Content Analysis of Museum Documentation with a Transdisciplinary Perspective

Günther Goerz, Martin Scholz
University of Erlangen-Nuremberg, Computer Science Department (8)
Erlangen, Germany
goerz@informatik.uni-erlangen.de

Abstract

In many cases, museum documentation consists of semi-structured data records with free text fields, which usually refer to contents of other fields, in the same data record, as well as in others. Most of these references comprise of person and place names, as well as time specifications. It is, therefore, important to recognize those in the first place. We report on techniques and results of partial parsing in an ongoing project, using a large database on German goldsmith art. The texts are encoded according to the TEI guidelines and expanded by structured descriptions of named entities and time specifications. These are building blocks for event descriptions, at which the next step is aiming. The identification of named entities allows the data to be linked with various resources within the domain of cultural heritage and beyond. For the latter case, we refer to a biological database and present a solution in a transdisciplinary perspective by means of the CIDOC Conceptual Reference Model (CRM).

1 Specific Goals of Content Analysis

When we speak of museum documentation, we address a wide variety of document types. First of all, there are acquisition and inventory lists or index cards, which contain more or less detailed records of museum objects. Often these are accompanied by photographs, restoration records, and further archival records. If curators prepare exhibitions, usually they provide catalogs by compiling data from sources, such as those just mentioned, and by contributing short articles on the exhibits. Last but not least there are scholarly monographs on museum objects.

With the introduction of information technology in museums and cultural heritage institutions, such records have been stored in (relational) database systems and content management systems. At the beginning — with the exception of bibliographic records — there were no metadata standards at all in the museum world. Since the 1990s, many metadata schemata have been proposed for the field of cultural heritage, some with very detailed classification features for specific object types[1]. There is still an active discussion about metadata schemata and their standardisation, as can be seen with recent proposals for CDWA Lite, museumdat and their combination (Stein and Coburn, 2008).

Today, access to museum documentation via the World Wide Web has become a matter of course, in particular, if the documentation has been the result of publicly funded research projects. Naturally, printed editions are still a very important medium of publication. However, in many cases the data are too voluminous, which means only abridged versions are published in print, while the full data are available only in digital form. Web access allows many means to retrieve and print the data, with very little cost involved. Using controlled language defined in terminologies and formal ontologies, different forms of "intelligent search" come within reach as well as interactive evaluation and visualisation methods. But it is not only access to the data alone; interactivity opens up possibilites for Wiki-style annotation and scholarly communication, as well as forums for the general public. Furthermore, the technology provides methods to link the data with other resources, e.g. authority files containing biographical or geographical data.

[1] cf. Getty Foundation's Metadata Crosswalk http://www.getty.edu/research/conducting_research/standards/intrometadata/crosswalks.html ;visited 03.12.2008.

Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education –LaTeCH – SHELT&R 2009, pages 1–9,
Athens, Greece, 30 March 2009.

A common situation in museum documentation is characterized by the fact that it is centered around museum objects, i.e. there is a database system or content management system, which contains structured descriptions of museum objects and further information about their creators, provenance, use, and so forth, according to given descriptive and administrative metadata schemata. Besides fields in such data records enforcing (more or less strictly defined) data types, e.g. for inventory numbers, there are free text fields which contain important background information about persons, objects, materials, stylistic festures, etc. without any further tagging. Basically, the free text fields are open for any kind of information which cannot be expressed in the strictly defined parts of the schema. Therefore, overall, the given data records at best provide a semi-structured representation.

The free text fields and their relations to other fields, in particular, indicate a clear need for content analysis. Firstly, named entitities must be identified, in particular person and geographic place names. For instance, there may be a data field for the creator of a work of art and another one for the place where this work was created, additionally one or more free text fields which talk about the artist's family relations, when he came to the mentioned place and how long he stayed there, etc. As this example indicates, at least a second type of linguistic expressions, time specifications in a variety of forms, ought to be recognized. In the future, we would like to identify event descriptions and how they are related among each other, for which the recognition of named entitites and time specifications is a first step.

