Accessing natural history

van Erp, M.G.J.

Publication date:
2010

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
van Erp, M. G. J. (2010). *Accessing natural history: Discoveries in data cleaning, structuring, and retrieval*. TICC Dissertation Series 13.

Accessing Natural History
Discoveries in Data Cleaning, Structuring, and Retrieval

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Tilburg,
op gezag van de rector magnificus,
prof. dr. Ph. Eijlander,
in het openbaar te verdedigen ten overstaan van een
door het college voor promoties aangewezen commissie
in de aula van de Universiteit
op woensdag 30 juni 2010 om 14:15 uur

door

Maria Godefrida Jacoba van Erp

geboren op 18 november 1982 te Breda

Promotor:	Prof. dr. A. P. J. van den Bosch
Copromotor:	Dr. P. K. Lendvai
Promotiecommissie:	Prof. dr. H. J. van den Herik
	Prof. dr. W. M. P. Daelemans
	Dr. R. W. R. J. Dekker
	Dr. E. Hovy

Netherlands Organisation for Scientific Research

The research reported in this thesis was funded by the Netherlands Organization for Scientific Research (NWO) in the project Mining for Information in Texts from the Cultural Heritage (MITCH), grant number 640.002.403. The MITCH project is part of the Continuous Access to Cultural Heritage Research (CATCH) Programme.

SIKS Dissertation Series No. 2010-30

The research reported in this thesis was carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

TiCC Dissertation Series no. 13.

ISBN 978-90-8559-027-9

Voor opa

Preface

In the spring of 2005, while I was finishing up my Master's degree at Tilburg University, I knew I wasn't quite done with academia and wanted to pursue a Ph.D. Unexpectedly (to me at least) I landed a Ph.D. position at Tilburg University in collaboration with the National Museum of Natural History Naturalis in Leiden. So, in the autumn of 2005, I moved to Leiden: to new adventures, and a fair amount of time on the train to Tilburg for the weekly meetings with my supervisor and promotor Antal van den Bosch and to hang out and learn from the members of the ILK group. I am very grateful to Antal for giving me the chance to do this project and start exploring the wondrous world of interdisciplinary research in such a 'gezellige' environment and to the ILK members for making it such a fabulous place.

At my other office, Naturalis, I want to thank the researchers and database managers for educating me and my fellow MITCH colleagues about the basics of natural history research and showing us around in their databases, in particular Pim Arntzen, Eric van Nieukerken and Dirk Houtgraaf during the early days of MITCH, and later on Marian van der Meij and René Dekker.

Somewhere halfway during my project, Antal and I realised that my work was moving away from core ILK business towards knowledge representation and ontologies. Antal thought this would be an excellent opportunity to gain knowledge elsewhere and see how another research group runs from the inside. He also knew a perfect mentor to help me explore these topics: dr. Eduard Hovy at the

Information Sciences Institute in Marina del Rey, California, USA. Again I was given an amazing opportunity, as Ed was willing to have me at ISI for a couple of months. So in September 2007, I moved to sunny California. This (ultimately) 7-month stint meant a huge turnaround for my research (and my private life, but more on that later) as ISI's bustling community provided me with new inspiration and research directions, as well as giving me an incredibly good time exploring Los Angeles bars and clubs, yoga, the beach, and American traditions. Thank you Ed, for giving me this opportunity, guidance, and for finding the time to be on my thesis committee.

I also wish to thank the other members of my thesis committee for their advice and comments on my text: prof. dr. Jaap van den Herik, prof. dr. Walter Daelemans, and dr. René Dekker.

Work hard, play hard. Luckily, my friends and family provided enough distractions from deadlines and the frustrations of working with computers in the form of concerts, surf trips, dance lessons, guitar sessions, dinner parties, trips to North End, and British comedy. Special thanks to Judith and Arthur for keeping me awake far past my bedtime in Leiden and to Steve for being a great programmer and friend. Without you this thesis would have looked very differently and my guitar skills would have been even worse than they are now.

I would never have gotten anywhere near a Ph.D. if it weren't for my family. I have amazing parents who raised me to be curious and open to the world, and who have supported me with every step. It was also a true joy to grow up with my brothers Frank and Hans, and my sister Sara. I hope that you all forgive me for moving away from Etten-Leur, but there's always space for you to crash in Amsterdam.

Fresh inspiration for my research wasn't the only thing I found in LA. I also found Paul, my soon-to-be husband. Thank you for your advice, distraction, bread-crusted Ahi tuna, coffee breaks, trips to exciting places far and near, music, smiles, and love.

On to new adventures!

Marieke van Erp
Amsterdam, 11 May 2010

Contents

1

Introduction

Data is a precious thing and will last longer than the systems themselves.

Tim Berners-Lee, Interview with the British Computer Society, 2006

An immense treasure of historical information is preserved in cultural institutions such as archives, museums, and libraries. Yet, it is often difficult to access. This thesis is about improving information access to the data in cultural heritage collections. Some cultural heritage institutions have been collecting historically relevant artefacts for several centuries and, as a result of the long existence of such collections, methods for storage and access of collection information have not always been kept up to date. This impairs information access. When collection information is difficult to access, users of the collection must spend a relatively large amount of time tracking down information before they can begin to use it. Storing information in a digital resource is a first step to improving information access. When information storage is digitised, it can be enriched: all information about one object can be linked so that in a single request all information on the object is retrieved, saving users time gathering information.

This thesis is built around three main themes, that each lead to more structure for and knowledge in the digital information collections at the Dutch National Museum of Natural History Naturalis[1] which provided the case study for the research described. The first theme is data cleaning. The second theme is data structuring. The third theme is retrieving the information contained in the data.

[1] http://www.naturalis.nl/

The three themes follow up on each other: it would not make sense to start using dirty data (i.e., data that contains errors) before one has attempted to correct it and enriched it with structure that aids its retrieval.

In the remainder of this chapter, the motivation for this research is given in Section 1.1. The main issues concerning information access in the cultural heritage domain together with the problem statement and research questions follow in Section 1.2. In Section 1.3, the research methodology used in this work is explained. This chapter is concluded by an outline of the thesis in Section 1.4.

1.1 Research Motivation

The work carried out for this thesis is motivated by the need from museums, archives, and libraries to have continuous access to their collections. Continuous access to an information collection means that, regardless of physical boundaries (such as storage conditions, closing hours, etc.), users now and in the future can find what they are looking for in the collection, either within a few mouse clicks or by entering a simple query in the user interface created to access the collection.

The above motivation is the driving force behind the work that is presented in the thesis. In the remainder of this section, a brief introduction of the collaborating institution is given, as well as the starting point of the research that has been carried out to facilitate the transition from an analogue collection of information resources to a cleaned and enriched digitised collection information resource.

The data used to test the approaches presented in this thesis is provided by the Dutch National Museum for Natural History Naturalis. Natural history is not always considered part of the definition of cultural heritage, as the goal of research on natural history collections is primarily to further biological research. However, natural history collections belong to the domain of scientific heritage, a subdomain of cultural heritage. These collections are important from a heritage perspective as they provide insights into the evolution of biology research, as well as insights into general history because many specimens within natural history collections were gathered in former colonies or in other culturally and historically relevant areas and circumstances. The data provided by Naturalis is available in the form of structured textual databases and semi-structured text. The details of the data sets are given in Chapter 2.

For the work carried out here, researchers at Naturalis provided input on

desired improvements in information access. The process of transforming an analogue resource to an enriched digital resource for advanced access consists of three steps, each providing a layer of greater utility of the data. The museum harbours more than twelve million objects consisting of animal, fossil and rock specimens, as well as samples of specimens, each with their own provenance. As the collection dates back to the 18th century, the majority of the information on the specimens is still in its original paper form. The first step in the transformation from an analogue resource to an enriched digital resource is the conversion from an analogue to a digital representation. The second step is directing the information in the digital representation into a format that is more suited for digital access, such as a database. Researchers at Naturalis have begun to enter data into databases straight from their analogue sources, effectively carrying out the conversion and reformatting task in a single action. However, manual data entry is a time consuming and monotonous task. Machines are particularly suited to carry out monotonous tasks, therefore this is typically a task in which machines could support experts to enable swift and correct data entry. In order to do so, the data entry process is split up again into two steps; (1) the conversion of an analogue resource to a digital representation and (2) the organisation of the information in the digital representation into a database. Current state-of-the-art technology does not permit automatic conversion of handwritten text by multiple writers to a digital text format, so this step was carried out manually for this project[2]. However, as the human transcribers only needed to type in the text and not interpret it, the conversion step is much faster than interpreting data and formatting it for the right database field[3]. A machine may then be employed to carry out the second step of the resource transformation task, namely automatically populating a database from the digital text. For this project, the transformation task was carried out by employing a machine learning algorithm using a small number of manually annotated examples. The experiments and results of this task are described in Chapter 3. This way, the human expert can focus on interpretation of the data and correcting the machine where it fails.

Two subsequent steps provide layers that offer flexibility in representing the information and integration with other resources. (1) Information in an ana-

[2]Advances in the field of Optical Character Recognition may in the future enable machines to carry out this conversion task.

[3]As an added bonus, it can be done by laypersons, lifting an immense strain off of the domain experts' time.

logue resource such as a book is tied to the format of the book (i.e., ordering of the chapters), whereas information in a digital resource can be presented differently to different users (e.g., a more restricted view of the data for some users compared to others). (2) Digital resources can also be linked or integrated with other resources. For example, when a database containing geographic information is linked to a dedicated geographic resource containing synonymous geographic terms or terms in different language, the database is transformed into a richer and better accessible resource.

1.2 Problem Statement and Research Questions

For a successful transformation from an analogue resource into an enriched digital resource at least the following three issues need to be resolved.

1. **Data quality** The presence of errors in data hampers data retrieval. From erroneously retrieved results an incorrect inference or conclusion may follow. For example, data in natural history institutions are used to assess whether a species is threatened. If this assessment is based on incorrect or incomplete data, the conclusions inferred from the data may be incorrect and thus a species could falsely be assumed as not threatened. This conclusion could mean that no action is undertaken to protect the species, in the worst case, causing it to go extinct.

2. **Integration** When data resources are not structured to enable integration, it is difficult for institutions to share their information with each other, limiting possibilities for more thorough and more complete information access. For example, when data on a particular species from all natural history institutions is used in an analysis, it is likely that the analysis provides more complete and possibly more conclusive results than when only data from one institution is included. It is also possible to link databases to taxonomic resources, in order to update the database automatically when the taxonomy changes. For this it is essential that data resources are integrated.

3. **Access** When data resources are not easily accessible, time is lost in tracking down vital pieces of information. For example, researchers at Naturalis have stated that precious time is lost by locating books from which

information for their research is needed. Also, for users from outside the physical building, such as researchers on expedition or from other institutions, it would be beneficial to be able to consult the institute's resources without having to go there or burden a person who is present with looking up the desired information. Digital resources can easily be copied and updated, providing the possibility of access to the information for many users at the same time, contrary to their analogue counterparts. Naturally, access should be controlled. Some data is sensitive and should thus not be shared carelessly. An example of this is geographic information on occurrences of rare species of ferns. Making such information public may cause poachers to visit that location and take the remaining specimens.

Resolving these three issues ensures that information will not be easily lost, as it becomes less fragile. It will also help users do their work as less time is spent on searching for information and more time can be spent on interpretation.

To help resolve the issues, this thesis uses techniques from the Natural Language Processing (NLP) field. Most NLP techniques are developed for unstructured, or free text, such as found in newspaper articles and journals. However, textual information comes in different forms. The data sets used for this work are made up of two other types of text: structured and semi-structured. The first, structured textual data, is found in databases, where the type of information is known by the database column it occurs in. The second, semi-structured text, is found in field logbooks and registers, and here it is known which types of information it can contain but not which part of the text contains which type of information. The types of information contained in these resources are for example Latin names (indicating to what species a specimen belongs), geographic names (indicating a location that is connected with the specimen), dates (positioning the specimen in time), and descriptions of circumstances of events linked to the specimen. Analysing and mining free text, such as found in journal articles and of which it is not known in advance exactly what types of information it contains is out of the scope of this thesis.

The approaches that contribute to improving information accessibility presented in this thesis can be classified as either soft-reasoning or hard-reasoning approaches. Soft-reasoning (or data-driven) approaches rely on implicit information that is contained in the data, such as conditional dependencies. Hard-reasoning (or knowledge-driven) approaches rely on explicit information about the domain,

such as manually constructed rules.

The three issues presented above are progressively addressed in Chapters 3 to 6. Resolving these issues is key to answering the problem statement of this thesis, which is as follows.

> **Problem Statement:** To what extent can manual and automatic soft- and hard-reasoning approaches improve the data quality, structure, and access to information in an analogue cultural heritage collection of natural history?

In order to answer the problem statement, three research questions (**RQs**) are investigated in this thesis. **RQ1** consists of two parts.

RQ1 a: Can data-driven and knowledge-driven approaches provide improvements to the data quality of structured textual resources describing collection objects?
b: To what extent are the data-driven and knowledge-driven approaches complementary?

RQ2 Do automatic methods for building ontologies provide different structure for data that is not achieved by manual ontology building?

RQ3: Can access to information be aided through a retrieval system enriched with domain knowledge?

To answer the research questions, methods have been developed that resulted in the three thesis contributions (**TCs**) that are listed below, along with the chapter they are presented in.

TC1: an automatic ontology-driven error detection and correction method for structured data (Chapter 4).

TC2: an instance-driven ontology construction method (Chapter 5).

TC3: an ontology-driven information retrieval system that automatically expands queries and ranks the results to improve access (Chapter 6).

Each of the thesis contributions addresses another aspect of the process of data accessibility improvement from a raw text resource to a digital resource for continuous access. In the following section, the general research methodology is outlined.

1.3 Research Methodology

The research methodology used in this thesis is empirical. For each research question, the methodology consists of (1) collecting the relevant literature about the task at hand, (2) analysing the findings, (3) reviewing the most promising techniques mentioned in the literature for their suitability on the Naturalis data. This is followed by (4) testing their application, (5) analysing and evaluating the results of the experiments according to standard practice for each of the different tasks. In every chapter, the evaluation metrics used are detailed. Where possible, a comparison between different techniques is made.

1.4 Thesis Outline

The remainder of this thesis is structured as follows. In the first part of this thesis (Chapters 2 and 3), background about the domain and data and the groundwork done in order to facilitate the experiments in the rest of the thesis are described. The second part of the thesis (Chapters 4, 5, and 6) present the main thesis contributions and the answers to the research questions. Chapter 7 concludes the thesis. In the remainder of this section, the contents of each chapter are detailed.

In Chapter 2, the field of natural history is introduced and some necessary background is given. Moreover, the resources involved in this work are described. The chapter includes data that are used in more than one chapter of this thesis; additional chapter-specific data are discussed in the chapters they are used in.

In Chapter 3, a manually constructed ontology for the natural history domain is presented, as well as experiments and results on the automatic population of a database from semi-structured text. The work done in Chapter 3 is not part of the major thesis contributions but describes tasks necessary to carry out the experiments for the main thesis contributions.

In Chapter 4, general issues regarding data that contains errors are discussed, followed by a discussion of data quality issues particular to the field of natural history. Two methods for the cleanup of databases are presented. The first method, named Timpute, is a statistical data-driven method. The second method, named Validato, is a hard-reasoning method that involves the construction of rules as well as domain knowledge from an ontology and external resources. Analyses of the impact on data quality of Timpute and Validato provide the answer to

RQ1.

In Chapter 5, an automatic ontology construction method is presented. Ontologies may enhance a data resource by providing a meta structure consisting of domain knowledge. However, manual construction of an ontology requires significant time and attention from domain experts. Automatic ontology construction methods can remedy this. The approach presented in Chapter 5, called TWIBIO, utilises implicit domain information present in the various data resources that describe objects in the domain as well as explicit domain information described in the online encyclopaedia Wikipedia. In the ontology construction process, this implicit information is made explicit by linking the data resources to Wikipedia and extracting relations between different data points from the resources. Comparison to the ontology constructed in Chapter 3 shows that the automatically constructed ontology provides a different view on the domain, and presents an argument for the co-existence of different ontological representations of one domain. The work presented in Chapter 5 provides the answer to **RQ2**.

In Chapter 6, improvements for data access are presented. Here, a system, named MIRA, is described in which the performance of retrieval of database records is improved by applying query interpretation, query expansion, and result ranking. Analyses of the results of the MIRA experiments yield the answer to **RQ3**.

Chapter 7 summarises to what extent the problem statement and each of the research questions are answered and provides conclusions and pointers for future work.

2

Background and Resources

In all things of nature there is something of the marvellous.

Aristotle, On the parts of animals, ca. 350 B.C.

In this chapter, the natural history domain and the data sets that are used throughout the thesis are described. Section 2.1 provides a brief overview of the natural history domain. In Section 2.2, the reptiles and amphibians data sets provided by Naturalis are detailed. External resources used for this thesis are described in Section 2.3. The chapter is concluded by a discussion of some limitations and future work on the Naturalis collections in Section 2.4.

2.1 Natural History

Natural history is concerned with the study of the earth and living things on it. In the narrowest sense it contains the subfields botany and zoology, in the broader sense, it also includes geology, palaeontology, climatology, ecology, and biochemistry. The main goal of natural history is to gain a deeper understanding of the natural world by describing its structure (i.e., how the natural world is organised into different species and how these evolved over time) and characteristics of its various species such as diet, reproduction, and social behaviour.

At the Naturalis research department, over twelve million animal and geological specimens are kept, that have been gathered during more than two centuries

of research expeditions. The zoologists, entomologists, palaeontologists, and geologists at Naturalis have been focussing mainly on collection-based research, describing the morphology (physical features), anatomy (functioning of the organs), systematics (investigating evolutionary history), and biogeography (distribution of biodiversity over space and time) of the species present in the collection.

To facilitate collection-based research, it is important that the provenance of each specimen is known and accessible. For biogeographical research, it is important to know where a specimen was collected and when. Therefore, to this research field within natural history, the written and digital resources that store such facts are as valuable as the object itself. These resources describe the history of the object in three stages. The first stage describes under what circumstances the object was collected (by whom? where? when? what were the climatological circumstances? in case of an animal specimen was it still alive or already dead?). The second stage is concerned with the object's introduction into the museum collection (when did it arrive? how is it preserved? what is its position in the taxonomy? who determined this? on what shelf is it stored?). In the third stage, everything that happened to the specimen after it came into the museum collection is recorded (was it lent to another institution? is there photographic material? are there publications written about the specimen? who entered data about the specimen into a database?).

The collection information was hitherto generally only available on paper, which has the limitations of not being easily reproducible, being only accessible to one person at a time and being fragile. Digitisation will ensure that the information contained in the paper resources can be more easily reproduced and thus preserved. Through more sophisticated information management, accessibility of the collection information can be increased. Improved access will help researchers find information about objects more quickly and possibly even help them discover new information in the data that would otherwise remain hidden. Some examples of biodiversity research will be given that have been made possible through the increased digital access to natural history collections.

The digitisation efforts carried out in connection with this thesis will primarily aid researchers in the fields of systematics and biogeography as these chiefly require access to the information about the specimen and its collection rather than to the specimen itself. However, also morphological and anatomical research may be aided by the outcomes of the work done for this thesis.

In the following subsections an introduction to biogeography (Subsection 2.1.1), systematics (Subsection 2.1.2), and biodiversity informatics (Subsection 2.1.3) are given. These subsections serve as background information to the domain and its data.

2.1.1 Biogeography

Biogeography is concerned with the study of the distribution of biodiversity over space and time. Traditionally, the field studied the geographic distribution of species over time. Since the late 1980s, biogeographers also try to answer questions regarding physiological, morphological, and genetic variation among individual specimens and populations, as well as differences in the diversity and composition of plant and animal life within localities.

A species' or family's distribution pattern is outlined through information on collection of specimens by biologists, historical facts (e.g., contintental drift and glaciation), and geographic constraints (e.g., rivers and mountains which can prevent a species from dispersing). Genetic and statistical analysis methods aid researchers in answering questions such as *Why are there more species of butterfly in Austria than in Norway?* and to discover the underlying principles that account for such observations. These principles can, for instance, be put to use in predicting the influence of human interventions on animal and plant life in a certain location.

Biogeography research often focusses on one species or family of animals or plants. The classification of organisms is studied in the field of systematics, which is introduced next.

2.1.2 Systematics

Systematics is defined by Michener *et al.* (1970) as follows:

> *"Systematic biology (hereafter called simply systematics) is the field that (a) provides scientific names for organisms, (b) describes them, (c) preserves collections of them, (d) provides classifications for the organisms, keys for their identification, and data on their distributions, (e) investigates their evolutionary histories, and (f) considers their environmental adaptations. This is a field with a long history that in recent years has experienced a notable renaissance, principally*

with respect to theoretical content. Part of the theoretical material has to do with evolutionary areas (topics e and f above), the rest relates especially to the problem of classification. Taxonomy is that part of systematics concerned with topics (a) to (d) above."

The data used in this thesis is mainly of taxonomic nature, a subfield of systematics. Taxonomy is the practice and science of classification. Originally, taxonomy only concerned the biological domain, but currently taxonomy can also refer to any type of categorisation and thus the term 'biological taxonomy' is used to refer to a taxonomy in the biology domain.

Biological taxonomy is based on the classification system that the Swedish botanist, physician and zoologist Carl Linnaeus (1707 - 1778) developed and described in *Systema Naturae* in 1735 and in subsequent material. For the sake of brevity throughout this thesis the term taxonomy will be used to denote the biological taxonomy.

In the traditional Linnaean taxonomy, seven major levels are discerned that are each more specific than the previous one[1].

1. Kingdom (e.g., *Animalia*, meaning organisms with eukaryotic cells having a cell membrane but lacking cell walls)
2. Phylum (for animals), Division (for plants and fungi) (e.g., *Chordata*, meaning animals with a notochord, a dorsal nerve chord, and pharyngeal gill slits)
3. Class (e.g., Aves, meaning birds)
4. Order (e.g., Passeriformes, the order of perching birds or songbirds)
5. Family (e.g., Passeridae, small passerine birds)
6. Genus (e.g., *Passer*, old world sparrows)
7. Species (e.g., *domesticus*, in combination with *Passer*, *domesticus* denotes the house sparrow)

In addition, each level may have several super- or sub-groupings such as Superorder for the level Order. Often the super- and sub-groupings will not be mentioned as the Linnaean seven-level taxonomy suffices to distinguish a specimen. Often when the subdomain in which the specimen falls is known, for example

[1] In zoological taxonomy, only the genus and species names are italicised.

in an article on only birds, the specification will be even further abbreviated to only its two-part, or binomial, name consisting of only the genus and species.

