consistent with that theory."[39]

Following Gruber's advice in practice can cause discussions about the model's comprehensiveness: experts may anticipate theories and further questions that could be applied onto the model in future. But bearing in mind Stachowiak's *Pragmatisches Merkmal*, a model is designed at a specific time for a specific reason from a certain perspective. Future research may have other

39 Gruber, "Toward Principles for the Design of Ontologies" 3.

methods at hand or, in the meantime, new or extended knowledge could have been gained, outdating the model's use. Thus, there is no need to overextend the current model with assumptions of possible applications.

Using Reference Models and Standards for Logical Data Modelling

In order to evaluate if the semantics were defined with the required care, Delcambre et al. advise mapping the constructed model to another (referential or standardized) model.[40] This step can be also understood as a part of logical data modelling in which the conceptual model is transferred into a machine-readable syntax. In the following sections, I illustrate how the model to be constructed can be aligned with existing models (e.g. in the form of metadata standards or reference ontologies) and how this can be used to transfer it into a machine-readable data model.

While also providing semantics for the description of objects and their domains, metadata standards[41] often come written in a machine-readable syntax, such as XML or RDF. The reuse of existing metadata standards does not only provide meaningful concepts and a vocabulary for the alignment of the constructed model but can also aid with logical data modelling. Using the same syntax as the referential metadata standard makes its application comfortable. However, another syntax can obviously always be used, as data can be converted into other data formats that may be more fitting to the model.[42]

To illustrate this, I use the project-specific application of the guideline of the Text Encoding Initiative (TEI)[43] for the development of the text corpus of the project *Text Database and Dictionary of Classic Mayan*[44] as an example.

40 Delcambre et al., "A Reference Framework" 34.

41 Durrell, *Data Administration*: "Metadata are structured, coded data that describe the characteristics of information-carrying entities for the purpose of identifying, researching, evaluating and managing the entities described."

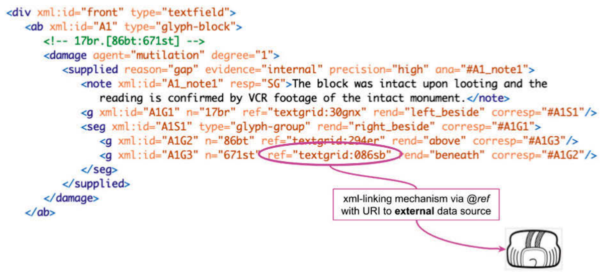
42 The right syntax to formulate the model is based on the structure of the conceptual model. If the structure is shown to be quite simple, there is no need to reach for an elaborate syntax: the rule of thumb can be 'form follows function.'

43 *TEI: Text Encoding Initiative*. https://tei-c.org/

44 The project *Text Database and Dictionary of Classic Mayan* is funded by North Rhine-Westphalian Academy of Sciences, Humanities and the Arts (Project No. I.B.17). Its goal since 2014 is the compilation of a dictionary for the script and language of Classic Mayan by creating a machine-readable corpus of all known inscriptions and codices. The author was part of the team from 2014–2018 and was responsible for the con-

The TEI Guidelines are a de facto standard for text-based research, particularly for editorial scholarship.[45] A huge community is applying those guidelines and also constantly improving them. The TEI does not insist on using XML as syntax for applying the guidelines, but many TEI-related frameworks and transformation routines rely on XML as the syntax of choice. The TEI guidelines provide semantics for describing a variety of characteristics of the textual material, for instance elements like <p> (paragraph) and <l> (line) can be used to describe the structure of a text. There are also elements to describe the condition of the text carrier like <damage> as well as editorial interventions with elements like <supplied> and <note>.

Figure 2: Excerpt of a TEI document from the corpus of the Text Database and Dictionary of Classic Mayan showing the encoding of a hieroglyphic text using the TEI Guidelines and the XML syntax.



