Text encoding and ontology—enlarging an ontology by semi-automatic generated instances

Amélie Zöllner-Weber
The Wittgenstein Archives (WAB), University of Bergen, Norway

Abstract

The challenge in literary computing is (1) to model texts, to produce digital editions and (2) to model the meaning of literary phenomena which readers have in their mind when reading a text. Recently, an approach was proposed to describe and present structure and attributes of literary characters (i.e. the mental representation in a reader's mind), to explore, and to compare different representations using an ontology. In order to expand the ontology for literary characters, users must manually extract information about characters from literary texts and, again manually, add them to the ontology. In this contribution, I present an application that supports users when working with ontologies in literary studies. Therefore, semi-automatic suggestions for including information in an ontology are generated. The challenge of my approach is to encode aspects of literary characters in a text and to fit it automatically to the ontology of literary characters. The application has been tested by using an extract of the novel '*Melmoth the Wanderer*' (1820), written by Charles Robert Maturin. For the main character, Melmoth, 72 instances were generated and assigned successfully to the ontology. In conclusion, I think that this approach is not limited to the theme of character descriptions; it can also be adapted to other topics in literary computing and Digital Humanities.

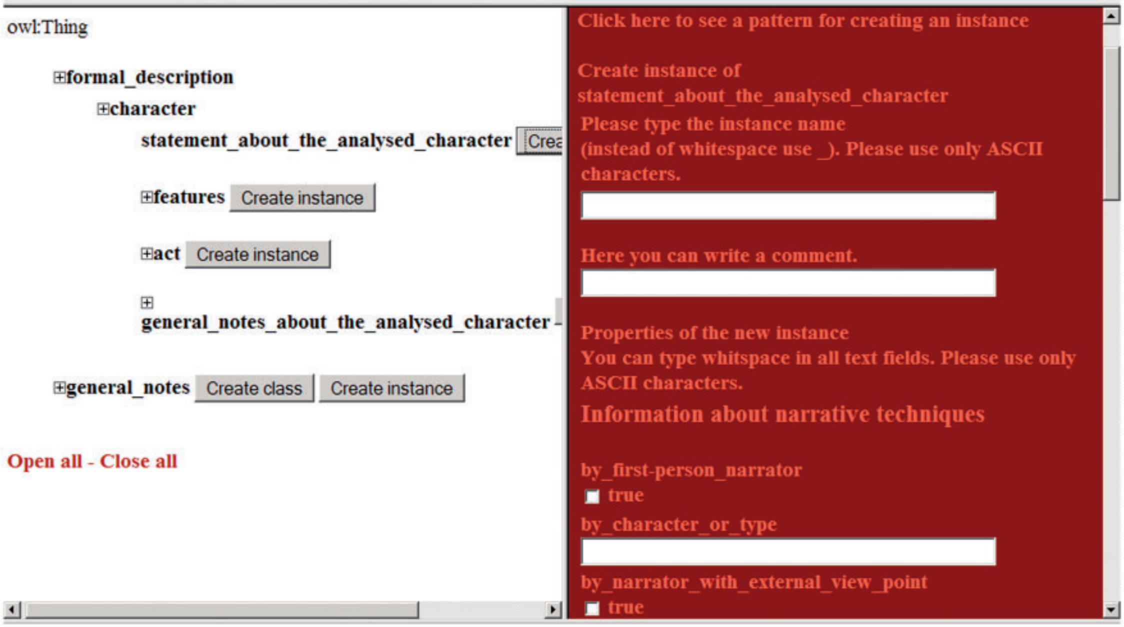
Correspondence:
Amélie Zöllner-Weber,
The Wittgenstein Archives
(WAB), Department of
Philosophy, University of
Bergen, Sydnesplassen 12,
Pb 7805, N-5020 Bergen,
Norway
E-mail:
amelie.zoellnerweber@gmail.com

1 Introduction

The challenge in literary computing is (1) to model texts, to produce digital editions (Meister, 2003; Jannidis *et al.*, 2006) and (2) to model the meaning of literary phenomena which readers have in their mind when reading a text. However, since text structures and mental representations often differ from each other, it is difficult to model them in the same way. In addition, interpretations depend on the cultural and social background of individual readers. It is, therefore, a challenge to create a model of these interpretations which will enable the development of descriptions, e.g. of literary characters so that these characters can be compared.