Within a species there can be variation (e.g., some specimens may be larger than others, or have a slightly different colour). Therefore, one or more type specimens are defined as the reference specimens from which a species is described. The description of one or more type specimens is published and referred to whenever another specimen of the same species is described in the literature or in a collection. When there is more than one type specimen, the set of type specimens is called a type series. The most important specimen from the type series is the species' holotype (or its name-bearing specimen). Additional specimens that were described along with the holotype are called paratypes. If there is no holotype specified for a species, then two or more specimens are defined as syntypes that make up the species defining specimen. Later one of the syntypes can be selected to serve as the name-bearing specimen, but as the species was not originally described by this specimen, it is not called a holotype but a lectotype. If the holotype is lost or destroyed, a new specimen can be selected as the name-bearing specimen, which is then called a neotype.

Although the foundations of taxonomy were laid in the 18th century, there is still a continuous debate over species' positions in the zoological taxonomy (Schuh, 2000). This provides interesting challenges for digitisation in this field. For instance, due to new methods such as DNA research, biologists find that certain classifications need to be revised (Stoeckle, 2003). Such was the case with toads in the genus *Bombina*. Prior to 1985 these were classified under the family Discoglossidae together with other disc-shaped tongue frogs. In 1985, biologists found that the species in genus *Bombina* were so different from the other species in the Discoglossidae family that they were classified in their own family called Bombinatoridae (Cannatella, 1985). An example of a specimen from the Discoglossidae family is shown in Figure 2.1(a) and an example of a specimen from the Bombinatoridae family is shown in Figure 2.1(b). A researcher who wants to know about all specimens of family Discoglossidae in a Natural History collection might not want to retrieve specimens of genus *Bombina*, but cannot be sure that she will not if the collection contains items from before 1985. In a properly enriched digital resource, it is possible to provide a link between specimens of family Bombinatoridae and specimens of genus *Bombina* from before 1985 so that they can both be easily retrieved in a search for Discoglossidae, and in reverse, that they can be

hidden when only specimens of true Bombinatoridae should be retrieved.

(a) Mallorcan Midwife Toad (*Alytes muletensis*). Image by tuurio and wallie from http://en.wikipedia.org/wiki/Majorcan_midwife_toad. Published under the Creative Commons Attribution ShareAlike 3.0 License

(b) European Fire-bellied Toad (*Bombina bombina*). Picture by Marek Szczepanek. Image from http://en.wikipedia.org/wiki/European_Fire-bellied_Toad. Published under the GNU Free Documentation License, Version 1.2

Figure 2.1: Example of a specimen belonging to the Discoglossidae family (*Alytes muletensis*) and of a specimen belonging to the Bombinatoridae family (*Bombina bombina*).

Next to the traditional disciplines within natural history research such as biogeography and systematics, new disciplines are emerging by the virtue of new possibilities brought on by digitisation. In the next subsection, an introduction to biodiversity informatics is given, the discipline to which this thesis aims to contribute.

2.1.3 Biodiversity Informatics

With the rising of awareness in the biodiversity research community of the benefits of information technology and information sharing a new field has emerged that was coined biodiversity informatics (Soberón and Peterson, 2004). Biodiversity Informatics "includes the application of information technologies to the management, algorithmic exploration, analysis and interpretation of primary data regarding life, particularly at the species level of organization." (Soberón and Peterson, 2004).

The aim of biodiversity informatics is to create a global biodiversity map, including the approximately 1.8 million species currently known and the information about their genes, proteins, behaviour, and morphology (Wilson, 2000).

Biodiversity informatics is necessary to be able to analyse a species distribution in full. Such an analysis can lead to understanding why a species is, for instance, in decline, what factors play a role in this decline, and most importantly, how to reverse it in order to prevent extinction. Guralnick and Hill (2009) present a framework and prototype of a unified biodiversity resource but also state that there is much work to be done as, for instance, obtaining data is a large bottleneck. This is partly due to the negative attitude towards data sharing that is still prevalent in some parts of the field (Krishtalka and Humphrey, 2000). Krishtalka and Humphrey's report on policies within some institutions that prevent data sharing is alarming, as data sharing and information integration are imperative to further biodiversity research. Krishtalka and Humphrey recommend that institutions become actively involved in constructing a "biodiversity informatics resource in an open, collaborative and community-based manner." Between the publication of Krishtalka and Humphrey (2000) and the time of writing this thesis much work has been done. For example, the Biodiversity Information Standards group has been created[2]. This group was previously known as the Taxonomic Database Working Group (TDWG), and was founded in 1985. It was primarily concerned with data standards for plant taxonomic databases, but this has gradually grown to encompass taxonomic databases in general and now has members in geology, zoology, and microbiology. As the focus of the work of the TDWG shifted towards developing standards for sharing biodiversity data, the name was changed to Biodiversity Information Standards Group in 2006 to cover the focus of the group better. The work of the Biodiversity Information Standards Group has, for example, resulted in the Access to Biological Collections Data (ABCD) Schema[3], a standard for the access to and exchange of primary biodiversity data. Currently, Naturalis is working on making their database systems compliant with this standard.

The integration of species presence data and geographical resources has made it possible, for example, to predict the geographical distribution of related species in similar climatological environments as research by Peterson and Vieglais (2001) shows. In this work, geographical regions were identified that shared similarities in precipitation, temperature, elevation, and vegetation. On the basis of these regions it was possible to predict the distribution of a species by knowing the

[2] http://www.tdwg.org/ Last visited 6 June 2009
[3] http://www.tdwg.org/standards/115/ Last visited: 4 September, 2009

distribution patterns of a related species.

Guralnick and Hill (2009) present a case study in which the Barcode of Life Data System (a database containing DNA sequences that serve as identification for species) was combined with the International Union for Conservation of Nature[4] red list ratings (indicating at what risk a species is of extinction). By doing so, it was possible to automatically prioritise the conservation need of different species. This would not have been possible without the availability of digitised resources containing information on species' DNA and extinction risks.

It must be noted that the field of biodiversity informatics is quite young and therefore much of the work is still in the early stages. Collaboration with the field of informatics could help speed up developments in biodiversity informatics. Biodiversity informatics could, for example, benefit from the use of machine learning algorithms to process and analyse large quantities of data.

In the next section, the main data sets on which the majority of the experiments for this thesis were done are introduced.

2.2 The Naturalis Reptiles and Amphibians Data Sets

Over one hundred manually created collection databases exist in the research laboratories at Naturalis. The choice was made to focus on the reptiles and amphibians (herpetological) data because of its accompanying collection of semi-structured text that partly overlaps with an existing database.

2.2.1 The R&A Database

The Reptiles and Amphibians (R&A) database is a flat database with manually assembled contents. It is pieced together by manual editing by researchers at Naturalis' research laboratories from field logbooks, registers, taxonomies, and their own knowledge about the animal specimens in the collection. As of February 2006, it contains 16,870 records with 47 columns. Of these 47 columns, 5 columns always contain a value, 10 columns are unused, and an additional 12 columns are sparsely used (less than 30% filled). The most important information is found in the remaining twenty columns, namely in the taxonomic columns (12 columns)

[4] http://www.iucn.org Last visited: 4 September 2009

and geographical columns (8 columns). The database contains information in a variety of languages among which Dutch and English are the most common, but German and Portuguese are also encountered. In Table 2.1, general statistics about the data in the Reptiles and Amphibians database are summarised. A data sample is presented in Table 2.2. See Appendix A for a more detailed account of the database.

# Columns	47
# Records	16,870
% Filled Cells	57.5
Collection Dates	1880 - 1998
Collection Coverage	Approximately 1/3 of specimens in collection
Geographical Coverage	Asia, Oceania, Amazonia, Southern Europe, Netherlands

Table 2.1: Statistics on reptiles and amphibians database

Column name	Example
Class	Reptilia
Genus	Gymnophthalmus
Author	Hoogmoed, Cole & Ayarzaguena, 1992
Town	Canaripo
Location	Riverine forest
Number of specimens	1
Country	Venezuela
Biotope	between dry leaves on edge granite plate
Special remarks	Slides 1978-10-13/15

Table 2.2: Example cells from a typical record in the reptiles and amphibians database

2.2.2 Field Logbooks and Registers

The field logbooks and registers used in this study are a set of 80 handwritten logs describing the acquisition of a reptile or amphibian specimen and its registration in the museum collection, respectively.

There are 47 field logbooks in which biologists have recorded information about the collection of specimens whilst in the field. Most entries contain the following

information: (1) identification of the specimen by the taxonomy (genus + species name)[5], (2) whether the specimen is feminine or masculine, (3) a finding date and time, (4) a finding location, and sometimes (5) additional remarks on, for instance, the climatological circumstances of the find.

There are 3,859 pages in total in the field logbooks, in which 17,818 different specimens found between 1881 and 1998 are described. An example photograph of two adjacent pages of a field logbook is shown in Figure 2.2.

What happened to a specimen after it entered the museum collection is described in the registers. The 33 herpetological registers contain information on when a specimen came into the collection, what its registration number is, whether it came into the museum through a purchase, an exchange with another institution, or whether it was collected by one of the staff members. Entries can be accompanied by an orange sticker indicating whether the entry concerns a holotype. Together the registers contain 3,813 pages that cover 21,870 specimens that were acquired for the collection between 1885 and 1992. As can be seen in Figure 2.3, the registers contain much shorter entries than the field logbooks.

Although the majority of the handwriting in the registers and field logbooks is clear and easily readable, optical character recognition (electronic translation of images of handwritten, typewritten or printed text into machine-editable text) was not an option (Schomaker, 1998). Hence, the books were transcribed by a team of people trained in handwriting recognition, this was particularly useful for some of the older logs that contained curly handwriting styles. The registers and field logbooks were written by at least twelve different persons, and the entries were not very standardised, sometimes not even within a book. Also, some characters were used (such as ♂and ♀) that could not be typed in a flat text editor. In order to make sure the transcribed texts were as standardised as could be the typists were provided with guidelines, such as[6]:

- write ♂ as [male], write ♀ as [female];
- insert two empty lines between consecutive entries;
- mark unreadable text as [unreadable] to avoid spending too much time on trying to decipher an illegible entry.

[5]Only if the animal species is known

[6]The complete instruction for the typists can be found on: http://ilk.uvt.nl/mitch/typistguidelines

To test the performance of the typist guidelines several test pages were transcribed by more than one typist. The differences between the resulting texts were small and few enough to judge the guidelines clear and sufficient.

Furthermore, the typists were encouraged not to correct any spelling errors they might spot as to obtain a parallel corpus that could serve as training data for handwriting recognition in the future, as well as to gain insight into the quality of the texts (i.e., whether it contains many errors).

2.3 Additional Resources

For some experiments described in this thesis, in addition to the core data from Naturalis, other resources are used. These additional resources are described in Subsections 2.3.1 and 2.3.2.

2.3.1 GeoNames

GeoNames[7] is an aggregated geographical database that is available and accessible through various Web services. It is released under a Creative Commons attribution license[8] which makes it free to use as long as it is cited.

The GeoNames database is an integrated collection of smaller databases such as the National Geospatial-Intelligence Agency's and the U.S. Board on Geographic Names, the U.S. Geological Survey Geographic Names Information System for geographic entities. On top of that it has information about elevation through resources such as srtm3[9] and gtopo30[10], population data through worldgazetteer[11], alternative names for locations through GNS[12] etc. In June 2009, these resources contained over eight million geographical names, of which 6.5 million are unique entities, making it the most complete free resource available at the time of writing this thesis and it is still growing. It is highly structured and contains alternative names in different languages. Especially the latter is a major advantage over, for instance, Google Maps[13] for usage in conjunction with

[7] http://www.geonames.org, Last queried: 15 July 2009
[8] http://creativecommons.org/licenses/by/3.0/, Last visited 4 June 2009
[9] http://www2.jpl.nasa.gov/srtm/ Last visited: 15 November 2009
[10] http://eros.usgs.gov/#/Find_Data/Products_and_Data_Available/gtopo30_info Last visited: 15 November 2009
[11] http://world-gazetteer.com/ Last visited: 15 November 2009
[12] http://earth-info.nga.mil/gns/html/index.html Last visited: 15 November 2009
[13] http://maps.google.com/, Last visited: 4 June 2009

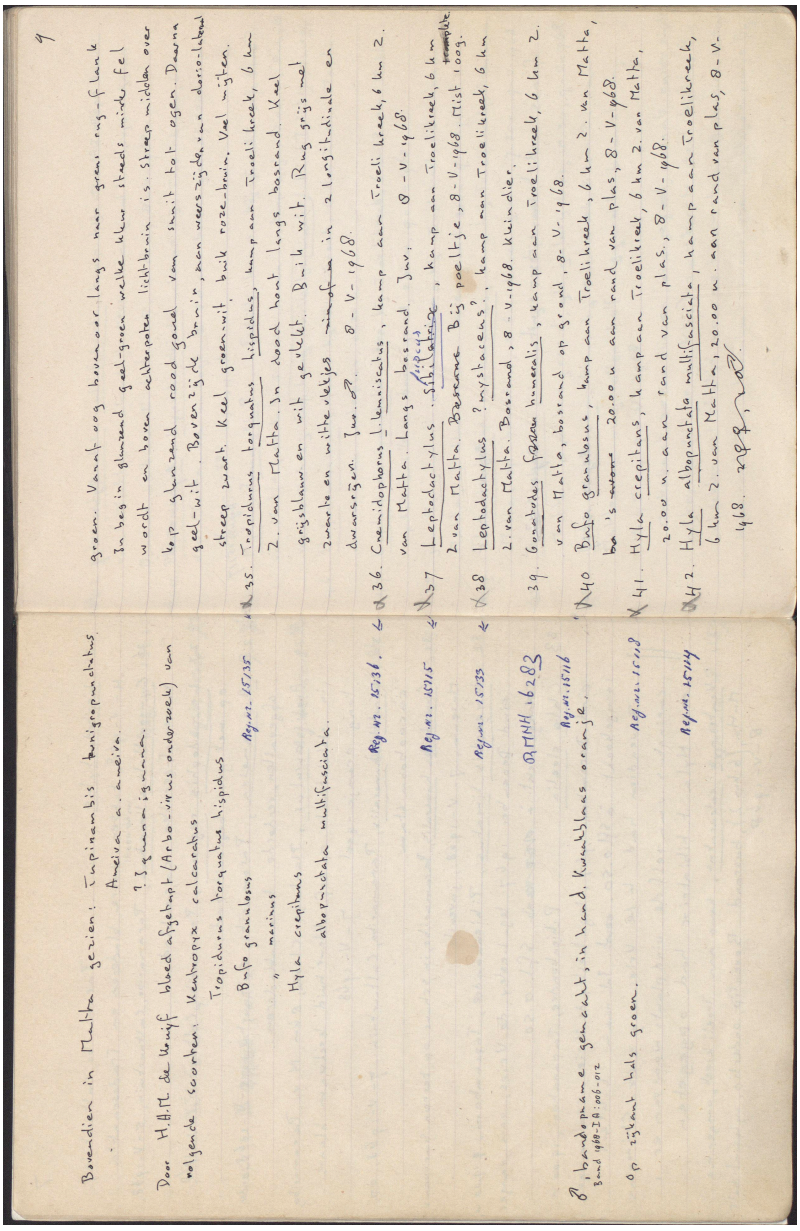

Figure 2.2: Photographic image of pages 8 and 9 of the field logbook describing collections of animal specimens in Suriname between 24 April and 5 August 1968. The right-hand page shows the initial comments recorded during the find, the left hand page shows additional comments added later.

Figure 2.3: Sample of page 12 and 13 of the second register herpetological register, describing museum acquisitions between 29 January 1932 and 5 July 1935 (register numbers 5442-6406). The orange stickers indicate that the described specimen is a type specimen.

the Naturalis data that is made up of several languages.

As the core of the GeoNames database comes from official public sources, quality may vary. Users can edit and improve the database through a Wiki, utilising user-contributed knowledge to ensure data quality and integrity. However, no extensive quality assessment has been done on GeoNames, as was, for instance, carried out on Wikipedia (Giles, 2005). Even without such an evaluation several research projects have been using GeoNames and have not reported any data quality problems (Auer and Lehmann, 2007; Jaffri, 2007; Halb *et al.*, 2008). One drawback of GeoNames is that it does not contain many archaic names for locations that do sometimes occur in the data from the Naturalis research laboratories. However, no other resource was found that covered this problem. Therefore, and because of the fact that GeoNames is easily accessible through various APIs have led to the choice of this resource to check the geographic information in the databases from Naturalis against.

2.3.2 Taxonomic Resources

To enrich and validate the taxonomic information in the reptiles and amphibians database, a reptile and an amphibian taxonomy were chosen as additional resources. Although the foundation for the zoological taxonomy that was laid by Linnaeus has not changed, there is a lively debate on the classification in the lower levels of the taxonomy. So, for each of the domains (amphibians and reptiles) there are several taxonomies in use, and some are used by one institution and some by another. Since the database at Naturalis was created following two particular taxonomic resources (one for reptiles and one for amphibians), the taxonomic resources used by Naturalis were chosen here to minimise inconsistencies arising from incompatibility with another resource.

Amphibians

For the amphibians the Frost taxonomy is used, as published online (Frost, 2009). For every species it contains:

- current scientific name and author and year of publication;
- original name, authorship, location of the type specimens and where they were collected;

- synonymous terms and location of type specimens known under these synonymous terms;
- published English vernacular names;
- known or inferred geographic distributions;
- comments on controversies or relevant taxonomic literature.

The version used in this work (version 5.3) contains descriptions of 6,433 amphibian specimens with references to the literature and synonyms.

Although the Frost taxonomy is an accepted resource there is much ongoing debate on choices made by its author, in particular on the latest version of the taxonomy that is also used for this thesis (personal correspondence with a researcher at Naturalis). However, this is inherent to the nature or systematics and the Frost taxonomy is currently the best to work with.

The resource is not as freely accessible as GeoNames, therefore it was only possible to use a hierarchical list based on the amphibian taxonomy and their accompanying synonyms and not the information on type specimen locality and geographic distribution.

Reptiles

The TIGR Reptile Database (Uetz *et al.*, 2008) is compiled from books, checklists, monographs, journals, and other peer-reviewed publications from the domain of reptile taxonomy. It is currently maintained by the Systematics working group of the German Herpetological Society (DGHT). It lists all species and their position in the taxonomy. For each of the 8,600 species the following information is stored:

- current scientific name and author and year of publication;
- synonymous terms and if relevant location of type specimens;
- known or inferred geographic distributions;
- additional comments or links to related information.

For about 3,300 species pictures are available, for 4,000 species type information is available and for 6,300 species their vernacular names are available too.

As with the Frost database, access to the online database is limited and thus only the taxonomic hierarchy including synonyms and vernacular names is used in the experiments.

2.4 Discussion

The natural history domain entails far more than only the herpetological collection. This collection was chosen because in Naturalis the most resources were available for it in digital form. Currently, the research laboratories at Naturalis maintain about one hundred databases containing information about their vertebrates, invertebrates, fossils, and insects collections. The database systems used for the registration of these collections do not comply with current demands on aspects of user friendliness, scalability, and data models. Within Naturalis, a project is underway to standardise these database systems and transport them to a single collection registration system (CRS). The CRS will help users enter data correctly by limiting what type of data can be entered (as Chapter 4 shows, the lack of this limitation is a common cause for errors in the data) and it will enable the analysis of data from all collections at the same time or quickly select subsets of collections. However, until this conversion is completed in 2010, the data selected for this work illustrates what problems and possibilities legacy specimen databases present to information access.

3

Preparatory Work

*In omnibus autem negotiis priusquam adgrediare, adhibenda est praeparatio diligens**

Marcus Tullius Cicero, De Officiis, 44 B.C.

In this chapter, work is described that enables the experiments necessary to answer the research questions. The studies presented in this chapter are not part of the main thesis contributions but do provide novel insights into the R&A data. There are two parts to the chapter; (1) automatic segmentation and labelling of segments of the R&A field logbooks and registers to resolve the data digitisation gap, and (2) work done to provide more structure to the flat R&A database by developing an ontology for it.

This chapter is structured as follows. In Section 3.1, work is described that was carried out for the automatic segmentation and labelling of the field logbooks and registers to automatically populate the R&A database with. In Section 3.2, the design process of an ontology for the natural history domain is described, after which the ontology is presented. Section 3.3 provides a chapter summary.

*Before undertaking any enterprise, careful preparation must be made

3.1 Automatic R&A Database Population

Structured data, as is found in databases, is a far more effective resource to access information via than a collection of unstructured data which is typically found on the Web or in a collection of books (Fayyad *et al.*, 1996). To transform the field logbooks and registers into database entries to automatically populate a database with, they need to be segmented and labelled. The groundwork for the results presented in this chapter was laid by Lendvai and Hunt (2008). The experiments presented in this chapter include a more thorough analysis and some adjustments to the original work.

Text segmenting and labelling is a procedure that covers various levels of granularity in Natural Language Processing (NLP) tasks. At the crudest level, it denotes the identification of different topics within a text, for example, topic boundary identification (Hearst, 1997; Beeferman *et al.*, 1999). At a finer level, there is a large body of work that is concerned with segmenting texts into smaller parts, to identify, for instance, different parts of bibliographical references (McCallum *et al.*, 2000; Han *et al.*, 2003) or postal addresses (Borkar *et al.*, 2001). The task in this work is closer to the splitting of bibliographical references than to topic boundary identification as the segments to be identified are short and it is known in advance what types of information may occur in an entry.

Prior to Lendvai and Hunt (2008), another approach had been investigated in order to segment and label the R&A field logbooks and registers automatically and populate the database. Canisius and Sporleder (2007) utilised the fact that the database is based on the field logbooks and registers, and tried to train a machine learning algorithm on data from the field logbooks and registers to further populate the database with. Canisius and Sporleder presented two approaches, namely an approach in which field log and register entries are restructured from the database values as training data for TiMBL (Daelemans *et al.*, 2004), and an approach that uses a Hidden Markov Model (HMM) (Rabiner and Juang, 1989). TiMBL is an implementation of the k-Nearest Neighbour (k-NN) algorithm that was introduced by Cover and Hart in 1967. The highest results they achieve are $F = 17.6$ for the TiMBL approach, and $F = 60.3$ for the HHM approach.