XML as a syntax consists of elements and attributes which specify the elements: e.g. <supplied> is specified by @reason="gap," @evidence="internal." The attribute @ana contains a reference to a XML identifier, which is used as a mechanism for linking elements in a document or even across documents (which is referred to as 'stand-off markup').[46] In Figure 2 the value of @ana in

ceptual development of the virtual research environment, including the realization of metadata schemas and ontologies. For more information see *Textdatenbank und Wörterbuch des Klassischen Maya*. https://mayawoerterbuch.de/

45 Along with the TEI, there are several initiatives specifying the guideline for special domains such as *EpiDoc* (subset of the TEI for epigraphic purposes). https://sourceforge.net/p/epidoc/wiki/Home/; *CBML Comic Book Markup Language*. http://dcl.slis.indiana.edu/cbml/; *MEI: Music Encoding Initiative*. https://music-encoding.org/

46 "20.4 Stand-off Markup." https://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html#NHSO

the <supplied>-element refers to a <note> with @xml:id="A1_note," in which the editor explains why it was possible to reconstruct a damaged glyph.

XML also provides a further linking mechanism using @ref, which also allows linking to an external data source. A particularity of the text corpus of Classic Maya is that it does not contain transcripted text.[47] Due to ongoing classification and deciphering tasks, it is not possible to provide transliterations for the hieroglyphic text. Instead, Uniform Resource Identifiers (URI)[48] are used to refer to the description of the used sign variant, which is stored in another database using another metadata schema (an ontology written in RDF).

This example illustrates how existing vocabularies and standards can assist with logical data modelling. Authorities curating data (like libraries) as well as research-driven communities (such as the Text Encoding Initiative) have developed a great amount of useful vocabularies, which can be useful when developing a project-specific data model. The challenging task may be to find the suitable ones, as available resources collecting those vocabularies (like Linked Open Vocabularies)[49] still lack comprehensiveness.

Challenges of Modelling Humanities Data—An Example

A modelling process needs sensible attention and a high degree of awareness. In the humanities we deal with complex knowledge. We deal with interpretations based on prior processes that generated knowledge at a certain time by pre-informed subjects. This is utterly different from how computational methods work, because algorithms are limited to exact assertions. McCarty states that "programmatic explicitness and precision is radically inadequate for representing the full range of knowledge."[50] Modelling means to be aware of what can be ignored and why. A model can only represent what it is built for, and further methods applied to the model depend on that. This is especially important when dealing with vague and uncertain information.

Since 'knowledge' about objects can be questioned or interpreted differently, it can be necessary to represent the various levels of knowledge in the

---

47 For further information on the encoding strategy of the text corpus of Classic Mayan see Sikora, "Interlinked."

48 "Uniform Resource Identifier (URI)." https://tools.ietf.org/html/rfc3986

49 *Linked Open Vocabularies*. https://lov.linkeddata.es/dataset/lov

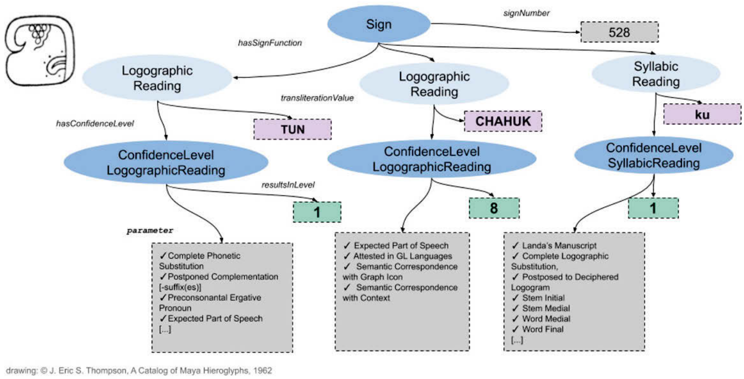
50 McCarty, *Humanities Computing* 25.

model in order to counteract distortions and to limit the knowledge base precisely in the sense of the defined ontological categories.[51]

To illustrate the complexity that has to be dealt with in digital humanities projects, I again use an example from the project *Text Database and Dictionary of Classic Mayan*.