In the following sections we describe our approach to address these problems. The next section outlines characteristic features of the data with a reflection on their typicality. Section three is the central technical part presenting the shallow text analysis techniques we use — word class tagging, recognition of temporal specifications, place and person names — and the utilization of name authorities for lexical and semantic enrichment. In the fourth section we show how the results achieved so far can be used to construct event-based shallow semantic representations related to the CIDOC CRM. Furthermore, the CRM is also the key to transdisciplinary approaches in museum documentation as outlined in the final section with an example between biology and cultural history.

2 Characteristics of the Data

We are working[2] with data which resulted from a project on goldsmith art in Nuremberg, executed at the German National Museum, providing descriptions of more than 6700 objects, 2290 artists, many locations, etc. Furthermore, with the museum's content management system we can access many more data records on sculptures and paintings — with a particular emphasis on the work of Albrecht Dürer — up to 1800. The latter corpora were accessed primarily to verify the general usefulness of the approach that will be presented in the following sections.

For many projects in the field of cultural heritage in Germany, a condition for public funding has been to use the MIDAS[3] data schema (Heusinger, 1989) in combination with a specific database implementation (HiDA). MIDAS defines a framework of record types with appropriate properties for terms (thesauri), time, place, artists, other persons and organizations, objects, content and signs, events, sources, and administrative data. The goal of MIDAS was to establish a de facto standard based on the current documentation practice in museums. Depending on what is to be documented, the appropriate record types can be selected. HiDA is a data administration system, which provides a graphical user interface for data input, editing, and search; it stores the records not in a database system, but in a system of files, one for each type, in a proprietary format. For this reason and problems in handling the user interface, many HiDA-encoded data are now being converted to an XML representation. For the free texts, we decided to follow the encoding rules of the Text Encoding Initiative (TEI) (Ide and Veronis, 1995)[4] for text bodies.

The actual encoding of the XML-transformed data sets is still very close to HiDA as far as the "classes" and properties are concerned. Currently, the data are in the process of being transformed to the emerging museumdat/CDWA Lite

[2]Research project "WissKI — Wissenschaftliche KommunikationsInfrastruktur"; funding provided by the German Research Council (DFG)

[3]Acronym for "Marburger Informations-, Dokumentations- und Administrations-System", not to be confused with the MIDAS heritage standard in the UK.

[4]Website: `http://www.tei-c.org/index.xml` ; visited 15.12.2008

standard (Stein and Coburn, 2008)[5], which in turn is compatible with CIDOC's Conceptual Reference Model (Doerr, 2003)[6]. The CRM is the formal reference ontology, which defines the conceptual background for the semantic representations resulting from content analysis. We refer to the CRM as a formally defined reference ontology because with the "Erlangen CRM"[7] we provided is a description logic version of the latest standard (ISO 21127:2009), implemented in OWL-DL (Goerz et al., 2008).

As for the content of the free text fields, the texts contain well-formed sentences in the linguistic sense, although in some cases, one can find elliptic formulations in telegraphic style. In most cases, the texts refer to defined data record fields (persons, creatorship, object properties, bibliographic data), providing additional information, for which there is no other place in the schema. A great deal of the texts talk about family and other relations between persons, about creatorship, techniques, actions of the mentioned persons other than the creation of the artwork, and the general cultural context. As in early modern German there is a great orthographic variation even in writing person and place names, many of the texts suggest disambiguations of different kinds. Nevertheless, there are still many writing variants of named entities. Furthermore, many texts contain quotations from reference works, some of which are too old to obey the actual orthographic standards.

It is important to notice that the actual data we have to deal with are nevertheless a typical example of the state of the art of documentation in many cultural heritage institutions. Hence, the techniques of content analysis and annotation presented in the following will be of a general utility in many similar projects.