The approach that Lendvai and Hunt (2008) took relies on annotated training data alone. Two different machine learning algorithms were investigated to segment the R&A field logbooks and registers, namely Conditional Random Fields

(CRFs) (Lafferty *et al.*, 2001) and the Memory Based Tagger (MBT) (Daelemans *et al.*, 2007).

The classifiers were trained on 300 manually annotated training examples and tested on 200 held-out entries. Lendvai and Hunt (2008) achieve an F-score of 69 for the CRF approach and an F-score of 84 for the MBT approach. Although the F-score for most fields is quite high, there is still room for improvement. In the next subsection, the overlap between the field logbooks and the registers and the R&A database is analysed in order to investigate whether it is possible to automatically generate training data for a classifier. In Subsection 3.1.2, preprocessing of the data to train the classifier on is described, followed by the experiments that were carried out in Subsection 3.1.3.

3.1.1 Overlap between Field Logbooks and Registers and R&A Database

As the database is derived from information in the field logbooks and registers an attempt was made to create more training data for MBT, by utilising the overlap between the registers and field logbooks and the database. In the field logbooks, 1,842 entries could be identified from which parts match the 'genus', 'species', and 'registration number' fields of an existing R&A database record. This results in very few matches but such exact matching is necessary, as registration numbers are not unique. Researchers at the Naturalis research laboratories estimate that at least 2/3 of the database records are directly derived from the field logbooks or registers and should thus be linkable. Other methods of matching, such as by computing the cosine similarity (Salton and McGill, 1983) between pairs out of the database and log entries did not yield satisfying results. This is possibly due to the fact that a database record often contains more information coming from publications or other additional resources, making it a more complete account of a specimen's history, but also removing it too far from the field logbooks and registers to use as training data.

Upon manual inspection of the entries obtained through the strict matching method, it was found that these were indeed all field log and register entries from which the database record was derived. However, they did not prove to be useful for automatic labelling from the database entries as in the transfer of the information from the field logbooks and registers parts were often rephrased or

even translated. An example database entry and its corresponding field log entry is given below[1].

Reg #	13879
Number	1
Author	Ruibal , 1952
Determinator	Arthur Dent
Town	Airstrip Sipaliwini , 4 km E
Genus	Leposoma
Species	Guianense
Sex	m
Collector	Dent, A.
Coll #	1968-MSH8787
Coll. Date	05-09-1968

Accompanying fieldbook entry:

Leposoma Guianense, Sipaliwini, 4 km o. van vliegveld, omgeving basiskamp, bosgrond tussen bladeren, 28-VIII-1968, 12.45 u. reg. nr. 13879

In the given example, the registration number cannot be completely labelled through the database because the indication 'reg. nr.' was deleted in the database cell. This could be solved by normalisation, for instance, by adding it to the database cell. More problematic, however, is the fact that not all parts of the entry have been typed into the database and some parts were even translated (*"Sipaliwini, 4km o van vliegveld"* from the field logbook is changed into *"Airstrip Sipaliwini, 4 km E."* in the database). This problem is known as the record linkage problem and it greatly affects the usability of databases as secondary resources in tagging of more material of primary resources (McCallum and Wellner, 2003; Bellare and McCallum, 2007; Snyder and Barzilay, 2007). The likely cause for such changes is that the field logbooks were first (partly) transcribed into the registers, which were then entered into the database, where missing information was looked up again in the field logbooks. These two translation steps caused the database to lack sufficient overlap with the field logbooks and registers to utilise as training data. Experiments with adding the parts of the field logbooks and registers that could be labelled via the database to the training set did not

[1]To preserve the anonymity of the people in the data, the names 'Arthur Dent' and 'Ford Prefect' are used. The data reflects different naming conventions used in the original name, for example Firstname Lastname and Lastname, Initial

yield any significant improvements in the F-score of 84 as reported in Lendvai and Hunt (2008).

The observation that did improve the results of the segmenting and labelling was the acknowledgement of the major differences between the field logbooks and registers. Therefore, in the remainder of this section, segmenting and labelling experiments following Lendvai and Hunt (2008) are presented in which separate experiments are carried out for the registers.

3.1.2 Annotated Training and Test Data

Both Canisius and Sporleder (2007) and Lendvai and Hunt (2008) use the same annotated training and test sets. The training set consists of 300 entries and the test set consists of 200 entries. These annotated sets contain only entries from the field logbooks. The registers are sufficiently different to warrant the creation of a parallel data set. Therefore, 500 entries from the registers were selected and annotated in similar fashion to the data sets from the field logbooks. This set consisting of 500 annotated register entries was split in a set of 300 entries for training and 200 entries for testing.

The main differences between the entries from the registers and the field logbooks are the length and number of segments. On average, the register entries contain 12 segments with a standard deviation of 6.05 and the average number of tokens is 34 with a standard deviation of 22.94. The field logbook entries contain 61 tokens (σ = 35.58) on average divided over 16 segments (σ =5.12). The distribution of tags in both resources differs as well; the registers, for instance, will often contain a segment indicating a donator as they more often deal with donated and purchased specimens, while the field logbooks exclusively describe finds and thus do not contain information about donations. A collection of statistics on the different data sets is given in Table 3.1.

3.1.3 Experiments and Results

Two sets of experiments were carried out that each consist of two parts. The first set of experiments consists of one experiment in which MBT is trained on the 300 annotated register entries and tested on the 200 annotated register entries and one experiment in which MBT is trained on the 300 annotated field logbook entries and tested on the 200 annotated field logbook entries. All scores presented

	Field book	register
Number of different segments	22	30
Avg. # tokens per entry (σ)	61 (32.58)	33 (22.93)
Avg. # segments per entry (σ)	16 (5.12)	12 (6.05)
Avg. # tokens per segment (σ)	4 (6.84)	3 (5.16)
Max. # tokens per entry	212	206
Min # tokens per entry	3	2
Max. # segments per entry	47	29
Min. # segments per entry	3	2
Max. # tokens per segment	133	201
Min. # tokens per segment	1	1

Table 3.1: Statistics on entries and segments in registers and field logbooks

in this subsection are computed on complete segments. The results of these two experiments are shown in Table 3.2.

Precision	Recall	$F_{\beta=1}$
Register entries trained on register entries		
75.96%	74.45%	75.20
Field logbook entries trained on field logbook entries		
82.98%	86.23%	84.57

Table 3.2: Results of experiments for register entries tested with MBT trained on register entries and for field logbook entries tested with MBT trained on field logbook entries

The second set of experiments consists of two experiments in which MBT is tested on a type of data set different from the one it was trained on, i.e., in the first experiment MBT is trained on the 300 annotated register entries and tested on the 200 field logbook entries, in the second experiment MBT is trained on the 300 annotated field logbook entries and tested on the 200 annotated register entries. The results presented in Table 3.3 show that the differences between the register and field logbooks negatively affect the performance of the tagger and illustrate the need to train a separate tagger for each data set.

The difference in F-score between the experiments in which the test data was of the same type as the training data and of the experiments in which the test data was of a type different from the training data were significant with a 95% confidence interval for both the registers and field logbooks. The significance

Precision	Recall	$F_{\beta=1}$
Field book entries trained on register entries		
47.11%	34.05%	39.53
Register entries trained on field logbook entries		
34.96%	33.15%	34.03

Table 3.3: Results of experiments for field logbook entries tested with MBT trained on register entries and for register entries tested with MBT trained on field logbooks

levels were computed using bootstrap resampling (Noreen, 1989).

The scores reported in Tables 3.2 and 3.3 show the averaged scores for all segments. However, some segments are easier to segment and label than others. Similar to the results in Lendvai and Hunt (2008), the classifier performs better on the shorter types of information (such as 'genus' and 'species') than on the longer segments (such as 'special remarks'). This is expected, as there is much variety in the latter type of fields. It is also not necessarily a bad thing that the classifier cannot predict the longer segments so well, as the information in the 'genus' and 'species' fields of the database are the most important ones. Most information requests from researchers on the data inquire after one or both of these fields. For 'genus' and 'species' MBT achieves F-scores of 90.53 and 91.37, respectively, on data from the registers and 94.36 and 88.13, respectively, on the data from the field logbooks.

3.2 A Manually Constructed Ontology

An ontology is an explicit conceptual specification of a domain (Gruber, 1993). An ontology represents knowledge about a domain by describing the classes and the relations between them for a given domain. The term ontology was first used in philosophy, denoting the systematic account of existence. It is derived from the Greek words $\acute{\omega}\nu\tau o\varsigma$ for being and $\lambda o\gamma\acute{\iota}\alpha$ for science, study or theory and used to denote the study of being.

In the context of artificial intelligence or computer science, ontologies still carry the meaning of describing the world, but they are usually focussed on a specific domain. In the remainder of this section, basic characteristics of ontologies are laid out in Subsection 3.2.1. Then, in Subsection 3.2.2, the resources

involved in building the manual ontology for the R&A database are described. In Subsection 3.2.3 the principles of ontology building are discussed, followed by the methodology used for the R&A ontology in Subsection 3.2.4. In Subsection 3.2.5, the ontology created for the R&A domain is presented.

3.2.1 Ontology Basics

There are different levels of formality for ontologies, depending on the application for which they are used. The ontologies developed for this thesis are not meant for direct automatic information integration with other resources. For that to be possible, there is not sufficient consensus on the data resources and structure at Naturalis, although this should change by the introduction of the new CRS. The main purpose of the ontologies developed here, is to conceptualise the natural history domain, and provide basic knowledge about it in order to facilitate several data cleanup tasks and aid information retrieval. Also, the data resources from which the ontology is derived are rather simple, i.e., they do not contain thousands of concepts that one would encounter in some other domains and there are not many sub- and superconcepts distinguishable.

An ontological class is an entity that encompasses all individual objects sharing particular properties that differentiate them from other classes. The notion of what properties are discriminative for the differentiation of the different classes may differ per domain and per intended use of the ontology. For example, for the R&A domain, a class describing provinces can be defined, which is used to indicate a region larger than a city but smaller than a country. To provide an indication of the locality of a specimen find for biogeographical research, the class of provinces is an acceptable level of granularity as the province is merely used to disambiguate the city name. However, for a dedicated geographical resource, such as GeoNames, such a class is too imprecise as GeoNames aims at providing detailed and fine-grained information about localities. In the data model for this resource, the types of information that are captured in the R&A class for province are split up into sixteen subclasses with a distinct hierarchy to describe whether a particular region constitutes an administrative division or a military buffer zone.

Ontologies can serve as a lingua franca between various parties operating within one domain. A layperson can, for example, use the word ‘snake’ to denote a certain concept in the domain, the domain expert will refer to it as ‘ophidia’,

and in the online registration system this concept can be denoted as 'X5087'. The aim of an ontology is to develop a framework of the domain where each of the users, whether human or machine, can use the extension of the ontology in such a way that there is common ground between their usage. By using a conceptual model such as an ontology within a domain, the data in the domain will become more transparent, and thus more inviting for other parties to contribute to and participate in (cf. Chandrasekaran, Josephson, and Benjamins (1999)).

As the natural history domain is changing (e.g., there are changes in taxonomy, research practice, and its role in society), the ontology presented in this chapter must be seen as a snapshot of the natural history domain at a point in time, because it is based on the resources available to the MITCH project. Ontologies cannot offer a complete specification of a domain for every user, but they can offer a consistent specification for every user to work with. Through the use of a commonly used conceptual reference model (see Subsection 3.2.2) and by basing the ontology on data constructed by the researchers at Naturalis, an attempt is made to construct a specification that is useful and acceptable to all parties.

A distinction is made between the type of knowledge that the ontology provides about the ontological classes and the individual objects. The generic information on the domain that is expressed by the description of the classes and their properties make up the intension of the ontological knowledge. The objects –in the natural history domain the specimens–make up the extension of the ontological knowledge.

3.2.2 Resources Involved

The domain knowledge that was used to construct the ontology was derived from the reptiles and amphibians database at Naturalis, as described in Subsection 2.2.1. Missing pieces of information such as what relation exists between the 'author' column in the database and a specimen stem from correspondence and interviews with domain experts at Naturalis (e.g., the 'author' column, as mentioned in Subsection 2.2.1, describes the publication in which the species was first defined, this cannot be deduced from the short database label without knowledge about the domain). The information on the herpetology domain was mapped to a standardised reference model for museum collections, namely the CIDOC-CRM reference model.

CIDOC-CRM

For the construction of the R&A ontology the choice was made to conform to the guidelines of CIDOC Conceptual Reference Model (Crofts *et al.*, 2008). The advantage of using a standard such as CIDOC-CRM is that the ontology can be easily integrated with other resources that use the same standard. Using a reference model also simplifies the ontology construction process somewhat as it provides constraints on the ontology to construct. The CIDOC-CRM is developed by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM) and accepted as ISO standard ISO:21127:2006.

The goal of CIDOC-CRM is to provide the underlying semantics of database schemata and document structure used in the cultural heritage to create a formal ontology. As such it does not define the terminology but rather the characteristics of relationships between different objects in a domain. The CIDOC-CRM aims to be cross-disciplinary, but it was designed from the perspective of the art and archaeological artefacts collections. Although there are differences between natural history and art collections, it is definitely possible to map animal specimen data to the CIDOC-CRM as Subsection 3.2.5 will show. This is backed up by the fact that other natural history institutions are also looking at CIDOC-CRM to represent their domain (see for example: Lampe *et al.* 2008).

The CIDOC-CRM reference model provides for the representation of far more types of concepts than those used for the R&A domain. Version 4.5.2 contains 86 entity classes (identified by labels E1 to E90) and 137 properties that define relations between classes (labelled P1 to P148). CIDOC-CRM contains for instance, concepts to denote birth dates and end of existence events which may be used to represent biographical information about persons or objects involved in the domain. Since this type of information is currently not recorded in the data sets and researchers did not state that this would be useful to them in the near future, such classes were not described in the R&A ontology. However, when necessary they can be added.

Protégé

To model the ontology the open source ontology editor Protégé version 3.4 is used[2]. Internally, Protégé uses OWL but it can also export ontologies to several

[2] http://protege.stanford.edu Last visited: 3 July 2009

formats such as RDF[3] and XML Schema[4].

OWL is a markup language for publishing and sharing data using ontologies on the Internet and is recommended by the World Wide Web Consortium[5]. OWL represents the meanings of terms in vocabularies and the relationships between those terms in a way that is suitable for processing by software.

Protégé allows for defining concepts, relations between concepts, and restrictions within the domain through a graphical user interface. Among its most important features are its prevention of the definition of duplicate concepts and automatic encoding of inheritance for relations between concepts that also hold for subsconcepts. For a complete overview of its features, the reader is referred to the Protégé manual (Horridge *et al.*, 2004).

3.2.3 Ontology Building Principles

Ontology construction has largely been a manual task and is thus dealt with differently by different parties. Chi (2007) has stated that ontology construction is an art or craft rather than a science, as an ontology also reflects the perspectives of the ontology constructors on the domain. With the growing interest in using ontologies for data sharing and integration, the use of standards and design principles has become necessary. Gruber (1993) defined the following five design principles which are widely used.

Clarity and objectivity: the ontology should effectively communicate the intended meaning of the defined terms. Definitions should be objective.

Coherence: an ontology should be coherent.

Extendibility: new general or specialised classes should be included in the ontology in such a way that existing definitions do not need to be revised.

Minimal encoding bias: the conceptualisation should be specified at the knowledge level without depending on a particular symbol-level encoding. This principle has become obsolete with the endorsement and acceptance of OWL as ontology description language in 2004.

[3] http://www.w3.org/RDF/ Last visited: 10 January 2010
[4] http://www.w3.org/XML/Schema Last visited: 10 January 2010
[5] http://www.w3.org/. Last visited: 14 July 2009

Minimal ontological commitment: the ontology should specify as little as possible about the meaning of its classes, giving the parties committed to the ontology freedom to specialise and instantiate the ontology as required. As has been noted by Gómez-Pérez (1998), this seems contradictory to the restriction that definitions should be as complete as possible but it sets a challenge for finding a balance between a generalisation and overspecification of a domain.

Some of Gruber's design principles are difficult to implement. Objectivity and coherence are, for example, hard to measure. Extendibility is also difficult to take into account when it cannot be foreseen what kind of new classes need to be added to the ontology in the future. Overall, the principles provide guidance and aid deliberate decision making in the ontology construction process.

Borgo *et al.* (1996) added a sixth design principle to ensure that hierarchies are clean and untangled. Their principle is as follows.

Ontological distinction: classes in an ontology should be disjoint.

3.2.4 Ontology Construction Methodology

Ontology construction is often carried out as a sequential bottom-up process in which several steps can be discerned. Every step provides a new level of deeper understanding and generalisation over the domain. The literature presents several variations (Uschold and Gruninger, 1996; Gómez-Pérez, 1998). For the R&A ontology, the approach that was used is taken from Noy and McGuinness (2001). This approach was chosen as it provides a clear step-by-step methodology for ontology building, it is commonly used, and is connected to the Protégé editor. The approach consists of seven steps:

Step 1: Determine the domain and scope of the ontology. The domain of the
R&A ontology is that of natural history, the scope is limited to reptiles and amphibians data, and pertains to the collection and storage of specimens.

Step 2: Consider reusing existing ontologies. When development of the R&A ontology started, there was no existing ontology that satisfied the inform-

ation need for the researchers at Naturalis[6], therefore, the CIDOC-CRM framework was adopted to ensure the ontology's easy integration with other resources and ontologies .

Step 3: Enumerate important terms in the ontology. The important terms in the ontology are taken from the R&A database, as well as from input from herpetologists. In the database, every column name is considered a term in the ontology. This does mean that some terms encompass rather broad topics and may be candidates for further splitting.

Step 4: Define the classes and the class hierarchy. The terms selected for inclusion in the ontology are matched to terms in the CIDOC-CRM reference model. Relations between the classes are defined after consulting a herpetologist who is familiar with the reptiles and amphibians collection and database.

Step 5: Define the properties of classes. The definition of properties of classes depends on the level of granularity of the previously defined classes. CIDOC-CRM defines classes up to a very fine level, where even the information given by a label connected to an artefact is considered a class instead of a property belonging to a class. The CIDOC-CRM view was adopted for this reptile and amphibians ontology. Therefore in some cases, what Noy and McGuinness (2001) consider a property of a class is defined as a class in CIDOC-CRM.

Step 6: Define the facets of the properties. Facets of properties describe what values properties can hold. If a 'Preservation Method' property were to be defined for the R&A ontology two facets could be 'alcohol' and 'dry' as these are common methods to preserve animal specimens. As the CIDOC-CRM guidelines are followed, 'Preservation Method' is considered a class, thus **Step 6** in Noy and McGuinness (2001)'s approach is not applied here.

Step 7: Create instances. The R&A database provides the instances to populate the ontology with. Therefore **Step 7** is straightforward for the R&A ontology. Since the ontology defines the classes important in this domain, instances from other databases can easily be imported into the ontology. By

[6]At the same time, a data model was being developed that will be used in the new CRS, which could have been adopted had the timing been different

populating an ontology with instances, both the intension and extension of the domain are specified and a knowledge base is formed.

3.2.5 Description of the Ontology

In this section, the ontology for natural history that was designed according to the CIDOC-CRM reference model is presented. First the seven types of classes of the R&A ontology are described, followed by the eight types of relations. The ontology is presented as a graph in Figure 3.1. The OWL source file can be found at `http://ilk.uvt.nl/~merp/NaturalisRAOntology.owl`.

Classes:

- **Specimen class:** The most important class in the domain is the animal specimen. Each database record describes all relevant facets of this object. According to CIDOC-CRM guidelines it is denoted as a class of type *E20 Biological object*. In Figure 3.1 it is located in the centre of the figure.

- **Person classes:** In the collection, acquisition, preparation, and description of the animal specimen several human actors are involved. One problem in representing these classes properly in the ontology is the accompanying actions may have been carried out by the same person, but they could have also been carried out by several persons. To resolve this, the choice was made to define the actors by the roles that need to be fulfilled for the specimen preservation rather than the persons as objects. These roles are defined through classes of type *E39 Actor* and describe the collector, recorder, determinator, donator, and author roles. Whether these roles are carried out by the same person is left to the extension of the ontology through the database instances. CIDOC-CRM does not cater for modelling of this type of information. This is probably due to the fact that is based on collection objects and persons involved in the acquisition and curation of the objects are of lesser importance.