For this problem, an ontology can be used. Only a few ontologies in literature studies exist. One such ontology, proposed by Lawrence *et al.*, covers narrative elements like events in literature (Lawrence and Schraefel, 2005). Zöllner-Weber (2007) uses an ontology to model mental representations in order to realize descriptions of literary characters. There, the character is regarded as a complex cognitive entity in the reader's mind. When using the ontology, readers can describe their own mental representations of literary characters in computer processable form. The goal of this approach is (1) to describe and represent the structure and attributes of literary characters, i.e. the mental representation in the reader's mind, (2) to

doi:10.1093/llc/fqr021 Advance Access published on 16 May 2011



Click here to see a pattern for creating an instance

Create instance of
statement_about_the_analysed_character
Please type the instance name
(instead of whitespace use _). Please use only ASCII
characters.

Here you can write a comment.

Properties of the new instance
You can type whitspace in all text fields. Please use only
ASCII characters.
Information about narrative techniques

by_first-person_narrator
□ true
by_character_or_type

by_narrator_with_external_view_point
□ true

Fig. 1 Extract of the client-server system. Here, users can manipulate the ontology, e.g. enlarging it by including information; yet, this has to be done manually

explore, and (3) to compare these different representations.

In order to expand the ontology for literary characters, users must manually extract information about characters from literary texts and, again manually, add them to the ontology (Fig. 1).

The contribution will present an application that will support users when they are working with ontologies within literary studies. A tool has been developed which takes the user's mark-up of a text and generates semi-automatic suggestions of instances to be added to the ontology. It is intended for users who are more familiar with text encoding than ontologies, and who are interested in topics of literary studies, especially literary characters.

By combining an application of text encoding with the ontology, the problem of manual manipulation of the ontology should be reduced. There have already been some approaches, e.g. HyTex (Lüngen and Storrer, 2007), dealing with texts, text encoding, and a knowledge base. However the challenge of the approach presented here is (1) to encode aspects of literary characters found in a text—which often have to be interpreted by the reader and (2) to match these automatically with the ontology for literary characters.

2 Methods

For the description of literary characters, an ontology that models characters by their mental representation was used (Zöllner-Weber, 2006). An ontology is a model especially developed to provide organisation and retrieval of information semantically. There are two comprehensions of the term of 'ontology': the pure formalistic approach of classifying objects as in the field of Artificial Intelligence (AI) and machine learning, and the more transcendental approach in the humanities and especially philosophy (Zöllner-Weber, 2009). In the following, the AI definition of ontology is used.

Several theories of literary characters are combined to create a basic description, e.g. features, actions, and speech performed by literary characters, using an OWL ontology (Nieragden, 1995; Grigoris and van Harmelen, 2003; Jannidis 2004). The main assumption of this approach is based on the theories of mental representations of characters proposed by Jannidis (2004) and Schneider (2000). The reader's representations of characters are complex, cognitive processes which are guided by literary texts; however, their mental structures do not often correspond with the texts.

I chose an ontology because its structure fits well with the mostly hierarchical structures of theories proposed to describe or analyse literary characters (Lotman, 1977; Jannidis, 2004). In general, an ontology is a hierarchy of 'classes'. The classes, in turn, contain instances that represent 'individuals'. 'Properties', which contain additional information, are attached to the individuals (Noy and McGuinness, 2001). By using this kind of structure, information is formally described.

