INTERDISCIPLINARY SCIENCE REVIEWS, Vol. 34 No. 2–3, 2009, 206–223

Weaving a New Fabric of Natural History

Antal van den Bosch, Piroska Lendvai, Marieke van Erp and Steve Hunt

Tilburg Centre for Creative Computing, Tilburg University, The Netherlands

Marian van der Meij, René Dekker

National Museum of Natural History, Naturalis, Leiden, The Netherlands

Natural history offers an interestingly rich mix of traditional and modern ways of organizing data, information, and knowledge. The Linnaean tradition still defines the basis of how taxonomic knowledge of organisms is organized, while at the same time complementary perspectives on databases and ontologies are developed and implemented, to provide enhanced access to natural history collection data to researchers in taxonomy and biodiversity. Some of this knowledge enrichment may even be automated, and bootstrapped from basic object metadata. This metadata is largely composed of natural language text, which is generally more noisy and ambiguous than numeric data. In this contribution, we present two methods for the automated discovery of metadata from textual object databases: first, the automatic detection of new metadata in existing free-text database columns, and second, the discovery of new ontological relations between metadata elements.

KEYWORDS Natural history, Taxonomy, Metadata discovery, Ontologies, Natural language processing

Introduction

One of the best known taxonomies in the world is what biologists and in fact many others simply refer to as 'the taxonomy': the system that Swedish botanist, physician and zoologist Carl Linnaeus (1707–1778) created to organize all of the planet's organisms. The Linnaean taxonomy distinguishes between seven hierarchical levels: Kingdom, Phylum or Division, Class, Order, Family, Genus, and Species (each of which may have several super- or sub-groupings). A natural history collection database will tend to list this information for each of its zoological or botanical objects, usually abbreviated

Published by Maney on behalf of the Institute

DOI 10.1179/174327909X441117

to the so-called *alpha taxonomy* that specifies for one organism only its Genus and Species. Besides this important anchor point, natural history collection databases contain much more information about the items in their collection. For instance, it is vital to know each object's scientific status; e.g., whether the specimen in question is in fact the *type* of its species, i.e. the particular physical example of an organism that is known to have been used when the species was first described. Also, the 'who, where, when' of the actual finding of the specimen is typically recorded, as well as time stamps of its arrival in the collection, its determination date, the author and date of the scientific publication associated with the object, the way in which it is stored, and under what identification number.

The Dutch National Museum of Natural History, Naturalis, of which we use the collection as the working example throughout this contribution, maintains various databases cataloguing the animal specimens in its collection, all containing roughly the same information of the aforementioned types. The main focus of our study is the collection of reptile and amphibian specimens, collected mainly in the Amazon rainforest and Indonesia, and the associated database that was created over a period of roughly two decades to capture about one-third of the museum's collection. In the database, each specimen is represented by one record. The database is a so-called flat database of size $n \times m$, where n is the number of records (the database rows), and m is the number of attributes (the database columns) by which all specimens are characterized. The database contains 16,870 records, and 39 columns. Some columns contain simple values such as a single term or a numeric value, while others contain longer stretches of free text. The information was manually entered by collection managers at the museum, and stems from various hand-written and printed sources such as field logbooks, museum registers, labels attached to objects, and publications. Table 1 shows some sample values of a record in the Naturalis database.

Even a fragment of a single record such as the one in Table 1 reveals much of how knowledge is organized in a typical natural history object database. Taken together, the attributes constitute a structured summary of the typical story of how a specimen was collected and identified. Yet, it is obviously not a fluent texual story. It leaves out many details, and stores exactly the

TABLE 1

SAMPLE COLUMNS OF ONE RECORD IN THE REPTILES AND AMPHIBIANS DATABASE

Column name	Value
REGISTRATION #	5529
ORDER	Sauria
SPECIES	pseudolemniscatus
COLLECTION DATE	06-11-1985
COLLECTOR	J. Doe
COUNTRY	Indonesia
ALTITUDE	1700 m
AUTHOR	Roux, 1911
SPECIAL REMARKS	died in captivity 23 September 1994

The collector's name is anonymized.

information that is deemed necessary for preservation and future research. Nevertheless, some leeway is offered in this strict template, and this leeway is heavily utilized. Columns such as SPECIAL REMARKS may be used (and misused) to store factoids and miniature stories that have no predetermined column of their own (such as the fact that specimen number #5299 died in captivity nearly 9 years after it was collected). The fact that there is no date of death column in the database reflects certain choices made in the past on what information types were considered relevant in the domain for preservation and research. However, as time goes on and the database is filled with records, it becomes clear from inspection that in fact particular types of data are regularly added to free-text columns, e.g. to SPECIAL REMARKS and BIOTOPE. These regular additions are usually phrased in natural-language text that is non-standardized and thus can only be interpreted by humans; that is, by speakers of the particular language used.[1] If we could identify some recurrent patterns in these texts by automatic means, we take a step towards solving a core issue of knowledge structuring for digital cultural heritage data: the generation of new metadata.

In the next section, we introduce an automatic metadata discovery method that makes use of the fact that the natural language used in free-text columns has some discoverable regularities, or grammar. Discovery is enabled by a grammar learning method that captures the fact that, for example, there are syntactic patterns from which the words BORN and DIED can be extracted and hypothesized as metadata label candidates. After validation by an expert, the metadata labels discovered in this way can in fact be incorporated in the current metadata structure (in our case: promoted to database columns), and filled automatically through the learned grammar rules.