- **Geographic classes:** In the R&A ontology, geographical classes pertain to the geographic locations that are connected to the specimen's find. These are denoted by *E53 Place* classes in the CIDOC-CRM hierarchy and in

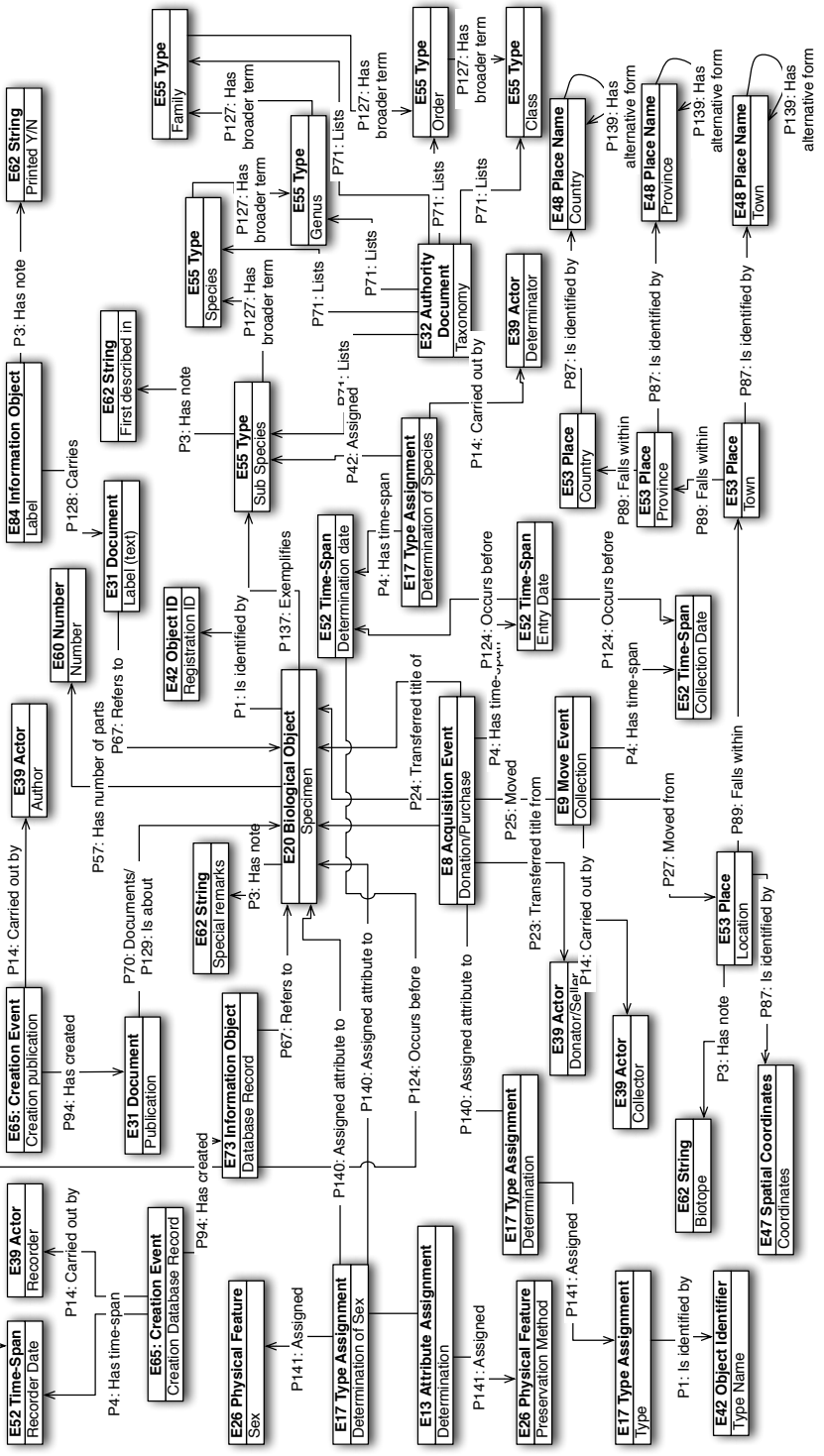

Figure 3.1: Manually Created Ontology for Natural History

the natural history ontology pertain to the location, town, province, and country that define the finding location of the specimen.

- **Events:** The processes involved in acquisition and curation of a specimen are denoted as events in CIDOC-CRM reference model. There are several types of events such as events that involve acquisition (*E8 Acquisition Event* for a donation or purchase) and *E17 Type Assignment* for an act such as the determination of the species of the specimen.

- **Dates:** Events are anchored in time by dates, which are denoted by *E52 Time-span* classes in the ontology.

- **Names and IDs:** Objects are identified by names and IDs. *E53 Place: Country* can, for instance, be identified by several names described by *E48 Place Name* identifiers. Here, one can think of several synonymous terms linked to the country object such as country names in different languages or a country code, these are linked to the preferred identifier by a *P139: Has alternative form* relation.

- **Attributes:** Additional information pertaining to the classes is described by various attributes. One such attribute is for instance *E26 Physical feature: Preservation method*. Documents, labels and any other classes such as *E62 String: Special remarks* are also considered attributes.

Relations

In CIDOC-CRM, relations are denoted as properties with a range and a scope. In this thesis, the term 'relation' is used instead of 'property'. The range is the class from which the relation originates. The scope is the class to which a relation can go. For example the range of the relation *P14: carried out by* needs to be a concept of type *E7: Activity* and its scope of type *E39: Actor* to describe a person involved in a certain activity. In CIDOC-CRM, nearly every relation also has an inverse relation. The relation *P4: has time-span*, that has a range of type *E2: Temporal entity* and a scope of type *E52: Time-span* for example, has the inverse property *P4: is time-span of* with the exact inverse range of type *E52: Time-span* and scope of type *E2: Temporal entity*. For the sake of readability these are not included in Figure 3.1, but they can be inferred from the relations shown.

- **Specimen - Event:** A specimen can be related to an event in several ways. CIDOC-CRM specifies different types of relations to link a specimen class to for instance an acquisition event or a move event. An *E9 Move Event* such as a specimen collection event is for example related to the specimen by a *P25: Moved* relation, as this event constituted the removal of the specimen from its habitat to the museum collection. An *E8 Acquisition Event* such as a donation has a different relation denoted by *P24: Transferred title of* as this event pertains more to the legal aspects of ownership. In most cases, prior to a donation the specimen was indeed collected somewhere (although some specimens were donated from for example zoos) but this is unrelated to the donation event as the specimen may have been part of a different collection prior to its entry in the Naturalis collection.

- **Specimen - Names and IDs:** A specimen is related to an ID such as a registration number by a *P1: Is identified by* relation.

- **Date - Event:** Dates and events are linked through a *P4: Has time span* relation.

- **Event - Event:** Events are related chronologically through a *P124: Occurs before* relation.

- **Person - Event:** Person classes often have an active role in the events described in the ontology for R&A. Hence their relation to an event is almost always of type *P14: Carried out by.* An exception is the relation between *E8 Acquisition Event: Donation/Purchase* and *E39 Actor: Donator/Seller* class which is a *P23: Transferred title from.*

- **Attribute - Event:** It is not possible to link a specimen directly to an attribute; this is done through an *E13 Attribute Assignment* object which on the one side is linked to a specimen through the *P140 Assigned attribute to* relation and on the other side is linked to the attribute through a *P141: Assigned* relation.

- **Geographic class - Event:** Events and geographic classes can be related through the *P7: Took place at* but also through more specific relations such as *P27: Moved from* which relates a move event to a location.

- **Geographic class - Names and IDs:** Similar to the *P1: is identified by* relation there is a *P87: Is identified by* relation especially for the identification of geographic classes. This distinction is made because the *E44 Place appellation* class can have attributes (e.g., spatial coordinates) that are different from those of the *E42 Object ID*.

The creation of the ontology via the CIDOC-CRM guidelines concludes the preparatory work. In the following chapters the main contributions of this work will be presented. Before doing so, this chapter is summarised and future improvements are discussed in the next section.

3.3 Chapter Summary

In this chapter, preparatory work was presented that makes the experiments that will be presented in the following chapters possible. In the first part of this chapter, the automatic segmenting and labelling of the field logbooks and registers was described. The segmenting and labelling task uncovered peculiarities regarding the method of manual data entry for the reptiles and amphibians database, such as interpretation of the data and choice of language. However, this barred utilising the overlap between the database and the resources as entries about the same object could only be identified by registration number and taxonomic information. Most other information had been reformatted and/or interpreted and translated in such a way that automatic annotation of the field logbooks and registers with information from its corresponding database entry was impossible. Yet, the Memory Based Tagger achieves high scores on the most important fields from the field logbooks and registers for the automatic segmenting and labelling task so that these results, with known levels of errors, could be added to the database. In particular, on the 'special remarks' and 'biotope' fields MBT's performance is low (with F-scores around 50), but as Chapter 6 will show, these are of secondary importance to index and retrieve information about the specimen collection. Although the approach taken to automatic segmenting and labelling in this chapter is not new, there is a significant improvement in results by acknowledging the difference between the field logbooks and the registers. It is a known problem in computational linguistics that results deteriorate when there is a discrepancy between the training and testing material used for the ap-

proach (Nothman *et al.*, 2009). So far, this problem has not been overcome, but it is gaining attention from the research community. However, as the results in Subsection 3.1.3 have shown, annotating a few hundred instances for a data sets such as the ones used has shown to be a small effort with a fair pay-off.

Contrary to the work performed by Canisius and Sporleder (2007) and Lendvai and Hunt (2008) different training sets were used for the field logbooks and the registers. This accounts for a significant improvement in the scores (from F=34.03 to F=75.20 for the registers and from F=39.53 to F=84.57 for the field logbooks).

In the second part of this chapter, a manually constructed ontology for the natural history domain is presented. The ontology is based on the CIDOC-CRM guidelines and provides information on how the concepts in the domain are related to each other. As the CIDOC-CRM guidelines are designed from an archive and collection management perspective the resulting ontology focusses on these aspects as well. In the ontology, several clusters are discerned that are more related to each other than others, such as a grouping of classes concerned with taxonomy and a grouping of classes related to geography. This grouping expresses a strong bias towards a hierarchical structure where each group of classes in the ontology describes which classes denote objects that are of the same type. The organisational perspective ignores the fact that the different types of information available in the domain can also be related in other ways. A particular genus of snakes occurs for example only in a certain geographical region. Such information can be useful to researchers in the domain. In Chapter 5, an automatic approach to ontology building is presented that does define such relations for the domain.

In the next chapter, improving data quality of the R&A database will be addressed.

4

Data Cleaning

To kill an error is as good a service as,
and sometimes even better than,
the establishing of a new truth or fact.

Charles Darwin, 1880

In this chapter, two data cleaning methods are presented to improve the quality of specimen databases. The first data cleaning method, called TIMPUTE, is a statistical method in which inconsistencies in data are identified through applying a machine learning algorithm. The second method, called VALIDATO, is a rule-based method in which the rules are inferred from the ontology presented in Subsection 3.2.2. Analyses of the impact on data quality of the two methods separately and jointly provide the answer to Research Question 1.

RQ1 **a:** Can data-driven and knowledge-driven methods provide improvements to the data quality of structured textual resources describing collection objects?

b: To what extent are the data-driven and knowledge-driven methods complementary?

The course of the chapter is as follows. Previous work on data cleaning is described in Section 4.1. In Section 4.2, the motivation for data cleaning is given as well as an overview of errors that can occur in databases. Section 4.2 is concluded by the results of a manual error analysis of the R&A database. It is

followed by a description of the normalisation steps carried out on the R&A database (Section 4.3). Section 4.4 presents the data-driven data cleaning method TIMPUTE. In Section 4.5, the ontology-driven data cleaning method VALIDATO is described. The chapter is concluded by discussion and conclusions in Section 4.6.

This chapter is based on the following publications:

- Sporleder, C., Van Erp, M., Porcelijn, T., and Van den Bosch, A. (2006d). Spotting the 'odd-one-out': Data-driven error detection and correction in textual databases. In *Proceedings of the EACL 2006 Workshop on Adaptive Text Extraction and Mining (ATEM-06)*, pages 41–48, Trento, Italy. ACL.
- Van den Bosch, A., Van Erp, M., and Sporleder, C. (2009a). Making a clean sweep of cultural heritage. *IEEE Intelligent Systems*, **24**(2), 54–63. Special Issue on Cultural Heritage.

4.1 The Essence of Data Cleaning

No matter how concentrated and precise manual labour is carried out, humans make mistakes. Redman (1997), Orr (1998), and Maletic and Marcus (2000) estimate that about 5% or more of information in manually created databases is incorrect. The errors are often small, but if not caught, they can harm computations and conclusions based on the data. The data mining community is aware that errors in data can be harmful and acknowledges the problem. However, often data cleanup is not the main topic of interest and is therefore brushed over as quick preprocessing step.

Database errors are introduced for a variety of reasons. An insufficiently clear database schema which confuses users can cause errors in data. This leads to different types of information being inserted in a single column because it is unclear where it should be entered. In the R&A database, wrong column errors indicate possible problems in understanding the database schema.

Error prevention is a topic to which much attention has been paid recently, in particular in the medical domain where prevention of errors can save lives (Cole *et al.*, 2006). However, there are also many data resources around for which error prevention methods were not in place at the time of entering the data, such as the databases used for this thesis. Therefore, in this thesis, the focus is on

error detection and not error prevention, although in Section 4.6 some pointers to improving the process of data entry into databases will be given. In addition to identifying errors caused by the database schema, this chapter will also deal with formatting inconsistencies, content errors, and spelling errors. The main cause for errors is the fact that the human mind is not suited for tasks that are repetitive and yet demand focus and concentration such as data entry (Amar, 2002).

In the literature on data cleaning, many different definitions are found, some more complete than others. In this thesis, the definition for data cleaning given by Chapman (2005) is adopted as it is geared towards the natural history domain.

> ***Data Cleaning:*** *"A process used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and omissions. The process may include format checks, completeness checks, reasonableness checks, limit checks, review of the data to identify outliers (geographic, statistical, temporal or environmental) or other errors, and assessment of data by subject area experts (e.g., taxonomic specialists). These processes usually result in flagging, documenting and subsequent checking and correction of suspect records. Validation checks may also involve checking for compliance against applicable standards, rules, and conventions."*

There is a fair body of work devoted to automatic cleanup of data (Rahm and Do, 2000; Wang *et al.*, 2005). Much of this work has primarily been concerned with the detection and elimination of duplicate records in databases (Bitton and DeWitt, 1983; Bobbarjung *et al.*, 2006). This is not a trivial problem, especially not in cases in which databases are merged (Hernández and Stolfo, 1998). There is an increasing body of research that addresses the problem of erroneous data within records (Knorr and Ng, 1998; Lee *et al.*, 2000; Jiang *et al.*, 2001; Kubica and Moore, 2003; Zhu *et al.*, 2004). Most approaches revolve around the detection of outliers, i.e., data points that do not conform to general patterns observed in the majority of the data which may indicate that the value is erroneous.

The research on automatic error detection and correction within database records is divided into four different types of approaches. First, statistical methods identify outliers through record value means, standard deviations, ranges etc. Such an approach is used, for instance, by Maletic and Marcus (2000), who select outliers as data points of which the values are too many standard deviations from

the mean. The second approach uses clustering methods to group records based on a distance metric: any record that cannot be added to a cluster is flagged as an outlier. An example is found in the work by Jiang *et al.* (2001) where the k-means algorithm is applied to identify data points that are at too great a distance from any of the cluster centres. These clusters are analysed and the data points contained in the smallest clusters are returned as outliers and marked as possible errors. A drawback of this method is that clustering algorithms have high computational complexity, and thus long runtimes. The third type of approaches concerns pattern-based approaches. Such approaches identify outliers through modelling the data according to patterns, and then analysing which data points do not comply with these patterns. The fourth type of approaches are are rule-based approaches. Rule-based approaches work in a similar fashion as pattern-based approaches as rules are specified to which the data must adhere (Bruni and Sassano, 2001). Pattern and association rule-based error checking consider outliers as data points that do not comply with a certain pattern. Patterns are different from association rules in the sense that patterns can be identified through partitioning, classification, and clustering methods (Aggarwal and Yu, 2001), whereas association rules often need to be hand crafted (Maletic and Marcus, 2006). The reader is referred to Pyle (1999) for a further analysis of possible problems present in data, which may harm a data mining approach.

4.2 The Importance of Automatic Data Cleaning

As data cleaning is not a well researched problem, it is hard to estimate the impact of erroneous or inconsistent data. What is known, is that incorrect data can contain crucial data points that can greatly affect outcomes of further usage. Errors can have an immense impact on modern economy (Redman, 1998), government (Karr *et al.*, 2006) and research (Veaux and Hand, 2005). Redman (1998) estimates that errors in data quality can cost a company about 10% of its revenue. Loshin (2001) even estimates that dirty data can cost an organisation 20-25% of its budget. Karr *et al.* (2006) regard issues arising from erroneous data from a government perspective, and present two case studies in which governmental data is analysed and it is shown how errors in data can impact policy or decision making.

In the natural history domain, erroneous data can lead to analyses based on incorrect information where it can, for example, cause an incorrect decision to be made regarding the preservation of a species. In the remainder of this section, the different steps in data cleaning are discussed in Subsection 4.2.1, followed by different error types that are encountered in data in Subsection 4.2.2. In Subsection 4.2.3, the error analysis of the Reptiles and Amphibians database is presented.

4.2.1 Data Cleaning Steps

The process of data cleaning can be broken down into a 5-step workflow. This workflow is based on (Chapman, 2005) and (Maletic and Marcus, 2000).

1. **Define and determine error types** Different data sets contain different types of errors. Manual inspection of a selection of the data is the main method to identify error types present in a particular data set (Dasu and Johnson, 2003).

2. **Search and identify error instances** Current data sets often exceed thousands or even millions of instances (Ramaswamy *et al.*, 2000). Hence, automatic error detection methods are imperative. For different types of errors, different detection methods have been developed

3. **Correct the errors** Once errors have been identified, correction is necessary when possible.

4. **Document error instances and error types** It is important to track which instances have been corrected and of what type. This provides insight into the error correction process and the proportion of error types in the data. Error documentation can also be used to recheck whether the error correction process was carried out properly and whether the adjustments made to the data were correct.

5. **Error prevention to reduce future errors** Preventing the introduction of errors is better than needing to detect and correct errors after data entry. Documenting error types can help the development of error prevention methods to put in place before more data is entered. To prevent formatting errors, one can think of restricting the entry field to only the

accepted format (e.g., for dates). Spelling errors and some types of inconsistencies can be prevented by the usage of accepted lists to restrict the possible values of a cell.

The data cleaning methods that are presented in this chapter are not actively concerned with the last two steps in the data cleaning process, as these are more a software and client-side business. It is stressed that any change to the data should be documented, therefore documentation of errors and error types to maximise prevention is imperative.

4.2.2 Types of Errors

Different types of errors can occur in data. In this section, different error types are described according to the classification proposed by Müller and Freytag (2005). The classification of error types is more specific than that given by, for instance, Han and Kamber (2001) who only discern between missing values, noisy data and inconsistent data. It is important to note that there are different types of noisy data, and inconsistencies that should not be treated equally. There are three main types of errors that can occur in data, that each have a subdivision of error types. Müller and Freytag's error types are described below.

A. Syntactic Errors

1. **Wrong column errors** are values that correctly belong to a particular database record, but they occur in the wrong cell. This is, for instance, the case if the information on the genus of a specimen is recorded in the column that is supposed to contain the province where it was collected.

2. **Domain format errors** are values that are not formatted properly. This is the case where the date format is specified as dd-mm-yyyy and the given value is formatted as yyyy-mm-dd.

3. **Irregularities** are cases in which values, units and abbreviations are not used uniformly. An example is a column where a part of the values is expressed in metres and another part in feet.

B. Semantic Errors

1. **Content Errors** are cases in which the value is simply incorrect. This is for instance the case if a record on a snake specimen contains the value *amphibia* in its 'taxonomic class' cell whereas it should be *reptilia*.

2. **Dependency Violations** are values that violate a dependency between different database cells. An example of this type of anomaly is the case where the value of the entry date precedes the value of the collection date, as a specimen cannot have been collected after it entered the collection.

3. **Duplicates** are two or more records representing the same object.

4. **Invalid tuples** are records that do not represent valid entities in the domain. This type of pollution could, for example, be a database record on a reptile specimen that is entered in the bird database. This type of error does not occur in the data sets used in this work which is due to the fact that there the different collections at Naturalis are treated by different departments and thus recorded in different databases.

C. Coverage Errors

1. **Missing values** are values that should have been recorded in the database but were not. Cases like this are found in the R&A database where the value for country of a specimen find is missing.

2. **Missing tuples** are database records that are completely missing. As mentioned in Chapter 2, the R&A database covers only 1/3 of the R&A collection of Naturalis, meaning 2/3 of the tuples are missing.

Most errors that are reported in natural history data are content errors that occur in the taxonomic, geographic or person name columns (Chapman, 2005). Errors in the taxonomic information regarding a specimen can be caused by an incorrect determination of the specimen. It can, for example, be the case that a specimen was determined quickly and imprecisely in the field, straight after collection. Sometimes errors in the taxonomic fields can be detected automatically as they are misspellings or inconsistencies with an accepted taxonomic resource. Some errors can only be detected through double-checking or revisiting the determination decision as part of collection maintenance.

Geographic errors are mostly induced by imprecise recording of a location in the field (e.g., 'Meyer's farm, 5km South of Sipaliwini'). There are geographic inconsistencies that can be detected automatically in a database. Those are the ones that pertain to changes in naming of locations (e.g., Ceylon vs. Sri Lanka, Bombay vs. Mumbai) or inconsistencies in the geographic hierarchy (e.g., Alaska, Canada). Modern technology such as GPS units have made it easier for collectors to record the precise location of their find.

Errors in person names are less frequent than errors in the taxonomic or geographic information about a specimen. The main error encountered is not of a semantic order but of a syntactical, namely inconsistent formatting. Person names are, for instance, given with or without initials and if given, initials are found before and after the last name. Citations in the R&A database are often incomplete, e.g., only an author is given (e.g., 'Kopstein') and the author is sometimes even abbreviated (e.g., 'L'. for 'Linnaeus, 1758'). One could argue that experts know to which publication such an abbreviation refers but for laypersons it is unintelligible and, due to its random nature, automatic indexing and linking to these publications is hampered.

4.2.3 R&A Database Error Analysis

To assess the quality of the databases at Naturalis, an error analysis was performed on the R&A database. For this study, a random sample of the database was manually checked for errors of three types: *content errors*, *irregularities*, and *wrong column errors*.

To estimate the proportion of content errors in the R&A database, all cells containing taxonomic or geographical information for 257 randomly chosen records were checked manually. The taxonomic and geographical information is spread over 10 columns, totalling 2,570 database cells to check. The choice was made to restrict the content error checking to taxonomic and geographical information because the majority of this type of information can be checked against published resources. For other types of content errors, such as preservation method, access to the specimen is necessary, which would slow down the inspection process.

Checking of the 2,570 cells took nine person hours and yielded 894 content errors (34.8%). 145 of these were caused by the use of a non-standard synonym and can for a large part be ascribed to an insufficient database structure. For

instance, in the 'taxonomic order' column the value *Sauria* occurs frequently but the correct value should be *Squamata* as Sauria is a suborder of the reptile order Squamata. However, there is no suborder field, and therefore suborder information is systematically entered into the 'taxonomic order' column. This type of inconsistency will very likely not be a problem for the researchers and curators who entered the data and use the database on a daily basis but it can cause problems for external researchers and for integrating the database with another data resource.

For spelling errors and wrong column errors, 10,000 random non-numeric filled database cells were selected and checked manually. This amounts to 3% of the database. The manual inspection took approximately three hours for one person and yielded 123 spelling errors (1.23%), and 435 wrong column errors (4.35%).

4.3 Normalisation

Normalisation provides a means to clear up data through standardisation of formatting. Normalisation is necessary to resolve ambiguity and to improve the structure of the information. Normalisation is mostly a rule-based step that occurs before other data cleaning steps.

Five types of normalisation were applied to the R&A database:

N1. Whitespace normalisation Trailing and leading whitespace within cells is removed and sequences of more than one whitespace character are replaced by a single space.