The framework of the mental representations is modelled by the main classes of the ontology, e.g. inner and outer features, actions on other characters and objects. The sub classes contain characteristics of special characters (special features of single characters or special characteristics of groups of characters). I decided to include single pieces of information that are taken from literary texts which are then entered as instances of the ontology. Instances of the classes represent direct information about a character given in a text. Properties contain additional information, e.g. type of narrator, author, annotation information, or reference to literature. Through the information provided by the class hierarchy, the instances and their properties, a mental representation of a character is modelled (cf. Fig. 2). Additionally, each instance is linked to a text part or text parts where the information is built upon. In this approach, individual description, which is a preparatory step for interpretation, is focussed upon. The main description categories secure a general classification making it possible to compare two different interpretations of one character, even when they are spread over different categories of the ontology.

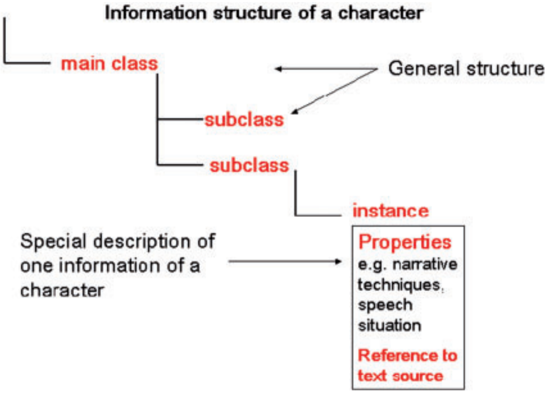

Information structure of a character
main class
General structure
subclass
subclass
instance
Special description of one information of a character
Properties
e.g. narrative techniques, speech situation
Reference to text source

Fig. 2 The structure of the ontology for literary characters: the main classes form the general structure of characters, whereas the subclasses contain special features. An instance represents a single piece of information about a character that is obtained from a literary text. Properties contain additional information, e.g. about narrative techniques

In order to fill the aforementioned ontology of literary characters in a more automated fashion, an encoding scheme has been developed. For the annotation, I compiled a TEI-DTD (Text Encoding Initiative 2003, http://www.tei-c.org/) using the TEI-DTD Roma server (http://www.tei-c.org/Roma/) consisting of the prose, the drama, the verse, and the analysis module. Elements of this encoding scheme were then utilized and rearranged so that it became viable for literary studies. This means that the usage of elements was enlarged by adding single, already existing TEI attributes to these elements. By using this special mark-up, a user can directly add interpretive pieces of information about a literary character to a text. Here, the annotation scheme is based on four main categories, 'description', 'statement', 'action', and 'speech', which classify pieces of information. All descriptions about a character that are stated by a narrator are subsumed under 'description'. The category 'statement' depicts commentaries by a character about another character. To mark non-verbal and verbal actions of a character, the categories 'action' and 'speech' should be used. In addition, a user should add, example, information about the type of narrator, the name of a character, and depending on the chosen category, additional information to complete the annotation. After the process of encoding, a user sends the marked texts via a web form to a server where the annotations are evaluated by an in-house developed programming algorithm (cf. Fig. 4). The algorithm tries to sort the encoded information about the character based on the four categories. The value of the attribute ‘ana’ is matched to the categories which also reflect the main classes of the ontology, and, thereby, the encoded information is assigned to the major classes of the ontology. If further encoded information is given by the mark up, the algorithm tries to generate a further subclassification. This is repeated until all available encodings are processed. Figure 3 depicts an example of this process. After successful processing, a user is presented with a list for all processed annotations that can form instances. Additionally, for all of these suggestions, a class assignment is also given.

In addition, the user is supported by an extracted list of surrounding classes of the ontology so that (s) he is able to inspect the environment of the new instance and its class. If a new instance requires a new class that does not yet exist, a user can also add a new class. Afterwards, (s) he can assign the instance to the new class.