3 Content Analysis: Shallow Parsing and Semantic Representation

The texts contained in the free text fields are encoded with the TEI Lite tag set, supplemented by tags of the module `namesdates` for person and place names. For processing, all texts in the free text fields of a MIDAS file — e.g., the "object" file containing all object descriptions in the "database" — are merged in a single multi-text TEI file. Each text from a data field is represented as a `<text>` element where the text proper without further annotations is contained in its subordinate `<body>` element. The association between the TEI text elements and the orginal MIDAS data fields is assured by unique XML identifiers in `xml:id` attributes. The "raw" text data are transformed automatically into the initial TEI representation in a rather straightforward way by a script. No further internal structuring is provided at this stage; annotations are added by subsequent processing steps.

Shallow parsing for place names and time specifications is based on sets of chunk rules implemented with the Definite Clause Grammar (DCG) formalism[8] which are executed by Prolog. There are grammars for person and place names and for time specifications; these sets of grammar rules define three partial "parsers". For the three parsers there is only one pass, and there is in principle no restriction on the order in which they are applied. The parsing results of each of the parsers are represented as feature structures, which are then converted to TEI tags and inserted into the file by a separate software component. At this stage, there is no recognition and resolution of anaphoric references, such as pronouns. In a second and independent pass, a lookup of person and place names in Name Authority files is executed and the results are collected in local files. There is no filtering applied to the lookup because, at this point, no special knowledge about these resources is available.

3.1 Tagging

First of all, the texts encoded conforming to the TEI guidelines are annotated with word class tags and lemmata (base forms) by a POS tagger. Lemmatisation is very useful in languages with a rich inflectional system, such as German. For POS tagging, we use the Stuttgart TreeTagger[9] with the STTS tagset which provides categories for German words and delimiters.

[5] cf. slide set by Georg Hohmann: `http://www8.informatik.uni-erlangen.de/IMMD8/Services/transdisc/cidoc2008_hohmann.pdf` ; visited 03.12.2008

[6] The actual version of the ISO standard and a lot of accompanying materials can be retrieved from `http://cidoc.ics.forth.gr/` ; visited 03.12.2008.

[7] `http://www8.informatik.uni-erlangen.de/IMMD8/Services/cidoc-crm/` ; visited 05.02.2009

[8] based on previous work by (Tantzen, 2004).

[9] Institute for Automatic Language Processing of the University of Stuttgart. The tagger is available at `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html` ; visited 03.12.2008.

The resulting TEI tags express morphosyntactic descriptions. Of particular interest are the tags `<w>` for encoding words and `<c>` for individual punctuation marks which are very well suited for encoding tokens: Both can accept an attribute `type` for the determination of the word or character class. Lemmata are encoded with the attribute `lemma`.

3.2 Time Specifications

The "temporal" grammar/parser recognizes a broad variety of temporal expressions built up from days, weeks, months, seasons of the year, years, decades, and centuries.[10] Time specifications may be given as absolute or relative.

Absolute time specifications describe unique time points or intervals on the time line, such as calendar dates (e.g. *28. August 1749*) and open or closed time spans (e.g. *bis 1832*, up to 1832). Furthermore, modifying particles are recognized, e.g. *Mitte 1749* (midyear 1749) or *Ende März 1832* (end of March 1832).

To determine the missing data in relative time specifications, such as *drei Wochen später* (three weeks later), a kind of anaphoric resolution method is applied. Therefore, we keep track of all occurences of temporal expressions. For resolution, we choose the most recently mentioned at the appropriate level (day, month, year).

3.3 Places

The normal case of place specifications in the goldsmith corpus consists of a geographic place name or a preposition followed by a place name. In some cases there are also enumerations of place names. We distinguish between a named entity and the corresponding linguistic phrase. Named entities are looked up in a local dictionary which is built from entries in Name Authorities.

Before lexical lookup, a procedure is executed which prevents the annotation of lower case words as named entities. It implements the simple heuristics that — even composite — named entities are designated by words beginning with a capital letter, but not each word beginning with a capital letter is a named entity as in English. In German, a noun must be written with its first letter in upper case.