The example in Table 1 exhibits a more or less flat perspective on the most salient properties of natural history objects, which does not fully capture that the object is related in many complex ways to people, places, and time spans, all of which in turn have other relations with other objects, such as other animals collected during the same expedition by the same collector in a certain region. In other words, a more realistic underlying conceptual representation of a natural history object database would be a connected graph of nodes representing domain entities such as animal specimens, collectors, and locations, where connections only exist between meaningfully related entities. For example, there exists a relation between a specimen and its collector, of the type 'collected by', whereas the relation between a specimen and its location of collection could be named 'collected at'.

In the subsequent section, we describe a method by which we are able to create relation labels such as 'collected by' on the basis of linguistic analysis and a large textual resource, Wikipedia. The method is straightforward: we look for the co-occurrence of entity pairs such as an instance of a genus name and a country name in sentences in Wikipedia. When a sentence is found that contains both, linguistic analysis is carried out on that sentence, extracting the verb connecting the instances in focus. This method thus constitutes a step in the automated discovery of a domain ontology.

After we have described and demonstrated our two data enrichment and structuring methods, we discuss how these automatic methods could be integrated in the actual workflow of a cultural heritage institution. One important conclusion is that, although the computer science part of our undertaking is about the automatic aspect, the expert should always be in the loop. Therefore, where we say 'automatic', we should always stress that, in fact we develop semi-automatic knowledge enrichment methods that are intended to help the human expert in his work, both in terms of quality improvement and time savings. We end in discussion with a reiteration of this important issue.

Discovering new metadata in existing object metadata

The rapid growth in the digitization of data, and the resulting increased demand on accessing this digitized data, has caused many curators, researchers, and data managers of cultural heritage institutions to turn to knowledge management systems to complement their initial database management systems. Using knowledge management systems typically causes them to think about the conceptual structure of their domain, which is usually captured in what is called an ontology. In such a structure, the key classes that define the object data and metadata, as well as their mutual relations are expressed. Traditionally, ontologies are created manually, but this is a time-consuming process that needs to be repeated for every new domain. Automatic construction of a domain ontology from texts is a heavily researched area in computational linguistics, where a number of advanced technologies have been created in the past decades, as surveyed in (Buitelaar *et al.* 2005). These approaches — depending on many factors such as the amount and quality of texts available, the desired granularity of the concepts and relations in the ontology — have been shown to yield successful practical applications in several real-world domains. Yet, most of the work is still carried out manually, as the process of specifying and organizing concepts and categories into a semantic network requires domain and world knowledge.

The ontologies underlying the structure of the typical databases in use in the natural history domain nowadays, are indeed often manually constructed. This manual design, necessarily fixated at some point to allow for data entry to start in the database that would retain its fixed structure, holds the risk of becoming out-of-date over time. When this happens, it is very hard to simply change a database to the new ontological design without a considerable amount of labour. To remedy this, a number of (ongoing) attempts have been made in the cultural heritage domain to create guidelines and standards for the preservation as well as generation of metadata; an overview of these, discussing the gap between requirements and implementation possibilities, is provided in Kanter (2008). As examples of initiatives that have found application over the past years, we mention the CIDOC-CRM standards (Crofts *et al.* 2008), which are well suited for representing field logbook

entries (our database rows); specifically in natural history, standards and protocols for sharing biodiversity data are currently under development by the Biodiversity Information Standards (TDWG) group.[2] One example standard the TDWG has published is the Access to Biological Collections Data (ABCD) Schema that 'promotes standardization of the terminology used to model biological collection information and provides a general format for data exchange and retrieval for biological collections', covering nearly 1200 domain classes. This number stands in stark contrast to the small number of columns in typical legacy databases. The development of the new standards shows that when the community invests a focused effort, the conceptual domain space turns out much richer than implied by the old database structures. When these standardization efforts are not implemented, old inferior choices are kept, and loosely defined database columns such as SPECIAL REMARKS or MISCELLANEOUS become the collection bin for a mix of attributes that do not have their own columns. Very often, these columns have relatively rich and verbose natural language content.

Since the contents of columns that contain unstructured descriptions are of lower semantic coherence than the clear-cut columns, information stored in this way is suboptimal for effective querying. The columns often contain mixes of several similar confusable data types, for example variants of numerical data types (coordinates, weights, counts), that describe different properties of an object without structured reference to the semantics of these. Free-text columns of the databases held at Naturalis are labelled SPECIAL REMARKS, BIOTOPE, LOCALITY, while columns specifying a single concept, entity, or measurement are, for example GENUS, SPECIES, COUNTRY, ALTITUDE, RECORDER. Consider the following example from the SPECIAL REMARKS column:

> Slides JD 1975-xviii-27/29, 1975-xix-20/25; tape recording 1975 II B 297–304. Acquired as a gift from the British Museum (Nat. Hist.), BMNH 1975. 1348

If one searches for tape recordings, or acquisitions from a certain year, it is inevitable to browse through several numbers, because retrieving tape recording IDs or gift acquisitions can only be performed by accessing the entire SPECIAL REMARKS column. The query would be more efficient if the various ID numbers of slides, tape recordings, and inventories could be separately searchable.

Metadata creation is arguably a knowledge-intensive procedure, because it involves both domain and organizational expertise. In the case of large and complex sets of (legacy) data, it is a tedious and often infeasible task to define and construct content metadata manually. Our method aims to alleviate this by mining the content of free-text database columns for metadata candidates in an unsupervised way. We use an algorithm that implements Alignment-Based Learning[3] (van Zaanen 2001), to induce the grammar of multi-word textual database columns. Grammar inference systems are language-independent and require no linguistic knowledge to induce structure from texts, which makes them suitable for processing specialized domain cultural heritage texts that often amalgamate several languages — not rarely in diachronically different variants — on which most natural

language processing tools trained on general corpora are likely to perform suboptimally. The procedure detects regularly occurring terms that express object or domain properties that, after validation, can be promoted to structural elements of the data model, to which values are automatically assigned. A simplified visualization of automatically generated metadata suggestions is provided in Figure 1. We can also regard the assignment of new metadata labels to content in complex columns as database column expansion.