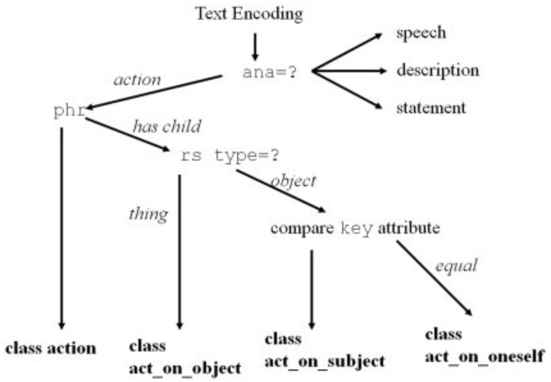


Fig. 3 Scheme of decision algorithm for adding instances to the classes of the ontology. The different nodes of the tree reflect the tags used in the encoding scheme, the leaves of the tree marked in bold depict possible matching classes; here the most right one (‘act_on_oneself’) will be the final assignment

3 Results

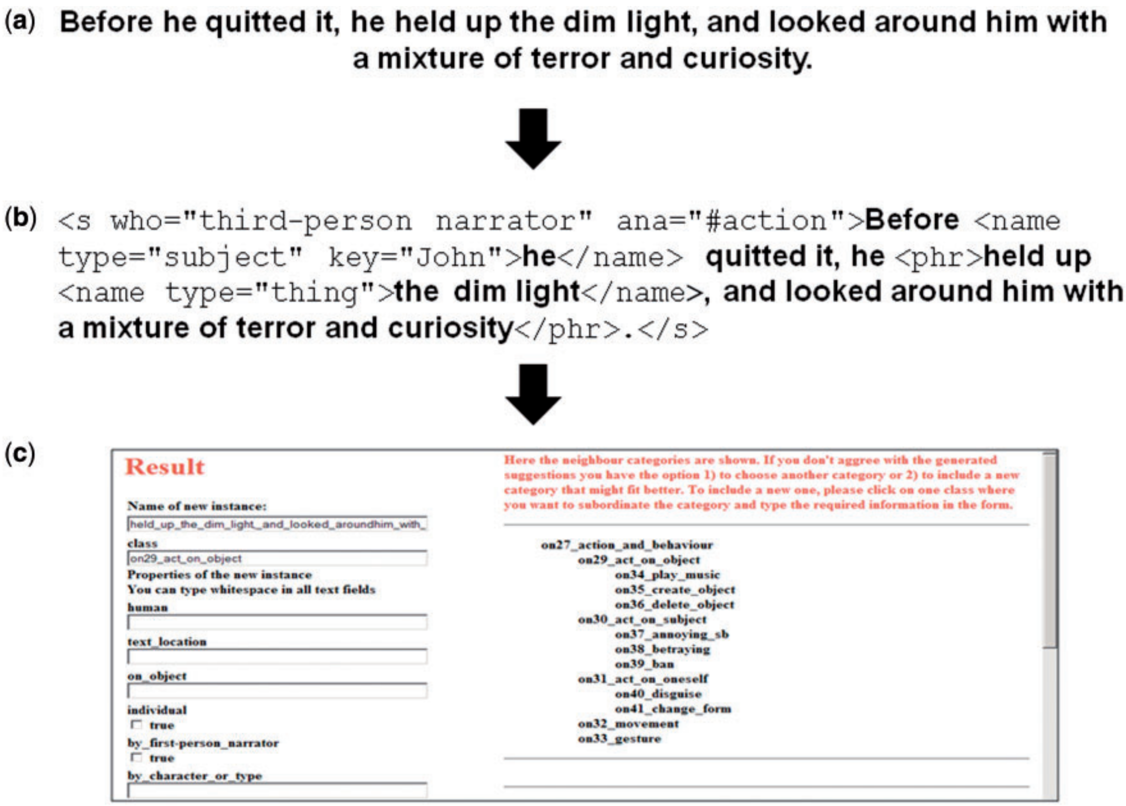
The application has been tested by using an extract from the novel ‘*Melmoth the Wanderer*’ (1820), written by Charles Robert Maturin. This novel was part of a larger study and comparison of evil and devil characters. Briefly, the literary task was to investigate how these characters differ between authors and how they developed over time. A more detailed outline of the findings of this study is given in Zöllner-Weber (2008).