Each named entity is associated with one out of ten geographical types to avoid aggregations of incompatible types as in *die Städte München und Berlin und Finnland* (the cities Munich, Berlin and Finland). On the other hand, certain words such as city, town, settlement, etc. are associated with such a type ("city") to be used as a constraint on subsequent proper nouns.

3.4 Persons

Parsing of person names is much more difficult because they are more complex and there is a considerably larger variation than with place names. Whereas, usually, composite place names are lexicalized, this is not a real option for person names. Every person in German speaking countries has at least one first and one surname, optionally amended by further forenames, appellations of nobility or ancestry or generation. We do not regard titles and forms of address such as *König* (king) or *Herr* (Mister) as parts of names — in spite of the fact that according to German law the title of *Doktor* (doctor) is a part of the name.

For name parsing, the constituents of names are divided into four categories: forenames, surnames, copula, and generation appellations. The class of copula subsumes many particles which serve as predicates of nobility or ancestry, e.g. *von*, *van der* or French/Spanish/Italian *de la*. The category of generation appellations contains words and numberings to distinguish persons with the same name, e.g. *Karl **V.***, *Albrecht Dürer **der Ältere***.

There are several sources of ambiguities with person names the grammar has to deal with, as well w.r.t. the correct interpretation of their parts as regarding their reference:

- Persons are often referenced not by their full name, but only by their first or surname.
- Many first names may also occur as surnames.
- Many surnames are also names of professions or places.
- There are several standards of equal range for the ordering of name parts.
- The use of a comma to separate surname and first name can be confused with an enumeration and vice versa.

Therefore we use dictionaries for the four categories of name parts. There are words, which may be members of several categories, if there are several possibilities of interpretation. The dictionaries for generation appellations and copula are

[10] The actual text corpus does not contain time of day expressions.

small and have been assembled manually. For first and surnames, several name lists were compiled into one dictionary file from lists available via Name Authorities and also from the WWW.

To recognize person names containing very rare first and surnames, as well as writing variants which are not contained in the lexicon, we use a system of syntactic and semantic cues — based on statistical analyses of the texts — which indicate the occurence of a name at a specific location (cf. table).

syntax of the trigger	**Example**
profession name	Goldschmied Samuel Klemm
appellation name	Frau Martha
preposition relation name	mit Meister Silvester
possessive pron. rel. name	Seine Tochter Katharina
relation `des/der` *name*	Tochter des Christian Mahler
relation `von` *name*	Sohn von Peter v. Quickelberg
relation `:` *name*	Lehrling: Lang, Joh. Christoph

Table 1: Rules for person name triggers. Words to be inserted for *profession*, *appellation* and *relation* are taken from hand-made lexica.

Statistical analysis of the goldsmith corpus has given clear evidence for three groups of words whose occurrence indicates an immediate following person name: Appellations of professions, appellations plus titles, and relations between persons. A relation between persons is regarded as a cue only if certain particles occur immediately before or after it. The word sequence *"Tochter von"* (daughter of) is a good example of such a cue for a subsequent person name.

In a first step, the name parts and the cues are labelled separately. In a second pass, whenever a cue or a name part is encountered, an algorithm to assemble the parts into complete person names is run. It tries to match the current word sequence with different patterns of name parts which constitute valid person names, i.e. it applies different finite state machines[11] to the word sequence. The longest sequence recognized by a finite state machine is assumed to be a name (see Table 2).