N2. Diacritics removal Diacritics were mapped to their ascii formats, converting for instance Müller to Mueller. As diacritics are not used consistently throughout the databases the problem of several forms for one name such as Mueller and Muller remains. Simple normalisation methods cannot tackle this problem as the alternative spelling variants may or may not refer to the same person.

N3. Date normalisation In the R&A database, dates occur in a variety of formats such as 18 Nov. 1982, 18-XI-1982, 18-11-1982, and in rare cases as 11-18-1982 (American style). Roman numerals indicating months and fully written out names of months are converted to their arabic counterparts and

all converted to dd.mm.yyyy format. In the case of British vs. American dates only dates for which it is possible to determine if a date is valid (when assuming most data is in British format, and a date such as 05.23.1969 is encountered) the date is converted to the British format. In ambiguous cases, such as 03.04.1938, automatic normalisation is not possible, without more information or consultation of the original sources.

N4. Person name normalisation Person names were normalised to 'lastname, firstname' or 'lastname, initials'.

N5. Tokenisation Punctuation characters are separated from adjacent characters by whitespace.

Normalisation is not feasible or useful on database columns such as the 'special remarks' column, as such columns do not contain data in a predefined format. For the columns that could be normalised the results of how many cells were changed was recorded and the number and proportion of corrections provided by normalisation is given in Table 4.1.

The scripts for the normalisation process were created by Caroline Sporleder and previously run as a preprocessing step. For the work done for this thesis, the normalisation process was revisited and the extent to which each of the database columns selected for normalisation was affected by the normalisation procedure is reported.

In Table 4.1, it must be noted that 100% of the 'recorder date' cells were normalised because, although these contained a consistent date format (yyyy-mm-dd), it was a format that was not used extensively elsewhere in the database. The amount of format consistency varies greatly; for some columns such as *Determinator* and *Donator*, 0.10% of the cell values do not comply with the preferred format, whereas for others, such as the *Determination Date*, almost half (47.28%) of the cells need to be reformatted to fit the preferred format. This strengthens the claim that normalisation, seemingly trivial, is a necessary step in data cleanup that corrects a fair amount of noise present.

4.4 Data-driven Data Cleaning

The main assumptions behind data-driven data cleaning are that (a) data elements in a database are conditionally dependent and (b) the majority of the data

Type of Normalisation	Column	# Filled	(%)	# Corr.	(%)
Diacritics					
	Author	15,043	(89.17)	1,342	(8.92)
	Collector	14,954	(88.64)	449	(3.00)
	Determinator	10,036	(59.49)	4	(0.04)
	Donator	4,395	(26.05)	50	(0.11)
	Recorder	16,870	(100)	4,209	(24.95)
Date					
	Collection date	14,288	(84.69)	4,789	(33.52)
	Determination date	2,432	(14.42)	1,150	(47.28)
	Entry date	9,144	(54.20)	497	(5.44)
	Recorder date	16,852	(99.89)	16,852	(100)
Names					
	Collector	14,954	(88.64)	1,674	(11.19)
	Determinator	10,036	(59.49)	10	(0.10)
	Donator	4,395	(26.05)	578	(13.15)
	Recorder	16,870	(100)	4,209	(24.95)

Table 4.1: Statistics on corrections provided by normalisation. The table shows the number and percentage of filled cells per database column (Filled) and how many of these were affected by the normalisation process (#Corr., given in numbers and percentages)

in a database is correct. In a specimen database such as the R&A database one can infer the information on a country in which a specimen was collected from the column containing information on the city in which a specimen was collected. A machine learning algorithm can be employed to make such inferences about data cells on the basis of the values in the other data cells in the database. Whenever the classifier predicts a value that differs from the actual value in the cell this cell is flagged as a possible error and needs to be checked by an expert. In the Subsection 4.4.1, the data-driven data cleaning approach that was developed in the Mitch project named Timpute is described. In Subsection 4.4.2 the experiments and results of error correction process on the R&A database with Timpute are presented and analysed.

4.4.1 Timpute

Following the main assumptions stated above, in the approach presented here a machine learning algorithm is applied to predict a value in a database cell, given

the values in the other database cells. Any predicted value that deviates from the original value is flagged as suspicious and presented to a user to check whether the original cell value is erroneous. This approach is inspired by Knorr and Ng (1998) who also utilise algorithms that use a distance function to assess a data point's similarity to other data points.

To check whether a particular database cell contains an incorrect value, the data cleaning problem is recast as a classification task. In this classification task, the training data for a machine learning algorithm is formed by all database records, except for the one in which a cell is to be predicted. The different database columns serve as features and the class that is to be predicted is the value of the target column in the target record (Sporleder *et al.*, 2006d). For instance, to predict the value of the 'country' cell for Object 3047 in the R&A database, a classifier is trained to predict this value based on the values of all other cells of Object 3047. All database records except the one for Object 3047 serve as training data. The prediction of the 'country' value for Object 3047 is based on the similarity of the values of other cells in the record for Object 3047 and the other records in the database. In Figure 4.1, a schematic overview of this approach is presented.

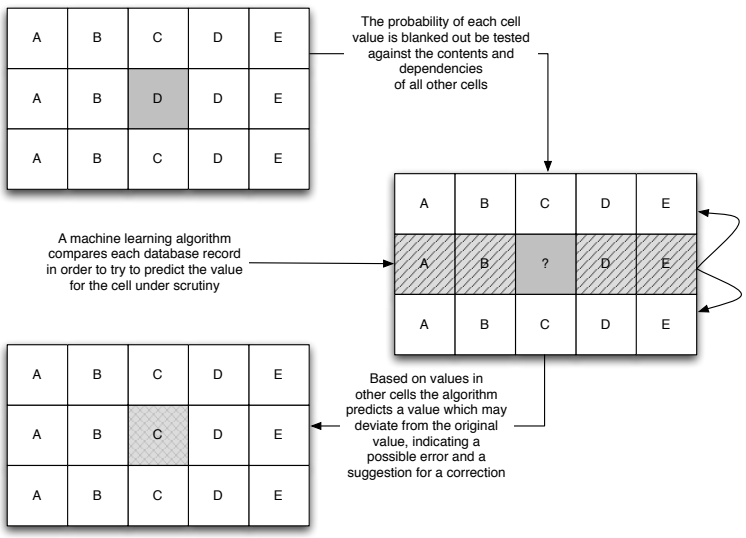

Figure 4.1: Schematic overview of data-driven error detection

This process is repeated for every database cell. In order to do this efficiently, the TiMBL implementation of the k-NN algorithm is chosen (see Chapter 3, Daelemans *et al.* (2004)). The main feature that makes this classifier particularly suitable for this task is that it does not abstract a model from the training data, instead it stores all training records in memory. In the classification phase it compares the database record to all stored records and assigns, according to some similarity metric, the class of the k most similar records to the record to-be-predicted. Because the k-NN classifier does not have to abstract a new model every time the training set changes (which is every time a new cell is to be predicted) the algorithm can perform this error detection task much faster than algorithms that do abstract a model from the training data. As in the experiments described in Section 3.1, the TiMBL implementation is used with standard settings.

It is not possible to check every type of data via this method as some database cells may not depend on other database cells. In the R&A database this is, for example, the case with the data that describes the database record, such as the time the database record was created and its ID. The independent columns are called extrinsic columns. Their opposite are intrinsic columns, which contain information that is closely linked to the specimen, such as its place in the taxonomy and its finding location.

After every database cell has been predicted by the classifier, cells in which the predicted value deviates from the original database value are presented (together with the value the classifier suggests) to a human expert for two reasons. First, it cannot be assumed that all deviations are actual errors as some database values might simply be extreme values (Chambers *et al.*, 2004). The second reason for keeping the human expert involved is because it is known that there are synonymous values in the data, which should be preserved and linked to each other rather than corrected. The expert has to judge the cases and determine (1) whether the predicted value is correct (the database value was erroneous), (2) whether the original database value is correct and the classifier made an error, (3) whether both were incorrect (this is the case where the algorithm detects an error but fails to predict the correct value), or (4) whether the classifier predicted a synonymous term.

4.4.2 Experiments and Results

In this subsection, the experiments and results of the TIMPUTE system on the R&A database are described. First, the experiments to measure precision are presented. Then the experiments that were carried out in order to estimate recall are presented.

Predictability and Precision

In this subsection, the results of the TIMPUTE experiments on the R&A database are presented. For completeness' sake Table 4.2 also shows the classifier's performance on the extrinsic columns to show indeed that most of these columns are problematic for the classifier. It is also not desirable to attempt to predict columns with a maximal entropy such as the 'record id' column, as this contains as many classes as instances and is thus perfectly unpredictable. When using this the TIMPUTE system it is therefore important to pre-select columns for the classifier on the basis of its entropy. As can be seen in Table 4.2, dependent intrinsic columns such as 'class' and 'order' can be predicted with a high accuracy. Independent and extrinsic columns such as 'collection #' and 'recorder time' are more difficult to predict for TIMPUTE.

In Figure 4.2, the accuracy of TIMPUTE is plotted against the entropy of the database column. As Figure 4.2 shows, the accuracy decreases as the entropy increases, meaning that columns with a higher entropy are harder to predict for TIMPUTE than columns with a lower entropy.

In order to check whether the deviations that were detected by TIMPUTE are true errors or mistakes made by TIMPUTE, the deviations were presented to a human judge. As checking these is time-consuming and for certain types of data requires access to the actual specimen (such as for the column 'preservation method') the precision checking was limited to only three columns namely 'taxonomic order','taxonomic class', and 'country', as checking these columns does not require access to the animal specimen and less than 100 deviations were flagged, limiting the time needed for the evaluation. The correction results are shown in Table 4.3.

As can be seen in Table 4.3, the algorithm only predicts the correct value in a minority of the cases (50% for Class, 12.5% for Order and 2.9% for Country), however, in the majority of the taxonomic cases it does detect errors and synonyms

Column	# Inst	# Classes	Type	Entropy	# Flagged	Pred.
Class	15,753	3	I	0.99	10	99.93
Order	9,367	13	I	1.75	88	99.06
Family	15,689	83	I	4.42	512	96.74
Genus	16,680	649	I	6.85	1,376	91.75
Species	9,841	780	I	7.56	887	90.98
Subspecies	2,435	176	I	5.22	97	96.02
Author	15,043	1042	I	7.56	1,602	89.35
Determination date	2,432	257	E	5.72	284	88.32
Determinator	10,036	151	I	3.93	240	97.60
Type-name	374	122	I	5.06	94	74.87
Type	513	28	I	3.24	194	76.22
Country	1,529	71	I	3.72	68	95.55
Country id	16,831	125	E	3.88	876	94.80
Province	9016	513	I	5.96	1051	88.34
Town/city	15,108	3564	I	10.02	4,109	72.80
Location	1,554	652	I	7.53	480	69.11
Coordinates	568	117	I	3.41	81	85.73
Altitude	2,428	333	I	6.92	442	81.80
Biotope	699	193	E	6.03	244	65.09
Collector	14,954	1055	I	6.02	2,069	86.16
Collection date	14,288	3239	I	10.14	4,686	67.20
Collection #	8,227	4095	E	11.03	3,507	57.37
Number	9,071	34	E	0.43	576	93.65
Preservation method	15,370	42	E	0.28	193	98.74
Donator	4,395	508	I	4.86	644	85.35
Entry date	9,144	811	E	6.08	1,070	88.30
Label information	8,795	1813	E	9.99	412	95.31
Printed	16,805	6	E	0.29	67	99.60
Sex	2,947	47	I	1.75	1,026	65.18
Publication	2,258	86	I	2.69	57	97.48
Special Remarks	9,611	2537	E	9.53	2,157	77.55
Inventory #	15,937	3	E	0.01	2	99.99
Reference #	15,937	2	E	0.00	1	99.99
Registration #	16,870	16769	E	14.03	16,827	00.25
Record ID	16,870	16870	E	14.04	16,870	00.00
Recorder	10,039	8	E	1.15	87	99.13
Recorder date	16,852	534	E	7.83	1,954	88.40
Recorder time	9,997	7554	E	12.62	9,512	04.85

Table 4.2: Measures of predictability for each database column in the R&A database. Column name is given, as well as the number of records that is filled (# inst), the number of different values (# classes), the type of the column (whether it gives intrinsic or extrinsic information about the specimen), the entropy of the class, the number of disagreements flagged by the classifier (# flagged), and the predictability of the column as given by the accuracy of the classifier (Pred.)

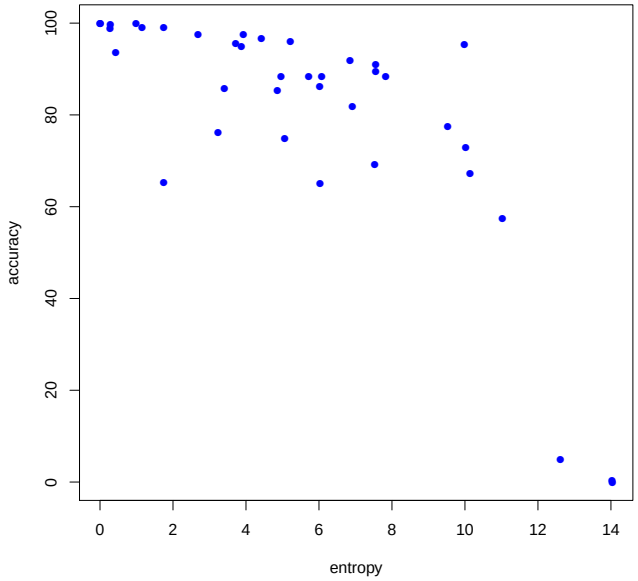

Figure 4.2: Accuracy of TIMPUTE plotted against the entropy for the reptiles and amphibians database columns

	Class		Order		Country	
# Flagged	10		88		68	
Corrected (%)	5	(50)	11	(12.5)	2	(2.9)
Detected	1	(10)	17	(19.3)	6	(8.8)
Synonyms	-		18	(20.5)	18	(26.5)
TIMPUTE Errors	4	(40)	39	(44.3)	37	(54.4)
Un- assessable	-		3	(3.4)	5	(7.4)

Table 4.3: Error correction precision, showing per column the number of disagreements flagged and presented to the expert (# Flagged), the number of cases in which the classifier predicted the right value (Corrected), the number of cases in which the classifier detected an error but did not manage to provide the correct suggestion on top of the corrected ones (Detected), the number of synonymous terms suggested by the classifier (Synonyms), the number of cases in which the classifier erroneously disagreed with the database (TIMPUTE Errors), and the number of cases that was un-assessable without access to the object (Un-assessable)

(60% for Class, 52% for Order). Input from domain experts is of vital necessity to link these synonyms properly and provide corrections for detected errors to come to a cleaned up and enriched the database.

The results presented in Tables 4.2 and 4.3 do not indicate how many of the errors present in the database are caught by the approach. As the manual error analysis in Subsection 4.2.3 showed, it is not desirable to manually detect all errors to measure recall. Therefore, a series of experiments was carried out in which the recall is estimated, these experiments are presented below.

Recall

Although the precision results of the experiments show that it is possible to improve the data, it is also important to know what portion of the data is improved, i.e., what proportion of the actual errors is caught. However, in order to measure this it would be necessary to know in advance what the errors are, which would require that the entire database be checked manually. As an alternative, the recall is estimated by introducing errors artificially into the data. Then TIMPUTE is run on the data with artificially introduced errors and the percentage of artificial errors detected by TIMPUTE is taken as an estimate of the recall performance.

The amount of errors that was introduced artificially for the recall estimation experiments is based on the manual error evaluation in Subsection 4.2.3. For content errors, 29.14% of the values of the taxonomic and geographic columns were swapped within a column. For wrong column errors, 4.35% of the values were swapped within a record for the taxonomic and geographic columns. For spelling errors, lists were compiled that contain possible options for spelling errors within a column. According to Reynaert (2005) approximately 80-87% of typos is found within Levenshtein distance 1. The Levenshtein distance is a metric that indicates how many mutations are needed to transform one sequence in the other (Levenshtein, 1966). However, certain spelling errors occur more often than others, for example, because certain keys are closer together on a keyboard. It is therefore not realistic to simply add, delete or swap a character from terms when artificially introducing errors for the recall experiment. Therefore, a list of possible spelling errors for every taxonomic and geographic database column was compiled that contained all terms as well as terms that occur in the same column at Levenshtein distance 1 to introduce spelling errors. This list was then used to automatically insert a spelling error into 1.23% cells of the taxonomic

and geographic columns. It must be noted that in the evaluation of recall error correction possible prediction of synonyms is not investigated.

In Table 4.4, the numbers and percentages of spelling errors and wrong column errors detected by TIMPUTE is given. In Table 4.5, the numbers and percentages of content errors detected by TIMPUTE is given.

Column	Pred.	# Typos	# Det.	(%)	WC	Det.	(%)
Class	80.21	193	193	(100)	745	745	(100)
Family	62.34	192	191	(99.48)	742	742	(100)
Order	67.01	193	193	(100)	744	741	(99.60)
Genus	59.02	205	199	(97.07)	788	788	(100)
Species	56.70	204	195	(95.59)	787	786	(99.87)
Author	57.38	185	172	(92.97)	711	710	(99.86)
Country	59.84	18	18	(100)	72	72	(100)
Province	55.77	110	104	(94.55)	426	426	(100)
Town/city	51.19	185	183	(98.92)	714	713	(99.86)
Location	67.05	19	19	(100)	73	73	(100)
Preservation method	93.12	189	189	(100)	727	720	(99.03)
Special Remarks	73.90	118	105	(88.98)	454	447	(98.46)
Biotope	62.96	25	20	(80)	92	91	(98.91)
Collector	82.61	184	181	(98.37)	707	703	(99.43
Determinator	91.79	124	124	(100)	474	468	(98.73)
Donator	80.02	54	54	(100)	207	206	(99.52)
Type	71.71	6	2	(33.33)	24	24	(100)
Publication	32.94	27	27	(100)	106	106	(100)
Recorder	60.93	208	208	(100)	797	786	(98.62)

Table 4.4: Results of TIMPUTE on recall estimate for typos and wrong column errors (WC) per column in number of detected errors (# Det.) and percentages (%)

As Tables 4.4 and 4.5 show, almost 100% of the artificially introduced errors are caught. However, the predictability of the columns deteriorates in comparison to the experiments presented in Subsection 4.4.2, so more possible errors are flagged. The decrease in predictability is probably due to the fact that the artificial errors that were introduced were added up to the existing errors which changed the distribution of values over the columns in such a way that the classifier had more trouble to discern patterns. The error correction is quite low when the classifier predictions are compared to the original cell value (i.e., before the artificial error was inserted). The scores for error correction lie between 0 and 50% for the correction of spelling errors, 0 and 41% for wrong column errors and 0 and 42%

Column	# Content errors	# Detected	(%)
Class	4590	4589	(99.98)
Family	4571	4571	(100)
Order	4589	4342	(94.62)
Genus	4860	4860	(100)
Species	4852	4852	(100)
Author	4383	4383	(100)
Country	445	445	(100)
Province	2627	2627	(100)
Town/city	4402	4402	(100)
Publication	657	657	(100)

Table 4.5: Results of TIMPUTE on recall estimate experiments for content errors

for content errors. This is not surprising, as the manual inspection of errors in Subsection 4.4.2 also showed that error correction percentages are low. The fact that almost 100% of the artificially introduced errors is detected indicates that TIMPUTE can provide significant improvements in data quality by reliably detecting database cells that may contain an error. This effectively reduces the workload for the human expert as she only needs to check a small number of cells per database column instead of the entire column.

4.5 Ontology-driven Data Cleaning

In addition to applying a data-driven, or *soft-reasoning*, data cleaning approach, a rule-based, or *hard-reasoning* data cleaning approach was tested. The approach that will be presented in this section, called VALIDATO, also utilises the fact that the information in some columns of the R&A database are dependent on each other, which poses constraints on the content of such columns. If the value in the 'taxonomic order' field of a record is *Anura* (=frog), for example, then the value for 'taxonomic class' should be *Amphibia* according to the zoological taxonomy. If this is not the case, then either the value for 'taxonomic order' or for 'taxonomic class' is incorrect, or both are incorrect. For the construction of the rules to test whether values in the database comply with constraints given by knowledge about the domain, the manually constructed ontology that was presented in Subsection 3.2.2 is taken as source of knowledge of the domain. Constructing rules from an ontology is a widely used approach in the Semantic

Web community (Berners-Lee *et al.*, 2001).

Similar to manual ontology construction, the construction of rules to assess consistency of a data set is a process that is not easily ported to other domains. Therefore, it is not applied to all classes in the natural history domain, but only to the concepts that were deemed the most common to the natural history domain (i.e., the temporal, geographical and taxonomic classes in the R&A database). In addition to this, the class had to be related to another class or other classes through a dependency relation. An example of such relation is the *Has broader term* relation that exists between the *Species* and *Genus* classes in the ontology presented in Chapter 3, as it is possible to check whether the value of the *E55 Type: Genus* of a particular instance is indeed a broader term of the value of the *Species* class. A relation that does not facilitate such checking is, for instance, *Is identified by* relation that exists between *Specimen* and *Registration ID*; it is not possible to infer the value of *Registration ID* from any of the other classes.

The remainder of this section is organised as follows. In Subsection 4.5.1, related work in rule-based error correction is described. In Subsection 4.5.2 the VALIDATO approach is described, followed by the experiments and results in Subsection 4.5.3.

4.5.1 Related Work

An approach that utilises the dependency between database columns is that of Kalashnikov and Mehrotra (2006) who exploit relationships between database columns to identify and disambiguate different references to one object, thus addressing the record linkage problem mentioned in Chapter 3. The work by Kalashnikov and Mehrotra can only take on one particular type of inconsistency, this is inherent to the method as rules are very precise but can only have a small scope (Kalashnikov and Mehrotra, 2006). Other work on ontology-based data cleaning has been carried out by Milano *et al.*; Kedad and Métais; Gong and Mu but with a different focus to the work done for this thesis. Milano *et al.* (2005) utilise an ontology to check a database structure and not its content. Ontologies are also used to facilitate merging or linking of different data sets as described by Kedad and Métais (2002). More relevant to the work done for this thesis, is the work by Gong and Mu (2000) who check the data values in a spatial database through rules based on relationships between objects that are checked against geo-

graphical data from other resources. The ontology-driven data cleaning method that is presented in the next subsection also relies on external resources, but in addition to geographic resources, also taxonomic resources are employed.