Buitelaar *et al.* (2004) suggest indeed that extracting concepts and their attributes is a productive way to populate ontologies. However, contrary to our approach, their method requires linguistically annotated texts, which is costly, error-prone, and available in automatic mode only for a limited amount of languages. A supervised approach to content analysis, applicable to syntactically well-formed museum object descriptions is described in the recent study by Goerz and Scholz (2009). A method that is reported by Hovy *et al.* (2003) for growing a domain model automatically can be related to our work because it utilizes heuristics as well: the observation that newly constructed databases are often accompanied by online explanatory textual material. They propose supervised machine learning techniques to process glossaries, i.e., structured concept descriptions written in natural language, harvested from websites.

Material

We prepared four distinct datasets from two databases, in order to verify that our method is generally applicable. The SPECRA dataset is drawn from the

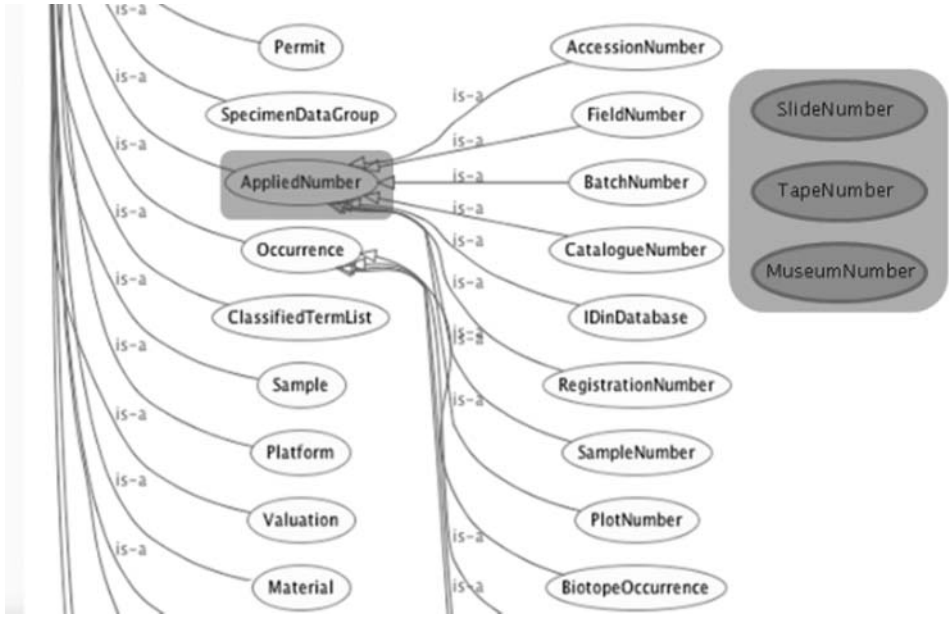

FIGURE 1 Introducing new metadata candidates to an existing data model.

SPECIAL REMARKS column of the reptiles and amphibians database, the BIORA dataset from the biotope column of the same database. The SPECCRU and BIOCRU datasets are gained from the corresponding columns of the database holding descriptions of crustaceans. The full content of each column belonging to these columns is regarded as a sentence. The words in the sentences are tokenized, and the datasets are filtered from duplicate sentences. In the SPECRA and SPECCRU sets, all occurrences of digits are collapsed to the symbol NUM. The top section of Table 2 reveals that the sets differ in size and complexity. In general, data originating from the SPECIAL REMARKS columns are more complex, shown by the larger vocabulary size and sentence length. As noted earlier, all sets contain a mix of languages.

The Alignment-Based approach

The Alignment-Based Learning algorithm (ABL) is an unsupervised, symbolic structure bootstrapping system, described in van Zaanen (2001). ABL finds the grammar that underlies a corpus of plain text sentences, without using any external sources of information. Its application is therefore attractive when no linguistic annotation and processing tools are available or applicable to a domain. Since our datasets are extracted from single database columns, they mostly contain grammatically elliptical sentences. Also, they are very domain-specific. ABL is nonetheless able to induce grammar from these elliptical, telegram-style texts, which would be hard, if not impossible, to achieve by off-the-shelf syntactic parsers trained on full sentences.

The first cycle of ABL aligns the sentences in the input corpus, creating a pool of hypotheses, where unequal parts of the sentences are hypothesized as grammatical constituents of the same type. The procedure can be likened to bracketing the words in each sentence by comparing it to every other sentence in the corpus, on the basis of various calculations of string edit distance, i.e. the number of operations required to transform one string of words into the other. For more details, ABL's manual can be consulted. The grammar can be visualized in terms of production rules from symbolic non-terminals to terminals. In order to maximize the amount of column expansion candidates, only this first cycle of alignment learning was used, with default settings, which allows for overlapping constituents in the hypothesis space.