3.5 Name Authorities

To achieve a normalization of appellations, person and place names are looked up in name authorities. There are several authorities, none of which can claim completeness, and each has its strengths and weaknesses. Up to now, we have used the following interfaces — however, further interfaces are in preparation: BGN: Board on Geographic Names (German places File)[12], Diskus "Geographie-Datei" (distributed with MIDAS)[13], Orbis Latinus (Graesse)[14], Getty TGN (Thesaurus of Geographic Names)[15], PKNAD (Person Names) by prometheus e.V.[16], and Getty ULAN (United List of Artist Names)[17]

There are two modes of use for name authorities in the process of named entity recognition:

1. Decision making: The data are used as dictionaries for the person name and place name parsers.

2. Enrichment with metadata in a second phase once the named entities are identified.

As there are not yet unique formats and inter-

Pattern	**Example**
s	Jamnitzer
$s\ g$	Jamnitzer II
$f^+\ s$	Hans Jamnitzer
$f^+\ g\ c\ s$	*Hans II von Jamnitzer
$f^+\ g\ s$	Hans II Jamnitzer
$f^+\ c\ s$	*Hans von Jamnitzer
$f^+\ g$	Hans II
f^+	Hans
$s\ ,\ f^+\ g$	Jamnitzer, Hans II
$s\ ,\ f^+\ c$	*Jamnitzer, Hans von
$s\ ,\ f^+\ g\ c$	*Jamnitzer, Hans II von
$s\ ,\ f^+$	Jamnitzer, Hans

Table 2: Recognized name patterns with examples showing the name of the goldsmith "Hans II Jamnitzer". `s` stands for surname, `f` for forename, `c` for copula and `g` for generation particle. The '+' sign expresses one or more occurences; the asterisk indicates that the name has been modified to fit the pattern with "von".

[11] Finite State Machines are formal automata which recognize regular expression patterns; i.e., both notions are equivalent.

[12] `http://earth-info.nga.mil/gns/html/namefiles.htm` ; visited 17.12.2008

[13] `http://museum.zib.de/museumsvokabular/index.php?main=download` ; visited 17.12.2008

[14] `http://www.columbia.edu/acis/ets/Graesse/contents.html` ; visited 17.12.2008

[15] `http://www.getty.edu/research/conducting_research/vocabularies/tgn/` ; visited 17.12.2008

[16] `http://www.prometheus-bildarchiv.de/index.php?id=56\&L=0\&skin=0` ; visited 17.12.2008

[17] `http://www.getty.edu/research/conducting_research/vocabularies/ulan/` ; visited 17.12.2008

faces for the mentioned name authorities, we implemented a querying interface for each name authority in both modes with the exception of the Getty vocabularies. These are not used directly as dictionaries, but only for metadata enrichment, because the data must be retrieved place by place from individual web pages due to the lack of an appropriate API.

3.5.1 Name Authorities as Dictionaries

Name authorities can be directly accessed through the dictionary interfaces of the place and person name parsers. To accelerate the search for entries, the retrieved data are stored in local dictionary files, one for each name authority. The dictionary files can be generated either during the recognition process or off-line. To keep the local data up to date, the generation process should to be repeated from time to time, at least for some of the mentioned resources.

3.5.2 Name Authorities for Metadata Harvesting

Metadata harvesting has been implemented as a separate process; it consists of the search for annotations of named entities in the TEI files, of querying name authorities and collecting the metadata through special interfaces, encoding in an appropriate format and storing in local files. We do not rank name authorities and the content of the metadata; its structure and degree of detail are taken as retrieved. However, with each data set the list of IDs of the tagged findings in the TEI file is stored.

3.6 TEI-Encoding of Named Entities

Temporal expressions are encoded with the `<date>` tag. For the attributes, the distinction between time spans and time points is represented by the attributes `from` and `to`, or the attribute `when`, resp.

The tag `<placeName>` is used to annotate place expressions as a whole. To label the named entities contained within, the TEI module `namesdates` provides six tags according to its geographical type: `<district>`, `<settlement>`, `<region>`, `<country>`, `<bloc>` und `<geogName>`; for some of them there may be a refinement by means of the ten geographic types mentioned in 3.3 with the attribute `type`.

For person names, the TEI tag `<persName>` and several subtags are defined, among which `<surname>`, `<forename>`, `<nameLink>` and `<genName>` correspond exactly to the name parts presented above.