Classic Mayan is an as-of-yet not fully deciphered script of the Maya culture, which inhabited pre-Columbian Yucatán. Due to the nature of the hieroglyphs, one sign can have multiple readings. These readings depend on the co-text the sign is used in, but can also be influenced by other factors, such as the type of text carrier it is written on, its temporal or spatial usage etc. Additionally, there are contradicting hypotheses on the deciphering of signs due to different approaches to decoding as well as the availability of material. The project's goal is to linguistically analyze the texts with respect to all reasonable hypotheses. A need arose for a way of formally assessing the level of confidence of the reading proposals. For that, a system was modelled that combines ontological modelling[52] and propositional logic (Figure 3).

Figure 3: Modelling of the criteria-based system for the qualitative assessment of reading proposals of the Text Database and Dictionary of Classic Mayan.



51 Diehr, "Modelling Vagueness" 37.

52 *Text Database and Dictionary of Classic Mayan*. https://classicmayan.org/

Based on rules of logic, the system assesses a qualitative level of confidence.[53] This helps with the linguistic analysis and ultimately contributes to the ongoing deciphering of the script.

The approach of the *Text Database and Dictionary of Classic Mayan* project also illustrates the possibilities that emerge when different data models are combined. It can also serve as a prototypical example for other applications. It is a concept of separate but interlinked ontological-based data models (Figure 4).

Each model fulfills a different purpose: encoding the text corpus, linguistically annotating and analyzing text, classifying linguistic signs, referencing bibliographical resources, and documenting text carriers. Together they form a system that provides rich functionalities that enable epigraphers and linguists to decipher the Classic Mayan script.[54]

This case may illustrate a specific case for the purpose of deciphering an ancient script, but it also serves as an example of dealing with diverse and complex materials and sources. Instead of documenting texts and their carriers, a model can be imagined, one that is used for researching complex entanglements of music and media history, describing and relating audiovisual documents as well as source materials.

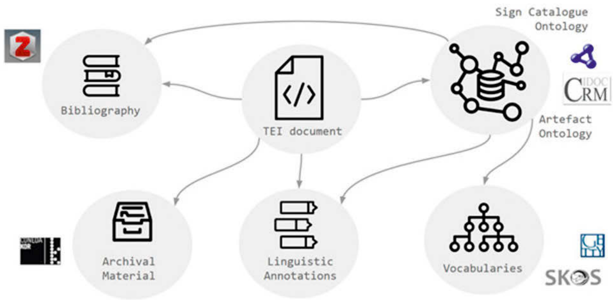
Conclusion

This paper laid out theoretical foundations and practical approaches to modelling (complex) knowledge for the purpose of research and analysis with the help of computational methods. As has been argued, modelling is regarded

53 Diehr, "Modelling Vagueness" 40.

54 The TEI/XML document forms the central part. Its encoding is enriched by multiple other resources to support specific functions and workflows for annotation, documentation and analysis: a Sign Catalogue for the classification of signs and graphs as well as their variants; an ontology for the documentation of the text carriers (both of them aligned with CIDOC CRM and written in RDF and supported by further knowledge organization systems, which are modelled in SKOS/RDF); the tool ALMAH for linguistic annotation and analysis; a project bibliography (for which Zotero is used); documentation and organization of archival material (which is managed by the DARIAH-DE service ConedaKOR). Bringing all those information sources together provides a holistic research environment for analyzing and deciphering the script of Classic Mayan.

Figure 4: The virtual research environment of the project Text Database and Dictionary of Classic Mayan uses multiple models and information sources.



as a hermeneutic method that builds an ontological understanding of the object of inquiry and the domain perspective(s) it has been regarded with. Or as Ciula and Eide put it: "In digital humanities we do not only create models as fixed structures of knowledge, but also as a way to investigate a series of temporary states in a process of coming to know."[55]

As illustrated, special attention should be paid to the modelling of vague and uncertain information. Unwanted bias built into the model can cause false observations. In contrast, the careful modelling of phenomena, the vagueness and uncertainties of which are anticipated, may enable one to see things from new and different perspectives and to draw new conclusions.