4.5.2 VALIDATO

Knowledge about a domain can be used to identify inconsistencies in data from a particular domain. This is the main assumption behind the ontology-driven data cleaning method presented here. As mentioned in Chapter 1, general knowledge about a domain can be captured in an ontology. This was put into practice by the development of an ontology for the natural history domain in Chapter 3. Constraints that hold for the classes in the ontology should also hold for the individual objects in the domain that are described by the records in the database. By making constraints from the knowledge explicit, database instances can be identified that do not satisfy these constraints, indicating possible errors in the data. This approach is different from normalisation in that it addresses content restrictions, whereas normalisation is mainly concerned with formatting restrictions and also does not exploit the interdependencies that exist between different database cells. The ontology-driven data cleaning approach is complementary to the data-driven approach presented in the previous section, as it, as Subsection 4.5.3 will show, selects different database values to check for an expert. In addition to that, as it is a hard-reasoning approach, it ensures that cases that do not comply with the rules are flagged, whereas the data-driven approach may let some inconsistencies slip that are not sufficiently significant to the classifier.

The knowledge used in the VALIDATO experiments comes from the natural history ontology as described in Subsection 3.2.2 and the geographic and taxonomic resources described in Subsections 2.3.1 and 2.3.2 respectively. A schematic overview of the system is given in Figure 4.3. The ontology is represented on the left-hand side. In this figure, the operators $>$, $<$ and $==$ and $!=$ are used to express possible relations in a domain. In the schematic overview of the system there is a $>$ relation between classes A and B and an $==$ relation between classes B and C. These relations are imposed on the database, which is represented on the right-hand side of the figure. The classes are translated to the database columns and the ontological relations as relations between the database columns. If values in the database do not comply with the relations or rules that hold between the

database columns, such as those between $a2$ and $b2$ and between $b1$ and $c1$, they are flagged as possibly erroneous and returned to the user to validate the system's decision.

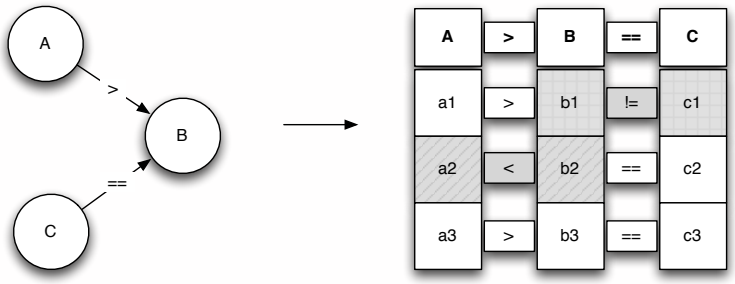

Figure 4.3: Schematic overview of the ontology-based error correction approach

In the remainder of this section, the experiments and results of the data cleaning study with VALIDATO are presented.

4.5.3 Experiments and Results

In the VALIDATO experiments, the consistency of two values from two different database columns were checked against the ontology and resources' restrictions. This means that instead of testing whether every value in a database record concerning a specimen's position in the taxonomy is consistent with the resource (i.e., 'subspecies', 'species', 'genus', 'family', 'order', and 'class'), one part at a time of this chain was investigated in a bottom-up fashion (e.g., 'subspecies' - 'species', 'species' - 'genus', 'genus'-'family', etc.). This choice was made for two reasons. First, it made it possible to zoom in on the fields of the database records for which the data did not comply with the knowledge about the domain, instead of having to check all fields in the record. The second reason is, that it is simpler to check from the bottom up than from the top down; it is easier to check whether a particular city lies in a particular country by looking up the city and then its country in an atlas than looking up the country and then going through all cities in it.

The experiments were divided into three groups: temporal, geographical, and taxonomic experiments and are presented separately in this subsection. The choice was made to limit the approach to these columns as these contain the most

interdependent data in the database and these types of information are more prevalent and therefore the results from the experiments presented here could more easily be reused than experiments on, for example, the concepts describing a publication related to a specimen.

Temporal

To identify inconsistencies in the temporally related information in the database, the database columns containing date information were selected for inspection. The four columns, 'collection date', 'entry date', 'determination date', and 'recorder date', correspond to *E22 Temporal Entity* classes in the manually constructed ontology and are interrelated by *P120: Occurs before* relations. The chronological order of the events related to these dates are summarised in Table 4.6. Inferred relations are also listed as the database is not complete and in some cases a value cannot be checked against its direct neighbour in chronology as it is missing in the database but it can be checked against the next neighbour. Such a relation is, for instance, present between the *Collection* and *Creation of Database Record* concepts. In the chronological course of the animal collection and registration process the *Collection* occurs first, after which the *Entry in Collection* takes place. The *Entry in Collection* is followed by the *Determination* event and then the *Creation of Database Record* takes place. However, this also means that the events in the chronological chain also inherit the order from the later or earlier events and thus it can be inferred that the *Collection* takes place before the *Creation of Database Record* event. This is indicated by the inferred relations in Table 4.6.

Event	Relation (→)	Event
Collection	Occurs before	Entry in Collection
Entry in Collection	Occurs before	Determination
Determination	Occurs before	Creation of Database Record
Inferred		
Collection	Occurs before	Determination
Collection	Occurs before	Creation of Database Record
Entry in Collection	Occurs before	Creation of Database Record

Table 4.6: Summary of chronological relations present in specimen data

The temporal information regarding *Collection* is described by column 'col-

lection date'. The temporal information pertaining to *Entry in Collection* is described by 'Entry date'. For *Determination* the temporal information is described by 'Determination date' and for *Creation of Database Record* the temporal information is described by 'Recorder date'.

The results of the consistency check on dates are presented in Table 4.7. The overlap with the results from the data-driven error detection experiments is also reported. When a constraint violation is detected by the rule, it is not possible to determine which date is the one containing an erroneous value.

In order to assess the overlap between TIMPUTE and VALIDATO, for both dates that were flagged as suspicious a look-up was performed to check whether TIMPUTE also regarded this value as suspicious. The results for this are reported as overlap with TIMPUTE in column 1 (OTC1), overlap with TIMPUTE in column 2 (OTC2), and overlap with TIMPUTE both columns (OTCB) in Table 4.7. Here, OTC1 reports the number of cases in which TIMPUTE flagged the same inconsistencies as VALIDATO for the first column, OTC2 reports the same but for the second column, and OTCB presents the intersection of these where TIMPUTE flags a possible error in both columns. For the 'collection date'-'entry date' pair this means that in OTC1 the number of overlapping cases for the 'collection-date' from TIMPUTE are reported, OTC2 reports the number of overlapping cases for the 'entry date', and OTCB reports the cases in which TIMPUTE flagged both the value for 'collection date' and 'entry date'. In total, TIMPUTE flagged 4,686 possible errors for the 'collection date' column, 1,070 for the 'entry date' column, 284 for the 'determination date' column and 1,954 for the 'recorder date' column, considerably more than VALIDATO.

Columns	Flagged by VALIDATO	OTC1	OTC2	OTCB
Collection - Entry	64	31	13	13
Entry - Determination	7	7	7	7
Determination - Recorder	26	5	8	2
Collection - Determination	5	1	5	1
Collection - Recorder	5	3	1	1
Entry - Recorder	2	0	0	0

Table 4.7: Results of ontology-based error detection experiments on temporal data

VALIDATO cannot suggest a correct date if a constraint violation is encountered as the domain has no rules about how much time there should be between the

different events. This has the additional effect that only cases are flagged in which a value is clearly violating constraints imposed by the ontology, and cases in which the value is incorrect but does not violate the constraints remain undetected. The data-driven approach presented in Section 4.4 may also not remedy these cases as they may be so inconspicuous that in order to detect these more information is needed. Such information could come from resources that describe the period during which expeditions took place or when a determinator was employed at the institute from which constraints with a finer granularity can be defined.

Geographical

To detect inconsistencies in the geographical information, such as a record that contains a value for a city and an incorrect country, e.g., city: Paris, country: Italy, the *Falls within* relations are translated to rules that flag pairs of database cells that do not comply with this restriction. In order to do so, the values from the different cells are looked up in the GeoNames database, and if there was no containment relation found in the returned records the database entry was flagged as containing possibly inconsistent geographic information.

The relations that hold between selected geographic classes in the specimen database are summarised in Table 4.8.

Class	Relation	Class
City	Falls within	Province
Province	Falls within	Country
Inferred		
City	Falls within	Country

Table 4.8: Summary of relations holding between the different geographic classes in the specimen data

Due to the multilingual nature of the data the rules left room for variation. If for instance the city-country pair 'swamp ca . 10 km E . of Parga'-'Griekenland[1]' was encountered, the city value was first stripped of all non-capitalised and numeric tokens and tokens shorter than 2 letters. This resulted in the value 'Parga', which was then queried against the GeoNames database resulting in 20 records returned (a part of these records are shown in Figure 4.4).

[1]The Dutch name for Greece.

20 records found for "Parga"

	Name	Country	Feature class	Latitude	Longitude
1	Parga Barga,Parga,Πάργο,Παρга,پارگه,پارغه	Greece	populated place population 2,379	N 39° 16' 59"	E 20° 23' 47"
2	Parga	Ivory Coast,	intermittent stream	N 8° 57' 0"	W 5° 38' 0"
3	Parga El Salvador,San Salvador	Spain, Galicia	populated place	N 43° 12' 0"	W 7° 50' 0"
4	Parga Estacion Parga,Estación Parga	Chile, CL.09	populated place	S 41° 13' 0"	W 73° 30' 0"
5	Parga San Breijo,San Brejome,San Bréjome	Spain, Galicia	populated place	N 43° 10' 0"	W 7° 49' 0"
6	Tītī Pārgā تیتی پارگا	Iran, Ardabīl	populated place	N 37° 14' 8"	E 48° 58' 54"
7	Río Parga	Spain, Galicia	stream	N 43° 10' 0"	W 7° 43' 0"
8	Parga Santa Cruz	Spain, Galicia	populated place	N 43° 12' 0"	W 7° 48' 0"

Figure 4.4: Screenshot of the first eight results returned by GeoNames for query "Parga"

The approach was quite lenient as for every returned result, all alternatives for the country name in the languages present in the R&A database were looked up and compared to the original country value 'Griekenland'. In this case, there is a positive match as 'Griekenland' is the Dutch word for 'Greece' and thus the record is not flagged as containing inconsistent geographic information.

Ideally, the database contains information in a single language, but this is not the case for the R&A database. It is imaginable that information in GeoNames is used to fill an extra database column with the name of a geographical entity in one particular language chosen for the database. For the 'Griekenland' example this could result in an extra column that contains the term 'Greece' so when a user searches for objects found in 'Greece', the objects that originally had 'Griekenland' in the 'country' field are also retrieved. This is outside of the scope of the work done for this thesis but looked into for the implementation of the CRS at the Naturalis research laboratories. In cases where no match between the city and country values is found in GeoNames the database entry is flagged as containing an inconsistency and the country name of the country for which most hits were found is returned as suggestion. A similar process is carried out for all 'province' - 'country' and 'city' - 'province' value pairs

In Table 4.9, the number of instances flagged per column is presented, as well as the overlap with the TIMPUTE results presented in Section 4.4. The little overlap indicates that the data cleaning methods are complementary.

Column	# Flagged by VALIDATO	#Flagged by TIMPUTE	Overlap
City	38	4,109	15
Province	23	1,051	5
Country	47	68	6

Table 4.9: Number of disagreements flagged per column in all geographic experiments in total (Flagged) and number of instances flagged by both the ontology-driven and the data-driven data cleaning approaches (Overlap)

The results of the experiments are presented per pair of columns in Table 4.10. The disagreements flagged by the ontology-driven data cleaning approach were analysed and classified as either being cases in which one of the terms could not be found in the geographic resource (NF), cases in which the value was correct but in the wrong column (wrong column errors, denoted by 'WC' in the table), cases in which a content error is detected (CE) and cases in which the database uses a synonym that is not found in GeoNames (SYN). The numbers in brackets indicate how many of the cases were unique errors.

Columns	# Flagged	NF	WC	CE	SYN
city - province	51	30 (6)	1	20 (4)	0
province - country	1	0	1	0	0
Inferred relations					
city - country	55	8 (6)	15 (4)	1	31 (4)

Table 4.10: Results of ontology-based error detection experiments on geographical information

The most prevalent cause for the system to flag a possible error is the non-standard usage of the 'city', 'province' and 'country' columns. In the 'city' column, values are found such as '4 km W. van airstrip Tafelberg' and 'Rechter kabalebo rivier, kamp keyzer, voet K. valle'. It is indeed a dilemma for the person entering the data as sometimes a specimen is not collected in a city, but somewhere near it. However, entering data in a '5km NW of' or 'near' format makes it very hard to disambiguate the location of a specimen find. Modern technology can aid in such cases as a location could unambiguously defined by the usage of a GPS device. However, to prevent this in older data such as the data at hand, the column should perhaps be renamed to 'city or nearest city' and a separate column could be devised in which the additional information such as '4km W. of airstrip' could be entered.

Most errors in the city-province experiments were a systematic mix-up of the two Surinam districts *Nickerie* and *Sipaliwini*, this is probably caused by many specimens being collected there and the erroroneous value proliferating. In the cases where the value in the city field could not be matched properly there was often a very common city name involved (such as *St. Jean*) and a province value that could not be matched (*Dep. Guyane*) because it was, for example, abbreviated in non-standard way and a non-English term is involved. This brushes upon the limitations of GeoNames as *Departement Guyane*[2] would have matched.

The fact that there is only one error found for the province-country combination is that the country field is fairly often empty. The entry that is flagged as erroneous contains the continent value 'Zuid-Amerika' (South America) in the country field and the value 'South America' in the province field.

The majority of the errors in the 'city' - 'country' experiments is caused by the fact that a term is used for the country name that is not present in GeoNames (e.g., *U.S.A.* for *United States*). In nine cases, the name from the city cannot be disambiguated properly by GeoNames, for instance through a typo. It occurs that the database contains *La Rochette - Luxemburg*, and the system suggests *La Rochette - Belgium*, whereas the value could also be *Larochette - Luxemburg*. For such cases, it is of vital importance that an expert checks the suggestions of the system. The approach also detects wrong column errors such as the value 'Kaapverdische eilanden' ('Cape Verde Archipelago') that is present in the 'country' field where it should be Brazil for a record with the value Ilha de Santa Luzia in the 'city' field.

Taxonomic

Taxonomic inconsistencies in the data are detected through a similar process as the detection of geographic inconsistencies. The taxonomic hierarchy in CIDOC-CRM is defined through the *Has broader term* relation. This relation applies to 'species', 'genus', 'order', 'family' and 'class' consecutively as shown in Table 4.11. The 'subspecies' level could not be queried as the data formatting of the resources prevented reliable identification of the 'subspecies' values.

As Table 4.12 shows, there is hardly any overlap between the disagreements flagged by the data-driven cleaning approach and the ontology-driven data cleaning approach. Only for the species column do the approaches agree on the majority

[2] Although the full term is Département de la Guyane.

Taxonomic Level	Relation	Taxonomic Level
Species	Has broader term	Genus
Genus	Has broader term	Family
Family	Has broader term	Order
Order	Has broader term	Class
Inferred		
Species	Has broader term	Family
Species	Has broader term	Order
Species	Has broader term	Class
Genus	Has broader term	Order
Genus	Has broader term	Class
Family	Has broader term	Class

Table 4.11: Summary of hierarchical taxonomic relations holding between the different taxonomic levels

of the disagreements. To investigate the cause of the difference in the approaches the flagged instances were analysed and classified as either being cases in which one of the terms could not be found in the taxonomic resources (NF), cases in which the information was correct but did not belong in that column (LE), and cases in which the system identified an content error (CE). The results are presented in Table 4.13.

Column	# Flagged by VALIDATO	# Flagged by TIMPUTE	Overlap
Species	220	887	206
Genus	3,261	1,376	286
Family	3,651	512	111
Order	8,514	88	79
Class	220	10	4

Table 4.12: Number of disagreements flagged per column in all taxonomic experiments in total (Flagged) and number of instances flagged by both the ontology-driven and the data-driven data cleaning approaches (Overlap)

The most peculiar result from the taxonomic ontology driven data cleaning experiments is the extraordinary number of wrong column errors found for the order column. Some 5,600 of these cases can be ascribed to the value *Sauria* being present in the 'order' column, whereas it denotes a suborder of reptiles of the *Squamata* order. This error is rarely caught by the data-driven approach as *Sauria* is used so systematically for *Squamata* that it is not an outlier in this

Columns	# Flagged	NF		LE		CE	
Species - Genus	4,122	3,035	(300)	0		1,087	(142)
Genus - Family	3,341	514	(81)	14	(1)	2,813	(124)
Family - Order	8,641	1,017	(23)			7,624	(66)
Order - Class	8,460	2,643	(6)	213	(6)	5,604	(2)
Inferred relations							
Species - Family	4,311	2,890	(215)	0		1,421	(91)
Species - Order	6,097	2,909	(202)	0		3,188	(362)
Species - Class	251	64	(7)	0		187	(3)
Genus - Order	8,583	515	(84)	0		8,068	(440)
Genus - Class	562	518	(81)	14	(1)	30	(11)
Family - Class	675	645	(21)	0		30	(9)

Table 4.13: Results of ontology-based error detection experiments on taxonomic information

database.

Incompleteness of the resources accounts for the majority of the cases in which the taxonomic name could not be found in the resource (e.g., the genus *Astylosternidae* is not described in Frost 2009, but it is listed in, for example, the Encyclopeadia of Life[3] and the Global Biodiversity Information Facility[4]). In a few cases, a value cannot be matched because of a spelling error such as *Alligatoridaer* instead of *Alligatoridae* or abbreviations such as *sp.* in the species field to indicate that the species has not been identified and that it could be any species in the genus indicated (in this case genus *Typhlops*). In some cases, the ontology driven cleanup uncovers an update to the taxonomy such as for the genus-family pair *Dendrobatidae-Mannophryne*. Here the approach suggests *Aromobatidae* as value for family which can be explained by a change in the taxonomy as in 2006 Aromobates was removed from the Dendrobatidae family into its own family, Aromobatidae (Grant *et al.*, 2006).

Overall, the approach detects a variety of error types and except for the cases in which the term is not present in the resource all cases it flags are real errors. As the prevalence of values of which the type (e.g., suborder vs. order) is quite high, the addition of a suborder column in the database might be considered.

[3] http://www.eol.org/ Last visited: 16 July 2009
[4] http://www.gbif.org/ Last visited: 16 July 2009

4.6 Discussion and Conclusions

This chapter has given an overview and analysis of data quality problems in collection databases. The context of quality problems will be different when data from other domains is investigated but the methodology and tools can be ported to the other domain as Van den Bosch *et al.* (2009a) have shown. In this article, TIMPUTE was applied to object databases from different cultural heritage institutions.

Through the experiments presented in this chapter and their analyses, **RQ1** can now be answered.

RQ1a: Can data-driven and knowledge-driven methods provide improvements to the data quality of structured textual resources describing collection objects?

The answer to **RQ1a** is positive as both TIMPUTE and VALIDATO have shown to be able to detect errors in a collection database. For TIMPUTE it was possible to estimate the recall of the method through an experiment in which errors were introduced artificially. The number of detected errors was high in the 90% range, hence TIMPUTE can be considered a usable tool for error detection. Error correction proves to be a more difficult problem as in many cases where TIMPUTE identifies an error it is yet unable to provide the correct answer. For the temporal VALIDATO experiments this was also the case, but due to the usage of external resources to check the taxonomic and geographic information in the R&A database with VALIDATO, for these types of information it was in most cases possible to provide a correction.

RQ1b: To what extent are the data-driven and knowledge-driven methods complementary?

TIMPUTE and VALIDATO are complementary in the sense that each detects a different type of errors, this is inherent to the underlying approaches. TIMPUTE focusses on outliers, whereas VALIDATO focusses on general patterns. It is for this reason that TIMPUTE cannot detect structural problems with the database such as data that is consistently entered into a wrong column. VALIDATO can detect such problems and thus implicitly propose alteration of the database schema. Hence, the answer to **RQ1b** is positive.

A limitation of both approaches is that they utilise the dependencies of data points within the object database. However, there are data sets which do not contain interdependent data. For these data sets error detection and correction is limited to statistical and manual methods.

Naturally, there are disadvantages to every approach. For VALIDATO, the first disadvantage lies in its lack of generalisability, i.e., for every new domain an ontology needs to be available, from which rules must be inferred and possibly the data should be linked to additional resources such as GeoNames and the taxonomic resources that were used here. A second disadvantage of VALIDATO is that the rules cannot always suggest corrections (such as the temporal rules) and can be heavily dependent on external knowledge, which should (a) be available and (b) be of high quality. When such resources are available the ontology-driven data cleaning can zoom in on the inconsistency and provide a suggestion for what the value should be. This could even be extended to correct missing values in some cases such as filling in missing country names if the city and province are known. Again the system is not perfect, as sometimes it matches to an incorrect country name in GeoNames or cannot be disambiguated, therefore a setup in which an expert checks the suggestions of the data cleaning system is recommended.

When high quality taxonomic resources are available, VALIDATO may also utilise the species locality information from the accepted taxonomic resource to check whether the geographical location given is compliant with the normal location of the species. If this is not the case then it could mean that there is an error in either the taxonomic information or the geographic information in the database record.

One area that deserves great attention is the prevention of errors, which was not explicitly investigated here as the data was already entered into the databases. However, the results from the work done in this chapter will be used to formulate recommendations for error prevention for future data entry projects. The recommendations will include the following topics:

Limit formatting variation When the entry field is constrained to only several options (e.g., the altitude field should always have a value in metres, dates should always be in dd.mm.yyyy format, the language should always be Dutch) formatting inconsistencies will be reduced drastically.

Spell check A typo is easily made, therefore a domain-specific spelling checker run in the background can prevent some errors.

Consistency check on taxonomic and geographical information The taxonomic and geographic database columns are among the most dependent chains of information found in the specimen databases. They are also the most important. Limiting the options for data entry from the highest level down can help prevent inconsistent taxonomic and geographic descriptions. When for instance for the taxonomic class the value reptile is entered (or chosen from a list) the system can prevent the user from entering a value in the order field that denotes an amphibian. To achieve this it is imperative that taxonomic resources be made available.

Timpute check It is thinkable that the data-driven error detection approach is run in the background during data entry. As soon as a value is entered that the classifier finds suspicious on the basis of the other information in the database it can prompt the user to check the value again. If the user decides to accept the value, TIMPUTE is retrained to accept also this value.