3.7 Evaluation Results

The three partial parsers are executed in sequential order. The best results were obtained in the order time – person – place:

On the goldsmith corpus with a test set of about 2000 word types, a precision of 81.8% and a recall of 72.6% was achieved with the described level of granularity, i.e., accounting for the distinction of first and last names and geographic types.

If these distinctions are dropped, as in many other systems, precision increases to 83.0% and recall to 82.6%.

A separate evaluation of the parsers (in parentheses: with distinctions) showed for

- time: precision 89.0% and recall 92.1%,
- person: precision 74.4% (71.6%) and recall 87.0% (75.5%),
- place: precision 78.9% (69.1%) and recall 76.9% (71.7%),

Depending on the choice of name authorities used for lexicon generation, and due to a high degree of ambiguity, too many words may be classified as place names. For this reason, BGN has been left out, because it led to a considerable decrease of precision and a slight increase of recall.

4 Building Blocks for Event Recognition

With parsing results for person and place names and time specifications, we have a first-level partial semantic representation of text chunks, which could be combined into larger representation structures. However, considering the characteristics of the given free texts and the state of the art in computational linguistics, it would be presumptuous to aim at a deep semantic analysis. Nevertheless, under the assumption of compositionality, i.e., the assumption that semantic representations of larger units are to be composed from those of their parts in a systematic way, it is possible to assemble partial semantic representations. In particular, we are interested in identifying events and the involved actors, objects, and instruments. Event recognition in texts has been an active research area in recent years, in particu-