The flexibility of computational models makes it possible to create and/or reuse models and combine them to fit specific needs. In doing so, holistic research environments can be developed. Nevertheless, the development of specialized models to address specific needs is recommended. A 'one model fits all' solution will not satisfy highly specialized research tasks and specific research interests.

As an interdisciplinary movement, digital humanities bridges the gap between computer science and traditional humanities studies. Particularly in the field of knowledge representation and the theory and practice of modelling, digital humanities should play a major role in training in the skills that

55 Ciula and Eide, "Reflections on Cultural Heritage" 37.

are needed for (not just) computer-aided but, first and foremost, data-driven approaches to humanistic research. The skill set should be grounded on a theory that comprises aspects of philosophy and logic, knowledge representation and engineering, as well as information and computer science.

In their survey Delcambre et al. asked experts from different fields about their ideas on (conceptual) modelling. Their participants particularly stressed the necessity of human involvement in the modelling process and thereby strongly emphasized a cognitive presence when creating knowledge. Nowadays, more and more tasks are being solved by artificial intelligence and machine learning algorithms, which are often referred to as 'black boxes' because the internal learning processes of the machines are not transparent to us. Nevertheless, they perform tasks quickly and produce satisfactory results. Those are profound arguments in favor of making use of such methods. When applying them to humanities research, we should carefully take into account the extent to which human intelligence and the processes of knowledge creation are needed when modelling. In this perspective, digital humanities may contribute to the discourse on digital ethics, critically examining the application of artificial intelligence for sensitive tasks of research as well as for general usage by humanity.

Bibliography (last accessed 16 April 2020)

Ciula, Arianna, et al. "Modelling Between Digital and Humanities: Thinking in Practice." *Digital Humanities 2016: Conference Abstracts – DH 2016*, Krakow, Poland (July 2016): 762–763.

Ciula, Arianna, and Øyvind Eide. "Reflections on Cultural Heritage and Digital Humanities: Modelling in Practice and Theory." *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, Madrid, Spain (May 2014): 35–41. DOI: 10.1145/2595188.2595207

Davis, Randall, et al. "What is a Knowledge Representation?" *AI Magazine* 14, 1 (Spring 1993): 17–33. DOI: 10.1609/aimag.v14i1.1029

Delcambre, Lois M. L., et al. "A Reference Framework for Conceptual Modelling." *Proceedings of the 37th International Conference on Conceptual Modelling – ER 2018*, Xi'an, China (October 2018): 27–42.

Diehr, Franziska, et al. "Modelling Vagueness – A Criteria-Based System for the Qualitative Assessment of Reading Proposals for the Deciphering of Classic Mayan Hieroglyphs." *Proceedings of the Workshop on Computational*

Methods in the Humanities 2018 – COMHUM, Lausanne, Switzerland (June 2018): 33–45.

Diehr, Franziska, Sven Gronemeyer et al. “Modellierung von Entzifferungshypothesen in einem digitalen Zeichenkatalog für die Maya-Schrift.” *Zeitschrift für Digitale Geisteswissenschaft*, special issue 4 (January 2019). DOI: 10.17175/sb004_002

Durrell, William R. *Data Administration: A Practical Guide to Data Administration*. New York: McGraw-Hill, 1985.

Flanders, Julia, and Fotis Jannidis. “Conference Report.” *Knowledge Organization and Data Modelling in the Humanities*, Brown Univ. Providence, USA (March 2012): 1–38. URN: urn:nbn:de:bvb:20-opus-111270

Geertz, Clifford. *The Interpretation of Cultures: Selected Essays*. London: Fontana, 1993 (1973).