The extraction of column expansion candidates draws on the observation that information type (henceforth INFOTYPE) metadata are typically syntactic heads modified by their actual information type value (henceforth INFOVALUE). For example, the birds database at Naturalis is a relatively newer database than the reptiles and amphibians or the crustaceans, and is structured on a finer granularity level, as collection managers decided to describe the specimens as thoroughly and as structured as possible. The birds database defines a general INFOTYPE metadata group, with categories such as SIZE, WEIGHT, ALTITUDE, EXPEDITION, PLUMAGE, containing textual and numeric values. As remarked earlier, in the reptiles and amphibians and the crustaceans databases, this type of information is usually lumped together in SPECIAL

REMARKS. It would be very helpful for Naturalis to organize this knowledge better, preferably in a semi-automatic way. We observe from the Bird database that most syntactic heads modified by a value can qualify as metadata, and use this fact in our approach.

Our procedure thus retrieves head-modifier instances from ABL's hypothesis space at the phrasal level: it spots *empty constituents*, i.e. indications of a possible but unrealised so-called *modifier* phrase (regarded as the value of the new metadata instance), and stores its context word (i.e., the modified *head*) as the metadata candidate. The value may constitute both the immediate left and right context of the candidate. Suppose the BIOTOPE column of three separate records contains the following texts:

1. *lagoon near sea, bottom with algae*
2. *creek in forest along sea, rock bottom*
3. *pitfalls in swamp forest; dead humid leaves.*

Here, the metadata candidate labels, the grammatical heads, will be 'lagoon', 'bottom', and 'forest', as these are the terms that are observed to be modified by empty constituents in the output of the ABL processing. For example, empty constituents are found by ABL for 'bottom' to its left (sentence 1) and right (sentence 2); based on this evidence, the modifying values 'rock' and 'with algae' are extracted as values of the candidate label 'bottom'. For the 'forest' label, we can identify the values 'along sea' and 'swamp'. All candidate instances are presented to a domain expert for evaluation, listed with the dependent value, and, if accepted, stored as metadata.

Quantitative evaluation

The bottom section of Table 2 describes the results of the candidate extraction procedure on the four datasets. Comparing the amount of tokens in each dataset and the number of proposed candidates, it is reasonable to assume that we have generated a potential reduction in time needed for a human expert to validate the candidates, as it is arguably more feasible to evaluate a few hundred candidate type-value patterns than having to browse thousands of columns manually, searching for regularities. The mean ratio of accepted and proposed candidates (i.e., precision) is then about 11%. The average number of accepted new labels is 25, a magnitude which is in line with

TABLE 2

TOP SECTION: STATISTICS OF THE COLLECTION DATABASE COLUMNS 'SPECIAL REMARKS' AND 'BIOTOPE' FROM THE *REPTILES AND AMPHIBIANS* (RA) DATABASE AND THE *CRUSTACEANS* (CRU) DATABASE. BOTTOM SECTION: METADATA EXTRACTION RESULTS WITH ABL AND HEURISTICS ON THE FOUR DATASETS

	SPECRA	BIORA	SPECCRU	BIOCRU
# Sentences	2641	694	665	781
# Words per sentence	11.8	6.5	8.5	4.7
# Vocabulary	2570	1047	1607	588
# Generated candidates	650	198	149	128
# Accepted candidates	37	25	15	24

manually expanded and linked database columns of the same cultural heritage institution. For example, the Birds database contains around 40 infoType labels in the additionalInfo column. Moreover, these are semantically on the same granularity level as those extracted from our experimental datasets.

When comparing frequency lists from the corresponding columns, around 65% overlap can be found between the candidate list and the 200 most frequent tokens; however, heads and modifiers could not be separately retrieved based on just frequency lists. Separating head terms from modifiers is important in establishing relations between concepts in the ontology. For languages such as English, where noun–noun modification (such as 'swamp forest') is a highly productive structure, our unsupervised procedure can serve a disambiguating role, as a word can often be both a head ('forest' in 'swamp forest') and a modifier ('forest floor') in a closed domain corpus.

Qualitative evaluation

To illustrate the semantic range of the output of our approach, Table 3 displays some of the metadata candidates generated from each dataset,[4] while Table 4 lists some of the actual values extracted per candidate. Note that the candidates cover a broad range of domain concepts, related to both natural history (e.g. types of biotope such as 'rock') and collection management (e.g. data carriers, preservation forms), as well as data types — digits used in inventories and IDs —, but also events (birth, transfer of custody).

What cannot be seen from the table is that often synonyms (e.g., 'formerly' vs. 'originally'), spelling variants, as well as semantically related word forms, are extracted, e.g. both the nominalized and the inflected verb form ('loan' and 'loaned'). We emphasize that the extraction output is seen as a list of candidate terms that may or may not be accepted by a domain expert, or may simply serve as a source of inspiration for creating metadata, based on evidence of regularity. Many of the candidates are possibly going to be collapsed into a single concept in a final manual authorization procedure that the institution chooses to implement. For the value assignment procedure, it is important that several syntactic or language variations of one and the same

TABLE 3

PARTIAL LISTS OF METADATA EXTRACTED FROM FOUR ANIMAL COLLECTION DATABASE COLUMNS

specRA	bioRA	specCRU	bioCRU
born	bush	colour	algae
died	forest	drawing	beach
formerly	ground	female	bottom
length	meter	male	cave
loan	pool	NUM	clay
museum	river	photo	coral
NUM	road	pot	creek
photo	swamp	see	forest
slide	vegetation	size	water
tank	water	tube	zone
...	...	...	...

TABLE 4

METADATA CANDIDATE ENTITIES WITH ASSIGNED VALUES (LEFT OR RIGHT MODIFICATION) FROM THE SPECIAL REMARKS AND BIOTOPE COLUMNS

Modifier	Head	Modifier
Hebrew	University	
Stanford	University	
South Australian	Museum	
British	Museum	
red	rock	
bare	rock	
	rock	surface of hill under stones
	rock	boulder near road
	rock	in savanna
scanty	vegetation	
Kalahari bushfield	vegetation	
scattered Pinus	vegetation	
	forest	of Quercus ilex
	forest	in moss cushions
	forest	on loamy soil
	road	through cultivation
	road	under fallen Cecropia leaves

term are detected, but these need to be manually linked by an expert. The reported precision score is based on a preliminarily validated and collapsed set of new metadata.