5

Automatic Data Structuring

No house should ever be on a hill or on anything.
It should be of the hill. Belonging to it.
Hill and house should live together each the happier for the other.

Frank Lloyd Wright, An Autobiography (1932)

In this chapter, The Wikipedia Instance Based Induction of Ontologies (TWIBIO) system is described, that was developed in the framework of the MITCH project. This automatic ontology induction technique links the R&A database, which contains implicit domain information, to Wikipedia, an external encyclopaedic resource that contains explicit domain information, to discover relations that hold between concepts in the R&A domain. The ontology created by TWIBIO is then compared to the manually constructed ontology presented in Chapter 3. This chapter provides the answer to Research Question 2.

RQ2 Do automatic methods for building ontologies provide different structure for data that is not achieved by manual ontology building?

The chapter is set up as follows. In Section 5.1, common methods in automatic ontology construction are described. In Section 5.2, the instance-based automatic ontology induction method that forms the second thesis contribution is presented. Section 5.3 presents a discussion of the differences between the automatically constructed ontology presented in Section 5.2 and the manually constructed ontology

presented in Chapter 3. The chapter is concluded by a chapter summary and discussion (Section 5.4).

The chapter is based on the following publications:

- Van Erp, M., Van den Bosch, A., Wubben, S., and Hunt, S. (2009b). Instance-driven discovery of ontological relation labels. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH - SHELT&R)*, pages 60–68, Athens, Greece. ACL

- Van Erp, M., Lendvai, P., and van den Bosch, A. (2009a). Comparing alternative data-driven ontological vistas of natural history. In *Proceedings of the eighth International Conference on Computational Semantics (IWCS-8)*, pages 282–285, Tilburg, The Netherlands

5.1 Automatic Ontology Construction

In the past decade, an increasing amount of work is invested in the development of support systems for automatic or semi-automatic ontology construction, with tracks and workshops devoted to this topic at several AI conferences such as ECAI and IJCAI (Buitelaar *et al.*, 2005). The task of automatically constructing an ontology from textual resources is also known as ontology learning, a term that was introduced by Maedche and Staab (2001).

Most approaches in automatic ontology construction follow a bottom-up approach, similar to the ontology construction guidelines presented in Subsection 3.2.4, i.e., starting by identifying specific terms and then generalising them into classes and defining the hierarchy and relations between them. This approach is visualised by the ontology layer cake in Figure 5.1 (Cimiano, 2006).

When taking the R&A database as a basis for the approach that is presented in this chapter (called Twibio), the database columns can be considered the classes in the ontology again. Therefore, Twibio will only focus on discovering relations between concepts. This collapses the task of identifying class hierarchies, relations, and relation hierarchies into one task. Due to their difficulty, the higher level tasks in ontology learning (identifying axiom schemata and general axioms)

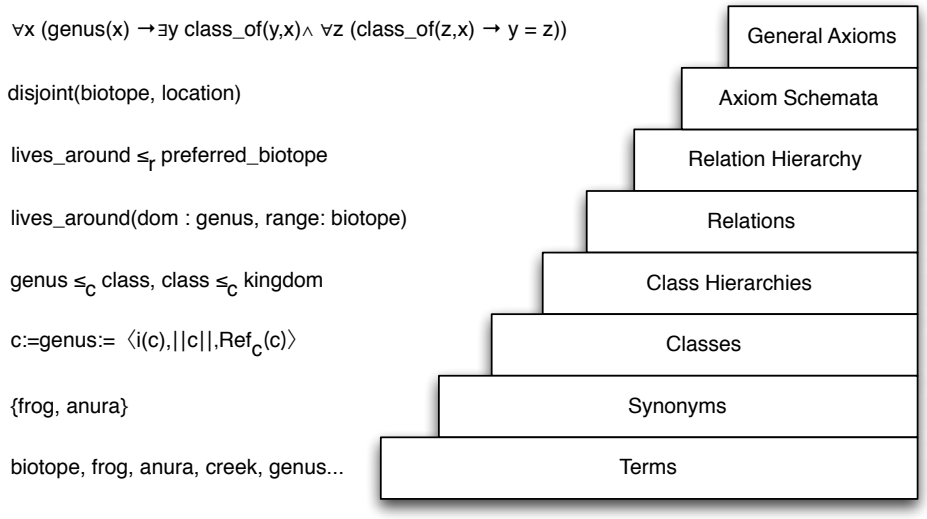

Figure 5.1: Ontology construction layer cake, based on (Cimiano, 2006)

have not been addressed extensively in previous research yet. These topics are also out of the scope of this thesis.

In the remainder of this section, the three tasks in automatic ontology construction that are addressed by TWIBIO, namely identification of class hierarchies, relations, and relation hierarchies, are discussed along with pointers to relevant literature.

Class Hierarchies

A class hierarchy provides the structure that represents which classes are parents of a class (i.e., which classes are more general, for example a 'taxonomic order' class is a parent of an 'genus' class) or which classes are siblings of each other (i.e., which classes share the same parental class, 'vernacular name' and 'taxonomic name' are both children of an 'identifier' class).

There are three approaches prevalent for the induction of class hierarchies from text: (1) application of lexico-syntactic patterns, (2) hierarchical clustering, and (3) analysis of co-occurrence of terms in the same sentence, paragraph or document. Lexico-syntactic patterns are mostly derived from the work done by Hearst (1992), who defined a set of lexico-syntactic patterns with which hyponymous

terms can be identified in a corpus. Hierarchical clustering is, like synonym identification, based on Harris' hypothesis that related words are often found in similar contexts, and therefore classes derived from word contexts are also found in similar contexts. An example of such an approach can be found in the work by Faure and Nedéllec (1998). They present clusters of classes that are ordered by generality in the hierarchy by their co-occurrence with a particular preposition that indicates a hyponymous relation. The third approach, analysis of co-occurrence within a sentence, paragraph or document of terms was introduced by Sanderson and Croft in 1999. In their approach, Sanderson and Croft define hierarchical relations automatically by term subsumption relations, which are derived from the occurrence patterns of terms in particular document sets. They assume that a term x subsumes a term y if the documents in which term y occurs are a subset of the documents in which term x occurs.

Relations

Relations define a further non-hierarchical structure to the classes in a given domain. In the natural history domain relations, such as 'genus is a broader class than species' or 'genus lives in a particular geographical area', can be found. The task of relation learning has been addressed by various approaches such as syntactic dependencies (Kavalec and Svátek, 2005), association rules (Maedche and Staab, 2000), and pattern identification (Pantel and Pennacchiotti, 2008). In the syntactic dependency approach used by, for example, Kavalec and Svátek (2005) and Ciaramita *et al.* (2005), syntactic patterns are identified that may indicate the presence of an ontological relation. The significance of the dependencies is tested by ranking them by frequency (Kavalec and Svátek, 2005) or by a significance test to investigate whether their occurrence is more frequent than what one would expect by chance (Ciaramita *et al.*, 2005). Association rule approaches, such as Maedche and Staab (2000)'s, start with identifying syntactic dependencies that may indicate an ontological relation, after which a generalised association rules algorithm (Agrawal and Srikant, 1995) is employed to extract conceptual relations from the selected syntactic dependencies. Pattern-based approaches, such as the ones employed by Pantel and Pennacchiotti (2008) and Ravichandran and Hovy (2002), use the web to identify useful patterns. Ravichandran and Hovy start off with a small list of seed patterns that are sent to a search engine to

retrieve more evidence for a pattern as a precise indication of a relation. Pantel and Pennacchiotti's approach is based on Hearst patterns (Hearst, 1992). Their approach takes a small number of seed instances of a relation that are sent to a search engine on a web to extract more examples of this relation. From the instances of class-relation pairs patterns are identified. These pairs are then ranked by the number of instances they cover to identify the most general patterns for a particular relation.

The approach taken in TWIBIO is an example of a syntactic dependency approach and aims to combine the induction of hierarchical and non-hierarchical relations. It is different from the syntactic dependency methods mentioned above in the sense that it utilises a specialised corpus and imposes more restrictions on the initial selection of data from which relations are to be extracted.

Relation Hierarchies

In a sense, relation hierarchies are already addressed by the ordering of relations between classes, therefore they are often not addressed separately. Some relations are subrelations of others, like some classes are subclasses of others. An example of a subrelation-superrelation is the 'preferred biotope'-'found in' relation pair. A preferred biotope for an animal species entails all climatological circumstances whereas 'found in' can be defined as only pertaining to particular geographical features. Some work on relation hierarchy discovery for verbs (verbs and verb phrases are often used to express relations) has been done by Schulte im Walde (2000). In this work, Schulte im Walde clustered verbs according to their alternation behaviour (i.e., the way verbs can be used in different subcategorisation frames in which their semantic meaning changes slightly) to find relation hierarchies.

5.2 TWIBIO

The automatic ontology construction approach presented in this chapter, called TWIBIO[1], exploits the content and structure of the reptiles and amphibians specimen database. Each database column contains up to 16,870 examples[2] of each class that is to be defined in the ontology. Similar to the approach in Chapter 3,

[1] The Wikipedia Instance-Based Induction of Ontologies system
[2] Depending on how many cells are filled

every database column is assumed to denote a class in the ontology. Through looking up the examples of each class in an encyclopaedic resource and finding phrases that may indicate a type of relation between two examples in the text of this encyclopaedic resource, relations can be defined between examples from two different database columns. TWIBIO generalises over the phrases that are found for each pair of examples from two database columns to define a relation label for a relation between two database columns, and thus two ontological classes.

In the remainder of this section, Wikipedia, the resource from which TWIBIO extracts relations is described in Subsection 5.2.1. Previous work on relation extraction from Wikipedia is described in Subsection 5.2.2. In Subsection 5.2.3, the preprocessing of the Wikipedia corpus and the query pairs for TWIBIO is detailed, followed by the description of the TWIBIO system architecture in Subsection 5.2.4. In Subsection 5.2.5, the evaluation methodology used in TWIBIO is described, followed by the results on the R&A data in Subsection 5.2.6.

5.2.1 Wikipedia

Wikipedia is the world's largest online collaboratively edited encyclopaedic resource. It has grown significantly since its beginning in 2001 (Voss, 2005). At the time of writing this thesis, there are versions of Wikipedia available for 267 languages of which 80 contain over 10,000 articles. The most popular version of Wikipedia is the English version of Wikipedia which contains 2,915,505 articles[3].

A key property of Wikipedia is that it is for the greater part unstructured text. A Wikipedia article can also contain three different types of meta-information objects: (1) categories, (2) wikilinks, and (3) infoboxes. Editors of Wikipedia articles are encouraged to supply their articles with categories. Categories can be subsumed by broader categories, creating a hierarchical structure. Editors can also link to any other article in Wikipedia through a wikilink, no matter if it is part of the same category, or any category at all. Some information in Wikipedia is presented in a more structured format, for instance through infoboxes, which have a table-like structure. The infoboxes format is not standardised for all Wikipedia articles, rendering it difficult to automatically extract information from. Figure 5.2 shows a screen shot of a Wikipedia article marked up with the three meta-information objects mentioned in this paragraph.

[3] http://en.wikipedia.org/wiki/Wikipedia:Statistics Last visited: 15 June 2009

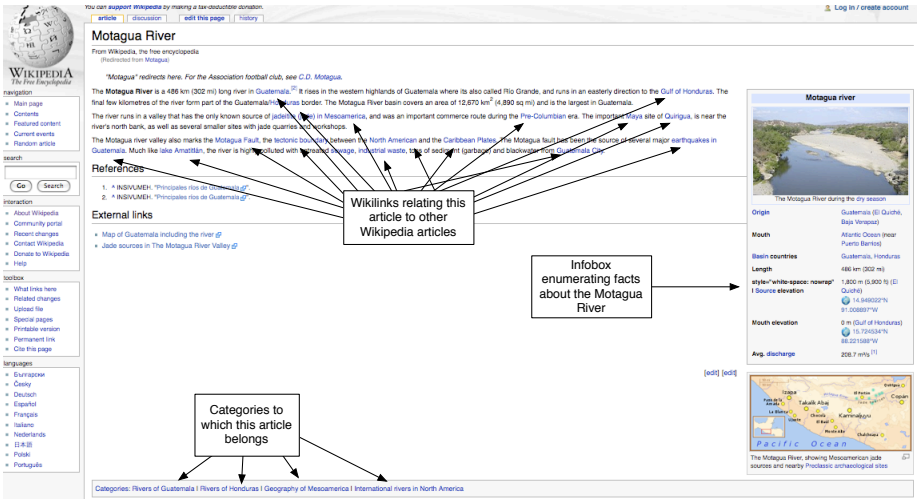

Figure 5.2: Screenshot of a Wikipedia Article showing wikilinks, an infobox and Wikipedia category links

Although there are many resources available for the natural history domain that contain general encyclopaedic information on this domain, such as Encyclopedia of Life[4], Wikipedia[5] is chosen as the external resource to extract relations from. The three main reasons for this choice are as follows.

1. Wikipedia contains information on a large variety of categories. This will be particularly useful when attempting to find relations between the geographic and taxonomic classes in the ontology. A strictly biological resource does not contain as much information on the geographic classes.

2. Wikipedia is a freely linked resource, meaning that articles on Wikipedia are not only organised via a hierarchical category structure but also through hyperlinks to other Wikipedia articles. The link structure provides far more flexibility for expressing relations between classes than more self-contained pages as often found on, for instance, Amphibiaweb[6] in which relations can only be expressed in text or through mark-up. The wikilinks will prove to be of crucial importance to TWIBIO, as will be explained in Subsection 5.2.3.

[4] http://www.eol.org Last visited: 16 July 2009
[5] http://www.wikipedia.org/ Last visited: 15 June 2009
[6] http://www.amphibiaweb.org Last visited: 15 June 2009

3. Wikipedia is more accessible than the taxonomic resources considered (e.g., Encyclopedia of Life, Amphibiaweb, Reptile Database), allowing for downloading of a copy for research purposes.

The status of Wikipedia as a dependable research resource is debated. This is partly due to its dynamic nature. However, research has provided evidence for the claim that Wikipedia is as reliable as a source that is maintained exclusively by experts (Giles, 2005). The quality of Wikipedia is maintained through the discussion between experts and non-experts worldwide. This global approach also explains Wikipedia's rapid growth and breadth, as any expert on any domain can contribute to the resource. This does mean that some categories are very well represented in Wikipedia whereas others are not.

Wikipedia is a popular resource in NLP research to extract relations from. Key examples of previous work on relation extraction from Wikipedia are presented in Subsection 5.2.2.

5.2.2 Related Work

Through the structure and breadth of Wikipedia, it is a potentially powerful resource for information extraction which has not gone unnoticed in the NLP community (Medelyan *et al.*, 2009).

Approaches that Utilise the Wikipedia Category Structure and Infoboxes

Preprocessing of Wikipedia content in order to extract non-trivial relations has been addressed in a number of studies. Syed *et al.* (2008) utilise 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 that are extracted through Wikipedia's category structure in combination with 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 extract relations.

Most studies to date have been focussing on the structured information that is explicitly available in Wikipedia, such as its infoboxes and categories (Auer and Lehmann, 2007). Although infoboxes provide rich structured information, their

templates are not yet standardised, and their use has not permeated throughout the whole of Wikipedia.

Approaches that Utilise the Unstructured Text in Wikipedia

Although the category and infobox structures in Wikipedia already provide a larger coverage of factual knowledge than resources such as WordNet, these structures do not express all semantic relations that are possibly relevant. In particular in specific domains, relations occur that are difficult to capture in infoboxes and categories. Therefore, an approach that exploits more of the linguistic content of Wikipedia is desirable.

Such approaches can be found in Nguyen *et al.* (2007) and Nakayama *et al.* (2008). In both works, sections of Wikipedia articles are parsed, entities are identified, and the verb between the entities is selected as the relation label. Through such methods, also relations that are not backed by a wikilink are extracted, resulting in common-sense factoids such as 'Brescia is a city'. For a domain-specific application this method lacks precision as high precision is more important for finding relations than recall.

The difference between previous work and TWIBIO is that the focus in this thesis is on a specific domain. The relevant Wikipedia articles for this domain are selected automatically through querying of instances from the specimen database. In this approach, not the structured content of Wikipedia is employed to extract relations from, but the textual content.

5.2.3 Data Preparation

The terms and the sentences in which they are contained are extracted from an offline copy of the English Wikipedia version of July 27, 2008. This version of Wikipedia contains about 2.5 million articles, including a vast amount of domain-specific articles that one would typically not find in general encyclopaedias. To speed up the look-up, an index was built of a subset of the link structure present in Wikipedia. The subset of links included in the index is constrained to those links occurring in sentences from each article in which the main topic of the Wikipedia article (as taken from the title name) occurs. For example, from the Wikipedia article on *Anura* the following sentence would be included in the experiments:

"The frog is an [[amphibian]][7] in the order Anura (meaning "tail-less", from Greek an-, without + oura, tail), formerly referred to as Salientia (Latin saltare, to jump)."

whereas this set would exclude the sentence:

"An exception is the [[fire-bellied toad]] (Bombina bombina): while its skin is slightly warty, it prefers a watery habitat."

as it does not include a reference to the topic of the article.

This approach limits the relations to only those between pages that are likely to be semantically strongly connected to each other. It is based on the assumption that the strongest and most reliable lexical relations are those expressed by wikilinks. Wikipedia author guidelines state that wikilinks should only be added when relevant to the topic of the article. Due to the fact that most users tend to adhere to guidelines for editing Wikipedia pages and the fact that articles are under continuous scrutiny of their readers and editors, most links in Wikipedia are indeed deemed to represent semantic relatedness between the topics of the linked pages (Blohm and Cimiano, 2007; Kamps and Koolen, 2008).

Besides removal of less relevant content, an advantage of pre-selecting content is that it reduces the amount of text to analyse, speeding up the relation extraction process.

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 relations are attempted to be found between values that are not attested in the R&A database. 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 field is filled, and some combinations occur more than once, this procedure results in 186,141 unique query pairs.

To assess whether the two terms that form a query pair are related, the semantic relatedness of those two terms is computed. Semantic relatedness can denote every possible relation between two terms, unlike semantic similarity, which typically denotes hierarchical relations such as hypernymy and meronymy and is

[7] The double brackets indicate Wikilinks

often computed using hierarchical networks like WordNet (Budanitsky and Hirst, 2006).

A simple and effective way of computing semantic relatedness between two terms t_1 and t_2 is measuring their distance in a semantic network. The wikilinks between different Wikipedia articles form the semantic network used by TWIBIO. The distance measures between terms result in a semantic distance metric, which can be inversed to yield a semantic relatedness metric. Computing the shortest path between terms t_1 and t_2 can be done using Formula 5.1 where P is the set of paths connecting t_1 to t_2 and Np is the number of nodes in path p.

$$rel_{path}(t_1, t_2) = argmax_{p \in P} \frac{1}{N_p} \tag{5.1}$$

In the TWIBIO corpus, each wikilink is one step in the path from one instance (represented by a Wikipedia article) to another instance (as represented by a Wikipedia article).

Only the shortest paths in the semantic network formed by the wikilinks are selected indicating that there is a strong relation between two terms. By indexing both incoming and outgoing wikilinks for each article a bidirectional breadth-first search can be used to find shortest paths between concepts (Wubben, 2008).

The main language of the R&A database is Dutch, still the English version of Wikipedia is chosen for this relation finding study. The consequence of this choice is that the relation labels that are discovered will be English phrases. The low coverage of the Dutch version of Wikipedia prevented using this resource, as it did not yield satisfying results for this method. The results will show that the approach does not suffer from using English instead of Dutch. Furthermore, articles in the English Wikipedia on animal taxonomy are far more elaborate than those contained in the Dutch Wikipedia, if available at all. Since the database contains Latin-based nomenclature, using the wider-coverage English Wikipedia yields a much higher recall than the Dutch Wikipedia. The values of the other columns mainly contain proper names, such as person names, geographic locations and dates, which are often the same or similar in Dutch and English. Here, the approach benefits from Dutch and English being closely related languages. In some cases differences in names that exist for different countries in each language, are matched because the database also contains entries partially or fully in English, as well as some in German and Portuguese. In other cases, TWIBIO fails to match

the database entry to the correct Wikipedia article. Here recall suffers slightly, but this is no problem as for each column there are thousands of chances to match to a Wikipedia article.

5.2.4 TWIBIO System Setup

In this subsection, the setup of the TWIBIO system is presented. As input to the system a term pair is selected. Each part of the term pair contains a value from one of two different database columns. TWIBIO processes each term pair in six steps. A schematic overview of the TWIBIO is given in Figure 5.3.

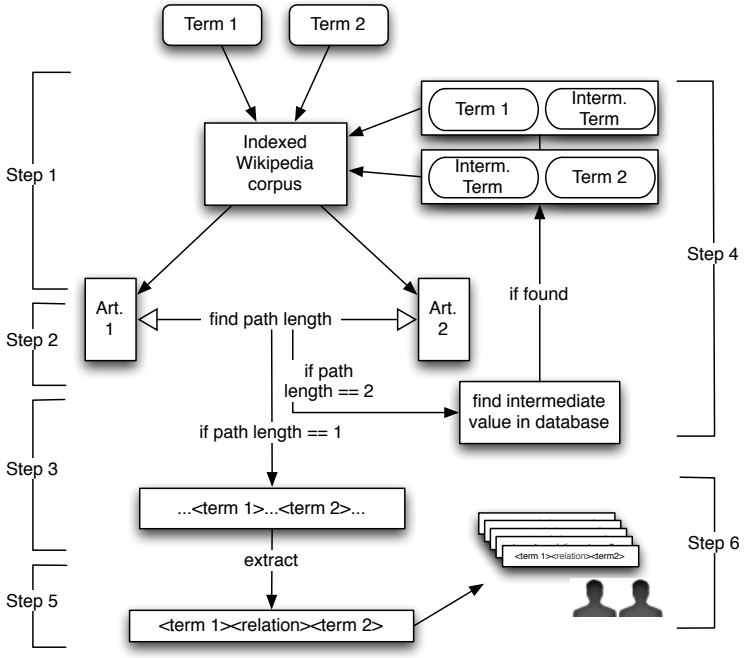

Figure 5.3: Schematic overview of TWIBIO

Step 1 For each term, the most relevant Wikipedia article is found, 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 article's main topic, it is assumed that querying only for article titles yields reliable results.