Gruber, Thomas R. “Toward Principles for the Design of Ontologies Used for Knowledge Sharing?” *International Journal of Human-Computer Studies* 43, 5–6 (November 1995): 907–28. DOI: 10.1006/ijhc.1995.1081

Hughes, R. I. G. “Models and Representation.” *Philosophy of Science* 64 (December 1997): 325–336. DOI: 10.1086/392611

Kuczera, Andreas, Thorsten Wübbena and Thomas Kollatz. “Die Modellierung des Zweifels: Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten.” *Zeitschrift für digitale Geisteswissenschaften*, special issue 4 (January 2019). DOI: 10.17175/sb004_013

Lakoff, George. *Women, Fire and Dangerous Things: What Categories Reveal about the Mind*. Chicago: Univ. of Chicago Press, 2008 (1987).

Mahr, Bernd. “Das Wissen im Modell.” *KIT-Report* 150 (2004): 1–21.

Mahr, Bernd. “Information Science and the Logic of Models.” *Software and Systems Modelling* 8, 3 (May 2009): 365–83. DOI: 10.1007/s10270-009-0119-2

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

Pickard, Alison. *Research Methods in Information*. London: Facet, 2007.

Reinhold, Anke. “Das Experteninterview als zentrale Methode der Wissensmodellierung in den Digital Humanities.” *Information-Wissenschaft & Praxis* 66, 5–6 (October 2015): 327–33. DOI: 10.1515/iwp-2015-0057

Rosch, Eleanor, et al. “Basic Objects in Natural Categories.” *Cognitive psychology* 8, 3 (July 1976): 382–439. DOI: 10.1016/0010-0285(76)90013-X

Sikora, Uwe, et al. “‘Interlinked!:’ Schriftzeugnisse der Klassischen Mayakultur im Spannungsfeld zwischen Standoff- und Inline-Markup in TEI-XML.” *DHd 2019 Digital Humanities: Multimedial & Multimodal: Konferenzabs-*

tracts, Frankfurt, Germany (March 2019): 143–47. DOI: 10.5281/zenodo.2596095

Silberschatz, Abraham, et al. *Database System Concepts*. [Place of publication not identified]: McGraw-Hill Education, 2019.

Sowa, John F. *Knowledge Representation: Logical, Philosophical and Computational Foundations*. London: Brooks Cole, 2000.

Stachowiak, Herbert. *Allgemeine Modelltheorie*. Wien, New York: Springer, 1973.

Uschold, Mike, and Martin King. *Towards a Methodology for Building Ontologies*. Edinburgh: Univ. of Edinburgh Press, 1995.

Online Sources (last accessed 16 April 2020)

CBML Comic Book Markup Language. http://dcl.slis.indiana.edu/cbml/

CIDOC CRM. http://cidoc-crm.org/

CLARIAH-DE. https://www.clariah.de

EpiDoc: Epigraphic Documents in TEI/XML. https://sourceforge.net/p/epidoc/wiki/Home/

Linked Open Vocabularies. https://lov.linkeddata.es/dataset/lov

"Linking, Segmentation, and Alignment. Stand-off Markup". *TEI: Text Encoding Initiative*. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html#SASO

MEI: Music Encoding Initiative. https://music-encoding.org/

"METS standard Version 1.12." *Library of Congress*. https://www.loc.gov/standards/mets/

"MODS standard Version 3.7." *Library of Congress*. https://www.loc.gov/standards/mods/

"Ontology of the Sign Catalogue for Classic Mayan." *Text Database and Dictionary of Classic Mayan*. https://classicmayan.org/documentations/catalogue.html

Text Database and Dictionary of Classic Mayan. https://classicmayan.org/

Textdatenbank und Wörterbuch des Klassischen Maya. https://mayawoerterbuch.de/

"Text Encoding Initiative: P5 Guidelines." *TEI: Text Encoding Initiative*. https://tei-c.org/guidelines/p5/

"20.4 Stand-off Markup." *TEI: Text Encoding Initiative*. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html#NHSO

"Uniform Resource Identifier (URI): Generic Syntax." *IETF Tools*. https://tools.ietf.org/html/rfc3986
"VD18 digital." *SUB: Göttinger Digitalisierungszentrum*. https://gdz.sub.uni-goettingen.de/collection/vd18.digital
Verzeichnis Deutscher Drucke des 18. Jahrhunderts – VD18. https://gso.gbv.de/DB=1.65/