Evaluating the quality of our current results can be challenging. A satisfaction score is difficult to create from the accepted term ratio, because the acceptance of terms might be biased by individual preferences. For example, a collection manager may disagree that certain candidates are helpful when directly searching the database, or may have preconceptions about how a database column structure should look. Yet, column expansion does not necessarily mean the physical modification of a database structure by adding more columns, rather, it is a method to induce additional layers of metadata, and, for example, link the primary layer to the content of the columns through domain concepts of various granularity. The method's goal is to allow for the induction of domain concepts that enable better structuring — and thus searching and mining of the contents — of a collection database.

Discovering new ontological views on natural history

In our contribution, we frequently use the technical term 'database columns'. The term masks the fact that we interpret database columns as ontological classes. If our reptiles and amphibians database contains 39 columns, we assume that the designer of this database had 39 ontological classes in mind. In a domain ontology, classes can typically be connected through relations. These relations are typically absent in a flat database, but they define the structure of a relational database, and they connect the ontological classes to form an ontological graph. In the working example of the reptiles and amphibians database, we do not know beforehand which columns are related,

and what the label would be for that relation. Again, as we demonstrate in this subsection, this process can be semi-automatically performed, much like our approach to discovering new candidate metadata columns as described in the previous subsection.

In a clean-slate situation where the classes are known, but their potential relations (and their names) are not, it is hard to think of a way to find and name these relations without resorting to human expert knowledge. Yet, the semi-automatic method we propose does attempt to do this. If we assume that the columns in our flat database represent the classes in our ontology, then the database columns provide us with many instances of values of each class. We can then attempt to find co-occurrences of pairs of instances in a knowledge-rich textual resource, e.g. an encyclopaedic resource, that we assume contains explicit relations between these instances, and thus between the ontological classes.

Any approach that is based on text to discover relations is dependent on the quality of that text. In this study, we opt for Wikipedia as the resource from which to extract relations between terms. Although the status of Wikipedia as a dependable resource is debated, in part because of its dynamic nature, there is some evidence that Wikipedia can be as reliable a source as one that is maintained solely by experts (Giles 2005). Wikipedia is also an attractive resource due to its size (currently nearly 12 million articles in over 250 languages). Additionally, Wikipedia's strongly hyperlinked structure closely resembles a semantic net, with its untyped directed relations between the concepts represented by the article topics. We are aware that there are more complete resources for the natural history domain such as for instance the Encyclopaedia of Life,[5] Amphibiaweb,[6] and the Reptile database.[7] However, these resources mainly express relations between instances of different taxonomic classes through a hierarchical layout, whereas we are looking for a textual label for relations, also beyond the taxonomical.

Due to its structure and breadth, Wikipedia is a potentially powerful resource for information extraction. Pre-processing of Wikipedia content in order to extract non-trivial relations has been addressed in a number of studies. Syed *et al.* (2008) for instance utilize the category structure in Wikipedia as an upper ontology to predict concepts common to a set of documents. In Suchanek *et al.* (2006), an ontology is constructed by combining entities and relations between those extracted from Wikipedia's category structure and WordNet. This results in a large 'is-a' hierarchy, drawing on the basis of WordNet, while further relation enrichments come from Wikipedia's category structure. Chernov *et al.* (2006) also exploit the Wikipedia category structure to which concepts in articles are linked to extract relations.

Extracting relations from Wikipedia

For this study, we used a database snapshot of the English Wikipedia of 27 July 2008. This dump contains about 2.5 million articles, including a vast number of articles dealing with natural history topics that one would typically not find in general encyclopaedias. An index was built of a subset of

the link structure present in Wikipedia. This subset of links is constrained to those links occurring in sentences from each article in which the main topic of the Wikipedia article occurs (as taken from the title name). For example, from the Wikipedia article on *Anura*, the following sentence would be included in the experiments:[8]

> *The frog is an [[amphibian]] in the order Anura (meaning 'tail-less', from Greek an-, without + oura, tail), formerly referred to as Salientia (Latin saltare, to jump).*

This approach is based on the assumption that the strongest and most reliable lexical relations are those expressed by hyperlinks in Wikipedia pages that relate an article topic to another page (Kamps and Koolen 2008).

In order to find meaningful relations between two database columns, query pairs are generated by combining two values occurring together in a record. This approach limits the number of queries applied to Wikipedia, as no attempt is made to find relations between values that would not normally occur together. This approach yields a query pair such as *Reptilia Crocodylia* from the taxonomic class and order columns, but not *Amphibia Crocodylia*. Because not every database cell is filled, and some combinations occur more often, this procedure results in 186,141 query pairs. Each query pair containing two values from two database columns is sent to the system. The system processes each term pair in four steps. A schematic overview of the system is given in Figure 2.