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<TEI>
  <teiHeader>
    ...
  </teiHeader>
  <text>
    <group>
      <text xml:id="kue00020e00029">
        <body>
    Er ist offensichtlich identisch mit dem Ornamentstecher
<persName xml:id="persName4815108">
  <forename>Theodor</forename>
  <surname>B.</surname>
</persName>
    und stammte wie
<persName xml:id="persName6059828">
  <surname>Bang</surname>
    ,
  <forename>Hieronymus</forename>
</persName>
<placeName type="zone" xml:id="placeName12514145">
    aus
  <settlement type="stadt">Osnabr&uuml;ck</settlement>
</placeName>
    (Verwandtschaft?) Kein Eintrag in den Eheb&uuml;chern
<date from="1600-01-01" to="1699-12-31" xml:id="date33491090">
    des 17. Jhs.
</date>,
    kein Eintrag im Totenbuch St.
<placeName type="zone" xml:id="placeName3113238">
  <district type="stadtteil">Sebald</district>
</placeName>
    bzw.
<placeName type="zone" xml:id="placeName9131644">
  <district type="stadtteil">Lorenz</district></placeName>
    bis
<date from="1623-01-01" to="1630-12-31" xml:id="date24591544">
    1623/30
</date>.
<date from="1611-01-01" to="1611-12-31" xml:id="date22562823">
    Von 1611
</date>
    stammt eine von
<persName xml:id="persName5006112"><surname>Bang</surname></persName>
    gestochene Ansicht
 <placeName type="zone" xml:id="placeName4837279">
    von
 <settlement type="stadt">Bamberg</settlement></placeName>.
 <persName xml:id="persName7446303">
   <forename>Balthasar</forename> <surname>Keimox</surname>
 </persName>
    verlegte von ihm eine Folge von
    12 Stichvorlagen mit reichem Arabeskenwerk.
        </body>
      </text>
    </group>
  </text>
</TEI>
```

Figure 1: Parsing result: annotated text in TEI encoding. (Layout has been rearranged for readability.)

lar in combination with text mining.[18] In previous work (Fischer et al., 1996; Bücher et al., 2002), we augmented a chart-based chunk parser with an incremental construction procedure for (partial) Discourse Representation Structures (DRSs). DRSs are semantic representations which contain a list of discourse referents, introduced by named entities or definite noun phrases, and a body, which consists of a possibly complex logical form representing the meaning of the given part of speech[19]. For events, we use a neo-Davidsonian representation, i.e., the corresponding verb is a one-place predicate whose argument is a discourse referent representing an event, conjoined with binary relations for the thematic roles. For example, the sentence *"Albrecht Dürer painted a self-portrait in 1500 in Nuremberg"* would get a semantic representation in which — extremely simplified — e would be the discourse referent for the event, *paint(e)* the representation of the event, and *actor(e,a)*, *object(e,s)*, *time(e,1500)*, etc. constitute the body, where a and s are the discourse referents for the artist and the self-portrait, resp. DRSs are reaching beyond sentence limits and can in principle be combined into larger and larger discourse structures. Therefore, they are appropriate representations on which reference resolution mechanisms, such as those described in (Fischer et al., 1996) can be built. In our current work, a central activity is to port this method and its implementation to the museum documentation domain and enrich it by collocational analysis as in (Smith, 2002).

The representation of events is not only an extremely important key to content analysis, but also the pivot which connects various objects, persons, places, with each other, making a variety of connections explicit, which are implicitly contained in the data fields and free texts of records of different types. It, therefore, becomes an obvious goal to enrich such relational structures with further information elements from other cultural heritage resources — beyond name authorities. In our particular application, access to Getty's Art and Architecture Thesaurus (AAT), to other museum and collection databases or online auction catalogs would be obvious. Unfortunately, many of these resources use idiosyncratic data formats just as MIDAS mentioned above. At best, they refer to a formal representation of their respective domain, in terms of a so-called "formal domain ontology", a representative hierarchical structure of concepts, properties and constraints of the domain. However, to satisfy the desideratum of linking diverse data collections, an intermediate level of interoperability is required. A well proven approach for such information integration tasks is to link the different domain ontologies to a generic reference ontology, which contains just the fundamental and most general concepts and properties for a wide variety of applications. In fact, for the field of cultural heritage, CIDOC's Conceptual Reference Model (CRM) is such a reference ontology. It is worthwhile to notice that, among other things, the CRM emphasizes the event-driven perspective, in fact, events are the glue in CRM which connects all documentation elements. As a first step, we have already implemented a generator for CRM instances from TEI-conformant texts with named entity annotations.

5 Transdisciplinary Aspects