I encoded the text with the mentioned TEI scheme and afterwards, by using the programming algorithm, I obtained suggestions for new instances. In Fig. 4, the process of generating an instance from a text passage is shown as an example. For the main character, Melmoth, seventy-two instances were generated and assigned to the ontology.

4 Discussion and Conclusion

In this contribution, a system has been presented which semi-automatically inserts information into the ontology for literary characters using a TEI encoding. Using this application, it is possible to add information about a single character to the ontology, as well as simultaneously process annotations for several characters. Time and work can be saved, as the whole text can be annotated at once and then transferred to the ontology. There is no need to go back and forth between text and ontology as one would have to do for a purely manual insertion of character information into the ontology. In addition, the suggested tags can be used not only for the purpose of the proposed ontology, but also for linguistic or other aims within literature studies.

Working with the client-server system of the ontology, it was observed that literary scholars had problems to capture the aspects of an ontology in general. But they did not have problems to detect text parts where information about a character is given and they could formulate and categorize this information. Text encoding is rather close to the working process of detecting and describing aspects; however, the complexity of the underlying



Before he quitted it, he held up the dim light, and looked around him with a mixture of terror and curiosity.

```
<s who="third-person narrator" ana="#action">Before <name
type="subject" key="John">he</name> quitted it, he <phr>held up
<name type="thing">the dim light</name>, and looked around him with
a mixture of terror and curiosity</phr>.</s>
```

Result

Name of new instance:
held_up_the_dim_light_and_looked_aroundhim_with_

class
on29_act_on_object

Properties of the new instance
You can type whitespace in all text fields

human

text_location

on_object

individual
□ true

by_first-person_narrator
□ true

by_character_or_type

Here the neighbour categories are shown. If you don't aggree with the generated suggestions you have the option 1) to choose another category or 2) to include a new category that might fit better. To include a new one, please click on one class where you want to subordinate the category and type the required information in the form.

- on27_action_and_behaviour
  - on29_act_on_object
    - on34_play_music
    - on35_create_object
    - on36_delete_object
  - on30_act_on_subject
    - on37_annoying_sb
    - on38_betraying
    - on39_ban
  - on31_act_on_oneself
    - on40_disguise
    - on41_change_form
  - on32_movement
  - on33_gesture

Fig. 4 The process of generating an instance for the ontology from an encoded text: (**a**) original part of the text, (**b**) encoded text, and (**c**) derived instance with the class assignment

ontology model seems to block this 'intuitive' process.

Therefore, using this approach, a user does not have to focus on technical details of the ontology and (s) he can concentrate on the literary aspects and their encoding or description. Furthermore, the result page of this application always presents the surrounding environment of a single suggestion, the instance, e.g. possible sibling classes and parent or neighbour classes (cf. Fig. 4). If a user is unsure, where to assign the instance, (s) he can inspect and compare descriptions of other users in the client-server-system.

Recently, several approaches to link encodings in TEI to ontologies were proposed. A mapping system from TEI to the CIDOC–CRM ontology, an ontology developed to describe cultural heritage information (CIDOC, 2003), has been provided by TEI Ontologies Special Interest Group (Ore and Eide, 2009). In comparison to the CIDOC–CRM system, suggestions for instances in the presented approach are generated from (a more linguistic) encoding rather than by mapping tags that appear in both, text and ontology. Another approach uses an ontology to represent the domain of the fine rolls of Henry III (Ciula *et al.*, 2008). Information taken from TEI documents were transposed by XSLT into the ontology. Additionally, ontologies and their applications are often linked to logic reasoning. However, incorporating such techniques into the present application might be difficult especially for untrained users, as shown elsewhere (Zöllner-Weber, 2009).

Therefore, I think that it is important to relate texts, encodings, and ontologies to each other so that the same information, which appears in these different resources, can be linked. This encourages users, who are familiar with encodings but not necessarily familiar with an ontology or logic reasoning, to add new information to an ontology without going too deeply into details. This approach is not limited to the theme of character descriptions; it can also be adapted to other topics in literary computing and Digital Humanities.