Step 1 We look for the most relevant Wikipedia page for each term, by looking up the term in titles of Wikipedia articles. As Wikipedia formatting requires the article title to be an informative and concise description of the

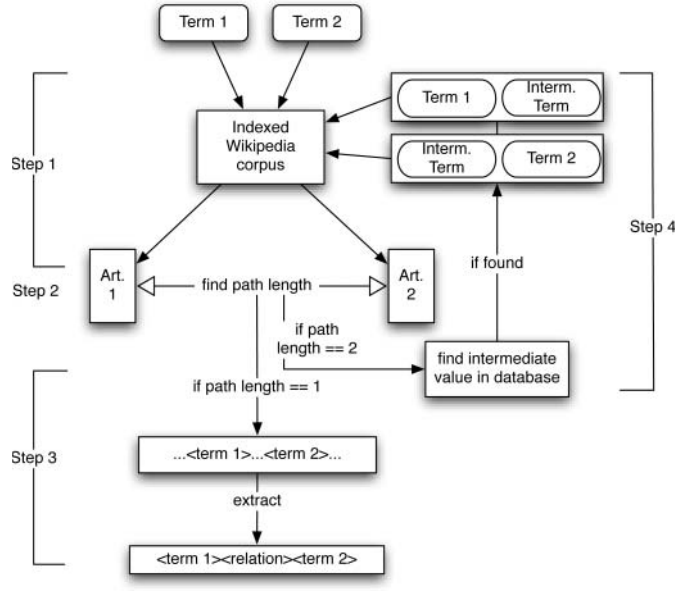

FIGURE 2 Schematic overview of the system.

article's main topic, we assume that querying only for article titles will yield reliable results.

Step 2 The system finds the shortest link path between the two selected Wikipedia articles. If the path distance is 1, this means that the two concepts are linked directly to each other via their Wikipedia articles. This is for instance the case for *Megophrys* from the genus column, and *Anura* from the order column. In the Wikipedia article on *Megophrys*, a link is found to the Wikipedia article on *Anura*. There is no reverse link from *Anura* to *Megophrys*; hierarchical relationships in the zoological taxonomy such as this one are often unidirectional in Wikipedia so as to not overcrowd the parent article with links to all its children.

Step 3 The sentence containing both target concepts as links is selected from the articles. From the *Megophrys* article this would be '*Megophrys is a genus of frogs, order [[Anura]], in the [[Megophryidae]] family*'.

Step 4 If the shortest path length between two Wikipedia articles is 2, the two concepts are linked via one intermediate article. In that case, the system checks whether the title of the intermediate article occurs as a value in a database column other than the two database columns in focus for the query. If this is indeed the case, the two additional relations between the first term and the intermediate article are also investigated, as well as the second term and that of the intermediate article. Such a bridging relation pair is found for instance for the query pair *Hylidae* from the taxonomic order column, and *Brazil* from the country column. Here, the initial path we find is *Hylidae* $\leftrightarrow$ *Sphaenorhynchys* $\rightarrow$ *Brazil*. We find that the article-in-the-middle value (*Sphaenorhynchys*) indeed occurs in our database, in the taxonomic genus column. We assume this link is evidence for co-occurrence. Thus, the relevant sentences from the Wikipedia articles on *Hylidae* and *Sphaenorhynchys*, and between articles on *Sphaenorhynchys* and *Brazil* are added to the possible relations between ORDER — GENUS and GENUS — COUNTRY.

Although no regulations exist for the names of ontological relations, often a verb or verb phrase head is taken, optionally combined with a prepositional head of the subsequent verb-attached phrase (e.g., 'occurs in', or 'donated by'). In this study, we assume that good candidate labels are frequent verbs or verb phrases found between instances from a particular pair of classes, and that this may sometimes involve a verb-attached prepositional phrase containing one of the two terms.

In order to extract these candidate labels, the selected sentences are POS-tagged and parsed using the Memory Based Shallow Parser (Daelemans *et al.* 1999). Evaluating semantic relations automatically is hard, if not impossible, since the same relation can be expressed in many ways, and would require a gold standard of some sort, which for this domain (as well as for many cultural heritage domains) is not available. Therefore the five most frequently recurring phrases that occur between the column pairs, where the subject of the sentence is a value from one of the two columns, are presented to the human annotators. The cut-off of five was chosen to prevent the annotators from having to evaluate too many relations and to present

only those that occur more often, and are hence less likely to be misses. Misses can for instance be induced by ambiguous person names that also accidentally match location names (e.g., *Dakota*).

Evaluating relations from Wikipedia

The four human judges were presented with the five highest-ranked candidate labels per column pair, as well a longer snippet of text containing the candidate label, to resolve possible ambiguity. The items in each list were scored according to the total reciprocal rank (TRR) (Radev *et al.* 2002). For every correct answer, $1/n$ points are given, where n denotes the position of the answer in the ranked list. If there is more than one correct answer, the points will be added up. For example, if in a list of five, two correct answers occur in positions 2 and 4, the TRR would be calculated as $(1/2+1/4)=0.75$. The TRR scores were normalized for the number of relation candidates that were retrieved because, for some column pairs, less than five relation candidates were retrieved.

As an example, for the column pair PROVINCE and GENUS, the judges were presented with the relations shown in Table 5. The direction arrow in the first column denotes that the GENUS value occurred before the PROVINCE value.

The human judges were sufficiently familiar with the domain to evaluate the relations, and had the possibility to gain extra knowledge about the class pairs through access to the full Wikipedia articles from which the relations were extracted. Inter-annotator agreement was measured using Fleiss's Kappa coefficient (Fleiss 1971).