Coming back to our project on goldsmith art documentation, we recognize clues in the data, which point beyond the domain of cultural history: there are goblets and centerpieces (epergnes) showing sculptered animals, such as lizards and beetles. Two of the documented objects exhibit a beautiful stag beetle, which induced interesting questions about those insects, not only on their iconographic significance, but also on their determination and classification in biology, the distribution of species, etc. This illustrates that there is a need to connect with further knowledge sources, such as resources from biology, biodiversity research, etc. For example, we may want to consult a database such as BIODAT, maintained by the natural history museum Koenig in Bonn. Considering the completely different scientific background and the different perspectives in description, this task seems to be very ambitious, to say the least. Whereas the stag beetle in the foot of the goblet is described in terms of art history and metallurgy, we find a completely different description of a pinned stag beetle in the BIODAT data base. We may be lucky to identify it there if we know the precise species name in advance, but in many cases, there is a significant chance that the match-

[18] To quote just one prominent example, cf. the TERQAS (Time and Event Recognition for Question Answering) Symposium, 2002, `http://www.timeml.org/site/terqas/index.html` ; visited 05.02.2009

[19] cf. (Kamp and Reyle, 1993)

ing task will fail. At this point in time, we can only provide a sketch in terms of an example how we would approach this challenge. But it seems obvious if we could find a general way to connect to different description systems, we would approach the long-term goal of an "epistemic web".

Recent efforts showed that there is in fact a way to a solution, indicated by the term "transdisciplinarity"; first results have been presented at the first meeting of the CIDOC working group on "Transdisciplinary Approaches in Documentation"[20]. Originating from philosophy of science (Mittelstrass, 2002), transdisciplinarity concentrates on problems, which cannot be solved within a single disciplinary framework. It takes a new view on the unity of science, focussing on scientific rationality, not systems. Taking into account that for all sciences there are common elements in the practice of argumentation and justification, transdisciplinarity is a research principle in the first place. Its emphasis on rational language use in science offers a clue to the field of documentation; as a starting point, our methodological focus is first of all on data integration . Taking into account that transdisciplinarity addresses the practice of research, this framework should support an action and event perspective on a generic level, i.e. for the tasks of classification, representation, annotation, linking, etc.

In fact, we claim that the CIDOC CRM can play the role of such a transdisciplinary framework; at least for the stag beetle on goblets and still life paintings, some other insects and also birds on drawings and paintings, the modelling task has already been performed successfully. For the birds — hooded crows in Dutch winter scenes in Brueghel paintings — our transdisciplinary modelling effort provided a nice result for biodiversity research as a side effect: During the "little ice age" hooded crows lived in Western Europe, whereas today they can only be found east of the Elbe river.

Acknowledgments

The authors are grateful for valuable hints and discussions to Siegfried Krause, Georg Hohmann, Karl-Heinz Lampe, and Bernhard Schiemann and to the anonymous reviewers for valuable suggestions.

[20] at the CIDOC 2008 conference; online materials are available via `http://www8.informatik.uni-erlangen.de/IMMD8/Services/transdisc/` ; visited 03.12.2008.

References

Kerstin Bücher, Günther Goerz, and Bernd Ludwig. 2002. Corega Tabs: Incremental semantic composition. In Günther Goerz and et al., editors, *KI-2002 Workshop on Applications of Description Logics, Proceedings*, volume 63 of *CEUR Workshop Proceedings*, Aachen, September. Gesellschaft für Informatik e.V.

Martin Doerr. 2003. The CIDOC Conceptual Reference Model: an ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3):75–92, September.

Ingrid Fischer, Bernd Geistert, and Günther Goerz. 1996. Incremental semantics construction and anaphora resolution using Lambda-DRT. In S. Botley and J. Glass, editors, *Proceedings of DAARC-96 — Discourse Anaphora and Anaphor Resolution Colloquium*, pages 235–244, Lancaster, July.

Günther Goerz, Martin Oischinger, and Bernhard Schiemann. 2008. An Implementation of the CIDOC Conceptual Reference Model (4.2.4) in OWL-DL. In *Proceedings of the 2008 Annual Conference of CIDOC — The Digital Curation of Cultural Heritage*, pages 1–14, Athens, Benaki Museum, September 15–18.

Lutz Heusinger. 1989. *Marburger Informations-, Dokumentations- und Administrations-System (MIDAS) / [1,2]*. Saur, München.

Nancy Ide and Jean Veronis, editors. 1995. *Text Encoding Initiative. Background and Context*. Kluwer, Dordrecht. Also in: Computers and the Humanities. Vol. 29, No. 1–3 (1995).

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer, Dordrecht.

Jürgen Mittelstrass. 2002. Transdisciplinarity — new structures in science. In *Innovative Structures in Basic Research. Ringberg-Symposium, 4–7 October 2000*, number 5 in Max Planck Forum, pages 43–54, München.

David A. Smith. 2002. Detecting events with date and place information in unstructured text. In *Proceedings of the 2nd ACM+IEEE Joint Conference on Digital Libraries*, pages 191–196, Portland, OR.

Regine Stein and Erin Coburn. 2008. CDWA Lite and museumdat: New developments in metadata standards for cultural heritage information. In *Proceedings of the 2008 Annual Conference of CIDOC*, Athens, September 15–18.

Svenja Tantzen. 2004. Ein Prologparser für temporale und lokale Ausdrücke in einem 'Geosem-System' für das Deutsche. Technical report, Friedrich-Alexander-Universität Erlangen-Nürnberg, Philosophische Fakultät II, Erlangen. Master Thesis.