References

CIDOC. (2003). *Definition of the CIDOC Conceptual Reference Model*/Produced by the ICOM/CIDOC Documentation Standards Group, continued by the CIDOC CRM Special Interest Group. ISO/DIS 21127. http://www.cidoc-crm.org/docs/cidoc_crm_version_4.2.1.pdf (accessed 4 April 2011).

Ciula, A., Spence, P., and Vieira, J. M. (2008). Expressing complex associations in medieval historical documents: The Henry III Fine Rolls Project. *Literary and Linguistic Computing*, **23**(3): 311–25.

Grigoris, A. and van Harmelen, F. (2003). Web ontology language: OWL. In Staab, S. and Studer, R. (eds), *Handbook on Ontologies*. Berlin: Springer, pp. 67–92.

Jannidis, F. (2004). *Figur und Person - Beitrag zur historischen Narratologie*. Berlin: Gruyter.

Jannidis, F., Lauer, G., and Rapp, A. (2006). Hohe Romane und blaue Bibliotheken. Zum Forschungsprogramm einer computergestützten Buch- und Narratologiegeschichte des Romans in Deutschland (1500–1900). In Lucas, M. G., Loop, J., and Stolz, M. (eds), *Literatur und Literaturwissenschaft auf dem Weg zu den neuen Medien*. Bern: germanistik.ch.

Lawrence, K. F. and Schraefel, M. C. (2005). OntoMedia - Creating an Ontology for Marking up the Contents of Fiction and Other Media. In 1st AKT Doctoral Colloquium, *14 June 2005*, Milton Keynes.

Lotman, J. M. (1977). *The Structure of the Artistic Text*. Michigan: University of Michigan Press.

Lüngen, H. and Storrer, A. (2007). Domain ontologies and wordnets in OWL: modelling options. *Zeitschrift für Computerlinguistik und Sprachtechnologie*, **22**(2): 1–17.

Meister, J. C. (2003). trans. Matthews, A. *Computing Action. A Narratological Approach*. Berlin, New York: Gruyter.

Nieragden, G. (1995). *Figurendarstellung im Roman: Eine narratologische Systematik am Beispiel von David Lodges Changing Places und Ian McEwans The Child in Time. HORIZONTE – Studien zu Texten und Ideen der europäischen Moderne*. Trier: Wissenschaftlicher Verlag.

Noy, N. F. and McGuinness, D. L. (2001). Ontology Development 101: A Guide to Creating Your First ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001.

Ore, C-E. and Eide, Ø. (2009). TEI and cultural heritage ontologies: Exchange of information? *Literary and Linguistic Computing*, **24**(2): 161–72.

Schneider, R. (2000). *Grundriß zur kognitiven Theorie der Figurenrezeption am Beispiel des viktorianischen Romans*. Tübingen: Stauffenburg.

Zöllner-Weber, A. (2006). Formale Repräsentation und Beschreibung von literarischen Figuren. In Braungart, G. G., Gendolla, P., and Jannidis, F. (eds), *Jahrbuch für Computerphilologie 7*. Paderborn: Mentis-Verlag, pp. 187–203.

Zöllner-Weber, A. (2007). Noctua literaria - A system for a formal description of literary characters. In Rehm, G., Witt, A., and Lemnitzer, L. (eds), *Data Structures for Linguistic Resources and Applications*. Tübingen: Narr, pp. 113–21.

Zöllner-Weber, A. (2008). *Noctua Literaria – A Computer-Aided Approach for the formals Description of Literary Characters Using an Ontology*. Ph.D thesis, Bielefeld University. http://bieson.ub.uni-bielefeld.de/volltexte/2008/1309/ (accessed 4 April 2011).

Zöllner-Weber, A. (2009). Ontologies and logic reasoning as tools in humanities? *Digital Humanities Quarterly*, **3**(4), http://www.digitalhumanities.org/dhq/vol/3/4/000068/000068.html.