Results and evaluation

As expected, more relations are discovered between certain columns than others. We presume that these columns have a stronger ontological relation than others. For some database columns, such as the COLLECTION DATE column, we did not retrieve any relations. This is not surprising, as even though Wikipedia contains pages about dates ('what happened on this day'), it is unlikely that it would link to such a domain-specific event such as an animal specimen collection. Relations between instances denoting persons and other concepts in our domain are also not discovered through this approach. This is because many of the biologists named in the database do not have a

TABLE 5

RELATION LABEL CANDIDATES FOR THE PROVINCE AND GENUS COLUMN PAIR, AND A SENTENCE FRAGMENT EXEMPLIFYING THEIR OCCURRENCE

Direction	Label	Snippet
→	is found in	…is a genus of venomous pitvipers found in Asia from Pakistan, through India,…
→	is endemic to	…Cross Frogs) is a genus of microhylid frogs endemic to Southern Philippine…
→	are native to	…are native to only two countries: the United States and…
→	is known as	…is a genus of pond turtles also known as Cooter Turtles, especially in the state of…

Wikipedia page dedicated to them, indicating the boundaries of Wikipedia's domain specific content.

From each column pair, the highest rated relation was selected with which we constructed the ontology displayed in Figure 3. As the figure shows, the relations that are discovered are not only 'is a'-relations as one would find in strictly hierarchical resources such as a zoological taxonomy or geographical resources. The numbers in the relation labels in Figure 3 denote the average TRR scores given by the four judges on all relation label candidates that the judges were presented with for that column pair. The scores for the relations between the taxonomic classes in our domain were particularly high, meaning that in many cases, all relation candidates presented to the judges were assessed as correct. The inter-annotator agreement was $\kappa = 0.63$, which can be taken to indicate 'good agreement'. If a relation that occurred fewer than five times was judged incorrect by the majority of the judges, the relation was not included in Figure 3.

The present study clearly leaves some room for improvement, for instance in the coverage of more general types of information such as dates and person names. For this, we intend to incorporate more domain specific resources, such as research papers from the domain that may mention persons from our database. We are also looking into sending queries to the web, while keeping the constraint of hyperlink presence. More details of this study are published in van Erp *et al.* (2009).

Discussion

In our first experimental study, we proposed a semi-automatic method to generate metadata candidates for structuring free text database columns that contain unstructured information; yielding, for example, corresponding labels added to the relevant spans of words, or the extension of the underlying

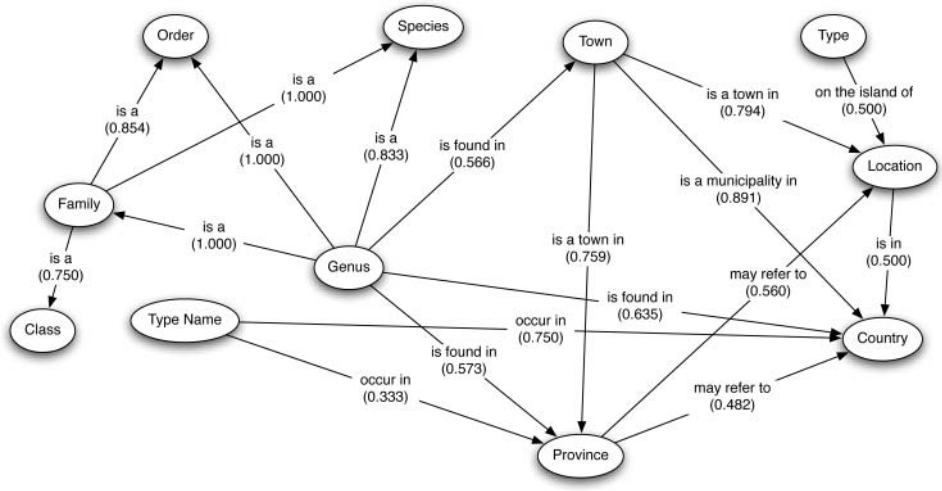

FIGURE 3 Graph of relations between columns, with TRR scores in parentheses.

metadata model. The goal of this study was to devise automatic means to mine metadata label candidates, and present the results to a domain expert for validation. We extracted metadata candidates with a heuristic method that draws on the output of alignment-based learning. No linguistic analysis is required for this process, making it attractive for cultural heritage materials that often feature several languages and distinct jargon or domain lexicons. Moreover, free-text database columns typically contain grammatically elliptical text, which is difficult to process with state-of-the-art (but generic) natural language processing tools.

We show that the method effectively finds head-modifier dependencies in a corpus of texts from a database column, and provides tangible output that qualify as elements of metadata. The extracted candidate labels are to be evaluated by a domain expert. The acquired set of terms can be directly used as a seed list in bootstrapped extraction of more candidates. Since the contents of a database column are semantically restricted, we conjecture that validated terms can be integrated in a domain ontology or data model.

In the second study, we have shown that it is possible to extract ontological relation labels to characterize the relation between a pair of ontological classes from Wikipedia. The main contribution that makes this work different from other work on relation extraction from Wikipedia is that the link structure is used as a strong indication of the presence of a meaningful relation. The presence of a link is incorporated in our system by only using sentences from Wikipedia articles that contain links to other Wikipedia articles. Only those sentences are parsed that contain the two terms we aim to find a relation between, after which the verb phrase and possibly the article or preposition following it are selected for evaluation by four human judges.

Our evaluations were partly performed by experts from Naturalis who provided working examples for the experiments in this contribution. To Naturalis, our work holds the promise that the data models that Naturalis now use will evolve and improve due to our enrichment procedures. Our methods should be seen as components in a semi-automatic process; our algorithms generate ranked suggestions of new metadata, which in the follow-up step, the human experts select and validate. The intended effect is two-fold: first, as the data model improves, access to the data and information retrieval from the data should improve as well. Second, in generating and prioritizing suggestions, we have potentially saved the expert from a tedious process with a high cognitive load, namely finding new metadata by inspecting the data and through considering, by introspection, candidate improvements. However, we are not there yet. First, what is needed to fully conclude our study is to collect evidence that information retrieval has in fact improved after our proposed metadata enrichment, and second, we need to qualitatively evaluate the complete process that the human expert goes through when selecting, validating, and implementing the changes our methods propose. Does it in fact save time, objectively as well as in the perception of the expert? This is a crucial question to be addressed in future work.

Another important, yet orthogonal issue that needs to be addressed in these types of studies, is the sociological issue of the acceptance of new technologies in a heritage environment. Almost by definition, heritage keepers are wary of incorporating technologies that threaten the sustainability of their data, be it primary or metadata. When introducing knowledge enrichment techniques, it is vital to stress the non-destructive and additive nature of the technology. Nothing is changed or destroyed; access is improved in a non-intrusive way. Given this reassurance, and given that access remains the undisputed reason for maintaining a high-quality data and metadata model, the hearts and minds of collection managers can be won in due course.

Acknowledgements

The authors would like to thank former team members Tijn Porcelijn and Caroline Sporleder for their contributions to the project, and to Sander Wubben for his collaboration on the Wikipedia study. We are grateful to Pim Arntzen, Jacob Leloux, and Ronald de Ruiter for providing us with datasets and valuable expert feedback. This research was funded by NWO, the Netherlands Organisation for Scientific Research, under the Continuous Access to Cultural Heritage (CATCH) programme.

Notes

[1] In our example database, Dutch and English are the prime languages used, but Portuguese, German and Latin can also be found.

[2] http://www.tdwg.org — last visited in March 2009.

[3] Available from http://ilk.uvt.nl/$\sim$menno/research/software/abl — last visited in March 2009.

[4] Labels are ordered alphabetically; some are translated from Dutch, some are the English original, such as 'loan' or 'formerly' in SPECRA; all labels in BIOCRU are the English original.

[5] http://ilk.uvt.nl/~menno/research/software/abl — last visited in March 2009.

[6] http://www.amphibiaweb.org — last visited in March 2009.

[7] http://www.reptile-database.org — last visited in March 2009.

[8] The double brackets indicate Wikilinks.

Bibliography

Buitelaar, Paul, Philipp Cimiano, and Bernardo Magnini, eds. 2005. *Ontology learning from text: Methods, evaluation and applications*. IOS Press.

Buitelaar, Paul, Daniel Olejnik, and Michael Sintek. 2004. A protégé plug-in for ontology extraction from text based on linguistic analysis. *Proceedings of the European Semantic Web Symposium (ESWS)*.

Chernov, Sergey, Tereza Iofciu, Wolfgang Nejdl, and Xuan Zhou. 2006. Extracting semantic relationships between Wikipedia categories. *Proceedings of the First Workshop on Semantic Wikis — From Wiki to Semantics [SemWiki2006] — at ESWC 2006*, 153–163. Karlsruhe, Germany.

Crofts, Nick, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff. 2008. Definition of the CIDOC conceptual reference model. Technical Report, ICOM/CIDOC CRM Special Interest Group.

Daelemans, Walter, Sabine Buchholz, and Jorn Veenstra. 1999. Memory-based shallow parsing. *Proceedings of CoNLL'99*, 53–60. Bergen, Norway.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76 (5): 378–382.

Giles, Jim. 2005. Internet encyclopaedias go head to head. *Nature* 438: 900–1.

Goerz, Guenther, and Martin Scholz. 2009, March. Content analysis of museum documentation in a transdisciplinary perspective. *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH — SHELT&R 2009)*, 1–9. Athens, Greece: Association for Computational Linguistics.

Hovy, Eduard, Andrew Philpot, Judith Klavans, Ulrich Germann, Peter Davis, and Samuel Popper. 2003. Extending metadata definitions by automatically extracting and organizing glossary definitions. *Proceedings of the 2003 Annual National Conference on Digital Government Research*, 1–6. Digital Government Society of North America.

Kamps, Jaap, and Marijn Koolen. 2008. The importance of link evidence in Wikipedia. *Advances in Information Retrieval: 30th European Conference on IR Research (ECIR 2008)*, Volume 4956 of *Lecture Notes in Computer Science*, ed. Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Rutven, and Ryen W. White, 270–82. Glasgow, Scotland: Springer Verlag.

Kanter, Norbert. 2008. From collection to museum management systems. A critical review of demands and features. *Proceedings of the Annual Conference of CIDOC.* Athens.

Radev, Dragomir, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating web-based question answering systems. *Demo section, LREC 2002*. Las Palmas, Spain.

Suchanek, Fabian M., Georgiana Ifrim, and Gerhard Weikum. 2006. LEILA: Learning to extract information by linguistic analysis. *Proceedings of the ACL-06 Workshop on Ontology Learning and Population*, 18–25. Sydney, Australia.

Syed, Zareen Saba, Tim Finin, and Anupam Joshi. 2008. Wikitology: Using Wikipedia as an ontology. Technical Report, University of Maryland, Baltimore County.

van Erp, Marieke, Antal van den Bosch, Sander Wubben, and Steve Hunt. 2009, March. Instance-driven discovery of ontological relation labels. *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH — SHELT&R 2009)*, 59–67. Athens, Greece: Association for Computational Linguistics.

van Zaanen, Menno. 2001. Bootstrapping structure into language: Alignment-based learning. Ph.D. diss., School of Computing, University of Leeds, UK.

Notes on Contributors

Correspondence to: Antal van den Bosch, Tilburg centre for Creative Computing, Faculty of Arts, room D343, Tilburg University, P.O. Box 90153 NL-5000 LE Tilburg, The Netherlands.

Email: Antal.vdnBosch@uvt.nl