Step 2 The system finds the shortest link path between the two selected Wiki-

pedia articles. If the path distance is 1, this means that the two terms are linked directly to each other via wikilinks in their Wikipedia articles. This is for example the case for *Megophrys* from the genus column, and *Anura* from the order column. In the Wikipedia article on *Megophrys*, a link is found that leads 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 as to not overcrowd the parent article with links to its children.

Step 3 The sentence containing both target terms is selected from the articles (if available). For example, from the *Megophrys* article the sentence "*Megophrys is a genus of frogs, order [[Anura]], in the [[Megophryidae]] family.*" is selected.

Step 4 If the shortest path length between two Wikipedia articles is 2, the two terms 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 the query pair *Hylidae* from the taxonomic order column, and *Brazil* from the country column. Here, the initial path that is found is *Hylidae* $\leftrightarrow$ *Sphaenorhynchys* $\rightarrow$ *Brazil*[8]. The article-in-the-middle value (*Sphaenorhynchys*) is found to occur in the R&A database, namely in the taxonomic genus column. This link is taken as evidence for co-occurrence. Thus, the relevant sentences from the Wikipedia articles on *Hylidae* and *Sphaenorhynchys*, and from the articles on *Sphaenorhynchys* and *Brazil* are added to the possible relations between 'order' – 'genus' and 'genus' – 'country', respectively.

Step 5. The selected sentences are POS-tagged and parsed using the Memory Based Shallow Parser (Daelemans *et al.*, 1999). Like MBT (see Chapter 3), this parser is based on TiMBL. The parser provides tokenisation, POS-tagging, chunking, and grammatical relations such as subjects and direct objects between verbs and phrases.

Step 6 The five most frequently occurring phrases that are found between instances of a column pair are presented to four human judges. It is hard, if not impossible, to automatically evaluate semantic relations as the same relation can be expressed in many ways. An automatic approach would also require a gold standard of some sort, which for this domain (as well as for many other domains) is

[8]The direction of the arrows indicates from which article the wikilinks originates.

not available. The cut-off of five relation candidates per column pair was chosen to prevent the judges from having to evaluate too many relations and to only present those that occur more often, and are therefore less likely to be false hits. These can, for example, be induced by ambiguous person names that also match location names.

5.2.5 Evaluation Methodology

The four judges were presented with the five highest-ranked candidate labels per column pair, as well as a longer example of a sentence found in Wikipedia that contains 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 1 correct answer the points are added up. For example, if in a list of five, two correct answers occur on positions 2 and 4, the TRR would be calculated as $(1/2 + 1/4) = .75$. The TRR scores were normalised for the number of relation candidates that was retrieved as for some column pairs fewer than five relation candidates were retrieved. The normalisation ensured that the number of relation candidates did not affect the TRR scores, by dividing every TRR score by the number of candidates which it was made up of.

As an example, for the column pair 'Province' and 'Genus', the judges were presented with the relation candidates shown in Table 5.1. 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).

In the next section the results of the experiments are presented.

5.2.6 TWIBIO RESULTS

The TWIBIO experiment shows that between certain columns more relations are found than between others. In total 140 (32%) relation candidates were retrieved

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

Table 5.1: Relation candidates for 'Province' and 'Genus' column pair

directly. 303 (68%) relation label candidates were retrieved via an intermediate Wikipedia article (**Step 4** in the system). It is assumed that the column pairs between which a larger number of relations is discovered are more strongly related than the column pairs for which only a few relations are discovered.

From each column pair, the highest rated relation was selected with which the ontology displayed in Figure 5.4 was constructed. As the figure shows, the relations that are discovered are not only 'is a'-relations that are typically found in strictly hierarchical resources such as a zoological taxonomy or geographical resource.

For some database columns, no relations were retrieved, such as for the 'collection date' column. This is not surprising, as although Wikipedia contains articles about dates (so-called 'what happened on this day' articles), it is unlikely that it would link to a domain-specific event such as an animal specimen find. Relations between instances denoting persons and other concepts in the domain are also not discovered through this method. The reason for this is that the majority of the biologists named in the database do not have a Wikipedia page dedicated to them, indicating the boundaries of Wikipedia's domain specific content.

Occasionally, a Wikipedia article for a value from a person name column is retrieved, but in most cases this mistakenly matches with a Wikipedia article on a location, as last names in Dutch are often derived from place names. A second problem induced by incorrect data is the incorrect match of Wikipedia pages on certain values from the 'Town' and 'Province' columns. Incorrect relation

candidates are retrieved because for instance the value *China* occurs in both the 'Town' and the 'Province' columns. Cleaner data and an extra disambiguation step in which it is checked whether the Wikipedia article is about a person or a geographical location could solve these problems.

The numbers below the relation labels in Figure 5.4 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 were particularly high, meaning that in many cases all relation candidates presented to the judges were assessed as correct. The inter-annotator agreement was $\kappa = 0.63$. This is not perfect, but reasonable. Most disagreements between the judges are due to vague relation labels such as 'may refer to' as found between 'Province' and 'Country' for example. If a relation that occurred fewer than 5 times was judged incorrect by the majority of the judges the relation was not included in Figure 5.4. This was the case for the discovered labels between the 'Location' and 'Species' columns for example.

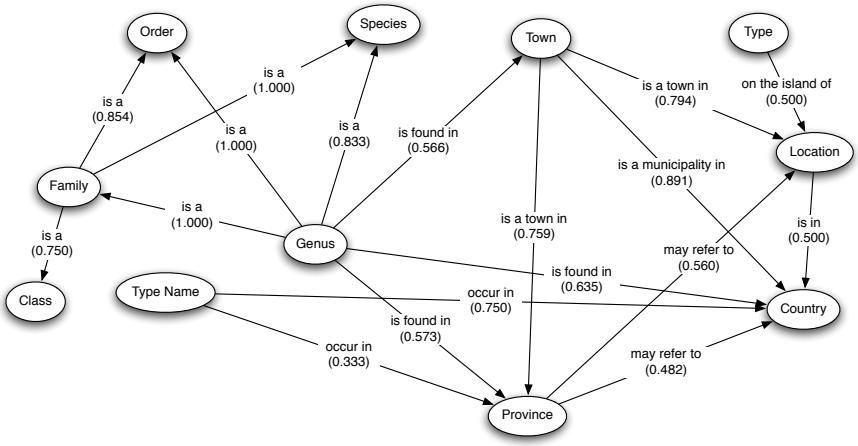

Figure 5.4: Graph of relations between columns. The arrows denote the direction of the relation, the TRR scores are shown in parentheses under the relation label

Post-processing of the results could filter out synonyms such as those found for relations between 'Town' and other classes in the domain. This would, for example, define one particular relation label for the relations 'is a town in' and 'is a municipality in' that the system discovered between 'Town' and 'Province'

and 'Town' and 'Country', respectively.

A duplication of this approach for the birds, crustaceans and shark databases from Naturalis yielded no satisfactory results as the Wikipedia articles for these subdomains were yet to be added or expanded at the time of performing these experiments.

This work has shown that it is possible to extract labelled ontological relations for domain-specific data from Wikipedia. However, the discovered relations that result in the ontology graph shown in Figure 5.4 present a view on the domain that different from the one in the ontology presented in Chapter 2. In Section 5.3, the differences between the two ontologies and the consequences of these differences are discussed.

5.3 Ontology Comparison

The ontology that is constructed via the R&A database and Wikipedia is quite different from the manual ontology that was presented in Chapter 3. This is illustrated in Figure 5.5, where the two ontologies are overlaid. In Figure 5.5, the uninterrupted lines denote the relationships defined in the manually constructed ontology. The dashed lines indicate the relationships between the different natural history classes as defined by the Wikipedia ontology. Out of the 25 relations present in total in Figure 5.5 there are only six cases in which the two ontologies define an overlapping relationship. The reason for this small amount of overlap is that the ontologies were constructed from different points of view. The CIDOC-CRM framework is constructed from an object-centred point of view with a focus on collection management and organisation, whereas Wikipedia is constructed with the aim of providing as much relevant information as possible on a given topic. This is reflected in the Twibio ontology containing more relations than the ontology constructed through the CIDOC-CRM model. Here, it is argued that this is not problematic and that each ontology provides a valid conceptualisation of the domain, but this does not mean they are interchangeable: each has its own uses. The remainder of this section is organised as follows. In Subsections 5.3.1 and 5.3.2 overviews of manual and automatically ontology comparison approaches are given respectively along with difficulties in comparing the manual and the Twibio ontologies. In Subsection 5.3.3 an argument for the coexistence of the two ontologies for one domain is given.

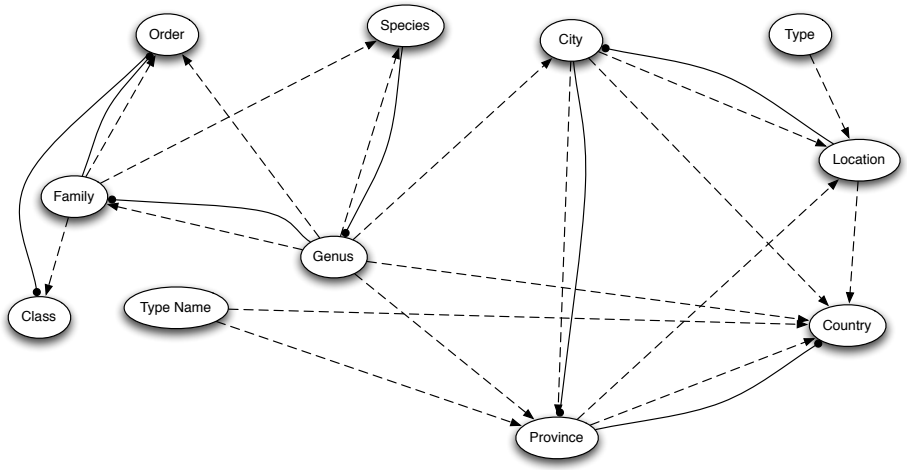

Figure 5.5: Overlay of manually constructed and TWIBIO ontologies. The relations in the manual ontology are denoted by the unbroken lines, the relations in the TWIBIO ontology by the dashed lines. The arrow heads point to the object of the relation.

5.3.1 Manual Ontology Comparison

Ontologies can be compared in different ways (Hartmann *et al.*, 2004). The most complete set of points to evaluate the similarity of ontologies on is provided by Noy and Hafner (1997), who discern eight facets for ontology comparison (see Table 5.2).

Noy and Hafner's approach can be seen as reverse engineering the ontology development process as it also takes into account decisions made in the design and development of an ontology. However, it does not provide a precise comparison of each concept and relation in two or more ontologies. Noy and Hafner acknowledge that it is not always possible to retrieve the reasoning behind design choices made. Therefore, this approach cannot be used to compare the manual and TWIBIO ontologies for the R&A domain.

5.3.2 Automatic Ontology Comparison

Ontology comparison metrics that only look at the content and structure of ontologies were developed by Maedche and Staab in 2002. In this work, Maedche

1. General	· The purpose the ontology was created for ·General or domain specific · Domain (if domain specific) · Easy integration possible into a more general ontology · Size: Number of concepts, rules, links, and so on · Formalism used · Implementation platform and language, if done · Publication, if done
2. Design process	· How was the ontology built? · Was there a formal evaluation?
3. Taxonomy	· What is the general taxonomy organization? · Are there several taxonomies, or is everything in the same one? · What is in the ontology: things, processes, relations, properties? · What is the treatment of time? · What is the top-level division? · How tangled or dense is the taxonomy?
4. Internal concept structure and relations between concepts	· Do concepts have internal structure? · Are there properties and roles? · Are there other kinds of relation between concepts? · How are part-whole relations represented?
5. Axioms	· Are there explicit axioms? · How are the axioms expressed?
6. Inference mechanism	· How is reasoning done (if any)? · What are some instances of going beyond first-order logic?
7. Applications	· Retrieval mechanism · User interface · Application in which the ontology was used
8. Contributions	· Major strengths and contributions · Weaknesses

Table 5.2: Facets for ontology comparison as defined by Noy and Hafner (1997)

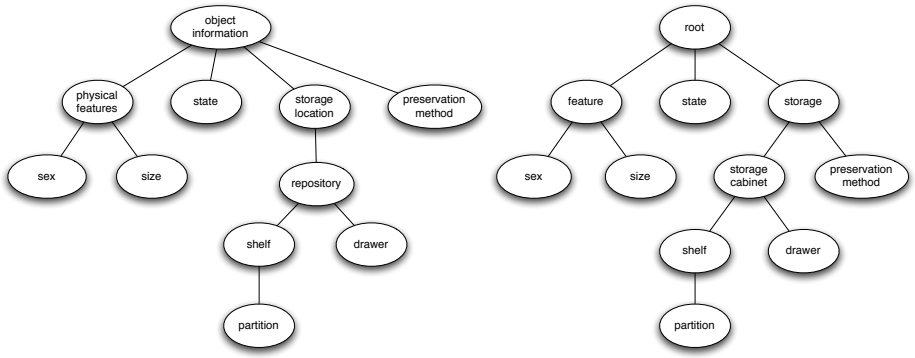

Figure 5.6: Comparison of two ontology hierarchies

and Staab present a set of similarity measures to automatically compare two or more ontologies to each other. The similarity measures address two aspects of ontologies and are presented as two different ontology-layers. The first layer is concerned with naming conventions, i.e., how classes are named in each ontology. In the comparison of the ontology induced from Wikipedia and the manually constructed ontology for natural history the terms used to identify the classes are identical, therefore this part of the evaluation can be discarded. The second layer Maedche and Staab discern is the conceptual layer which pertains to the relations, i.e., semantic structure, in the ontologies. For this second layer two measures are developed: one measure that compares the hierarchies of the ontologies and one that compares relations within the ontologies. Both conceptual layer measures are based on comparison of sub- and superclasses within an ontology. An example of the type of ontologies the measure was designed for is given in Figure 5.6. This figure is based on the example given in Cimiano (2006).

In the example ontologies in Figure 5.6 the difference in structure under the 'object info' on the ontology on the left and under 'root' in the structure on the right can be quantified. However, the TWIBIO ontology is not organised via a hierarchical structure, but it resembles a directed graph structure more; thus, it is not possible to compare it to another ontology via these measures. Even for the more hierarchical parts of the TWIBIO ontology, such as the taxonomic classes, it is not possible to use these metrics as the hierarchy is not preserved. As Figure 5.4 shows the 'Family' class is related directly to the 'Species' class, skipping the

'Genus' class which defines the level between them. This is a consequence of the choices made in the approach for constructing the TWIBIO ontology. It also uncovers a fundamental difference in what type of information each ontology expresses about the domain. As the manually constructed ontology is designed from an archive organisation perspective it defines mainly hierarchical relations between classes. An encyclopaedia is designed to provide relevant definitions and context to a topic. By extension an ontology derived from such a resource will present not only hierarchical relations, as these are often part of a definition or context description of a topic, but also any other possibly relevant relation.

5.3.3 Manual vs. TWIBIO Ontology

In this thesis, it is argued that it is conceivable to have different ontologies for a single domain that each present a different perspective on the domain. For certain tasks, a more formal and elaborate ontology is required, whereas for other tasks a simpler conceptualisation of the domain that only contains the most important classes and relations may suffice. This does not exempt the ontologies from needing to be evaluated. As the existence of different ontologies for one domain is defended by their utility in applications, their usefulness is assessed in the retrieval experiments presented in Chapter 6. In these experiments, the influence of an ontological structure on ranking of results is investigated through comparing results of the experiments with and without ranking through the two ontologies, as well as a comparison of the results of the rankings through the different ontologies. In the next section, this chapter is summarised and the answer to Research Question 2 is given.

5.4 Discussion and Conclusions

In this chapter, an automatic approach for ontology construction was presented. The novelty of the approach is that it utilises the fact that the reptiles and amphibians database provides a large number of instances for each ontological class, whereby the ontology construction method only needs to focus on identifying relations between the ontological concepts. To retrieve ontological relations the online encyclopaedia Wikipedia was used because it contains extensive and broad information on herpetological taxonomy. The resulting ontology is quite different

from the manually constructed ontology that was presented in Chapter 3.

The comparison of the manually constructed ontology and the TWIBIO ontology provides the answer to **RQ2**.

RQ2 Do automatic methods for building ontologies provide different structure for data that is not achieved by manual ontology building?

RQ2 is answered positively. As Figure 5.5 shows, the two ontologies are quite different from each other. However, as it is not possible to compare the two ontologies quantitively through the presented manual or automatic ontology comparison approaches. An external comparison, through using both ontologies in an application will be done in the next chapter.

Duplication of the experiments for other databases such as crustaceans and birds from Naturalis was barred by the unavailability of Wikipedia articles for these. However, articles are continuously added to Wikipedia, likely alleviating this bottleneck in the future. One issue that is not solved by this is the fact that for some types of information from the specimen databases, no information is, and most likely will be, available in Wikipedia, namely the dates and persons involved. To remedy this, the usage of external resources, such as publications about the specimens described in the database could be investigated.

An important feature of the approach is that it takes advantage of the implicit expression of relatedness of Wikipedia articles through wikilinks as only relations between database instances are retrieved that are connected through a wikilink and a reference to the Wikipedia article title. This limits the search space for relations but ensures a higher precision. A second feature of Wikipedia is that, as an encyclopaedia, its articles are meant to explain concepts and their context thus increasing the chances of there being a relevant relation described. Experiments that attempted to retrieve ontological relations through a search engine on the web, showed that the data found there was too varied in language and mark-up to be used for this approach to automatic ontology construction.

As the individual relations of the instance-driven approach are rated by human judges, only relations they deemed relevant are included. In this way, not the entire ontology is evaluated instantaneously, but the individual relations are evaluated. A comparison to a gold standard ontology, such as the manually constructed ontology presented in Chapter 3, is difficult as ontology comparison techniques are designed for hiearchical structures which are not available as the

Wikipedia ontology. This is due to the fact that the TWIBIO ontology it is derived from a network-based resource which more resembles a directed graph than a hierarchy. However, it need not be a problem that there are two different ontological viewpoints on one domain, as both express different features of the domain.

In the next chapter, the ontologies are both applied to rank results from an information retrieval experiment.

6

Data Retrieval

The good of a book lies in its being read. A book is made up of signs that speak of other signs, which in their turn speak of things. Without an eye to read them, a book contains signs that produce no concepts; therefore it is dumb.

Umberto Eco, The Name of the Rose, 1980

Access to information hinges on two conditions: (1) the availability of information and (2) the ease with which the information can be retrieved. Due to the backlog of digitisation in many domains, the first condition is often not fulfilled. As institutions have limited manpower and an enormous amount of data, the fact that data often only exists on paper is still a major bottleneck. For the herpetological collections at Naturalis, this bottleneck was alleviated in Section 3.1 by automatically segmenting and labelling semi-structured data from the reptile and amphibians field logbooks and registers for inclusion in the database. By automatically growing the database, a significant amount of precious time from experts is saved and the recall of information retrieval is boosted considerably. Access to this digital resource is provided by the Mitch Information Retrieval Appliance (Mira[1]) which aids users in accessing information by interpreting and expanding queries automatically and ranking results by relevance. To test the query expansion and ranking modules developed for Mira, a series of experiments was

[1]Meaning 'Look!' in Spanish

carried out in which the performance of the different modules was evaluated on one hundred real world domain specific queries. These experiments were not only run on the herpetological data, but also on a birds database, with another set of one hundred queries, to test whether the approaches would also apply to a similar but different data set. The results and analyses of the MIRA experiments provide the answer to Research Question 3.

RQ3: Can access to information be aided through a retrieval system enriched with domain knowledge?

This chapter is set up as follows. In Section 6.1, an introduction to the field of Information Retrieval is given. In Section 6.2, the birds databases are described. In Section 6.3, the queries used for testing MIRA are discussed, followed by the different search modules and metrics implemented to improve and evaluate the retrieval experiments in Section 6.4. In Section 6.5, the evaluation metrics used in this study are described. The chapter is concluded by a presentation of the results in Section 6.6 and a chapter summary and discussion in Section 6.7.
The chapter is based on the following publication:

- Van Erp, M. and Hunt, S. (2010). Knowledge-driven information retrieval for natural history. In *Proceedings of the 10th Dutch-Belgian Information Retrieval Workshop (DIR 2010)*, pages 31–38, Nijmegen, The Netherlands

6.1 Information Retrieval

For as long as people have created large collections of data, they have been searching for means to access information in these collections (Paijmans, 1999). As many digital data collections grow more rapidly than analogue collections, many researchers have investigated means to harness the problem of finding relevant pieces of information in large collections (where relevant denotes how well a retrieved set of documents, or a single document, meets the information need of the user). The field that has emerged from this is called Information Retrieval (IR), a term coined by Mooers in 1948 (Mooers, 1948; Gupta and Jain, 1997). IR is concerned with identifying documents from a collection that are relevant to a particular query. The work presented in this thesis does not deal with search on full-text documents but structured data from a database, therefore, in the strict

sense of IR this would need to be described as data retrieval (DR) (van Rijsbergen, 1979). However, in the work done for this thesis, the database, field logbooks and registers will be linked to several external resources, and approaches from the IR domain are investigated to improve the retrieval of relevant database records to a query, putting this work somewhere between IR and DR and closely related to XML retrieval or XML information retrieval.

The database systems used at Naturalis do not aid the user in retrieving records that contain synonymous terms to the query term entered, or rank database records that are more relevant to the query before records that are less relevant. The current database search functionality also requires that the user knows the database structure in order to fully access it. To make it easier for users to retrieve the relevant entries from the databases for a particular information request, three different types of improvements over simple database search are investigated: query interpretation, query expansion, and result ranking. Query interpretation facilitates a more precise formulation of the query to improve the retrieval performance, this interpretation can be done automatically or manually, as Section 6.3 will show. Query expansion facilitates retrieval of database records that do not only contain the terms present in the original query but also synonymous terms. Ranking aims to present more relevant results first followed by the less relevant results.

6.2 The Naturalis Birds Databases

The birds database that is used in this work is assembled from three databases on birds that were created by researchers at Naturalis. The first two databases were compiled within the *Building the databases of life*[2] project at Naturalis which was aimed at storing information on the collection in databases. The first database holds 117,649 records pertaining to the passerines (songbirds) that are preserved as study skins in the collection. The second database was created between 2002 and 2004 and contains 68,341 records describing the non-passerines (not-songbirds). The third database was created later by researchers and contains 34,028 records that describe specimens in the collection that were not included in the previous two databases (such as the mounted passerine specimens). The combined database (henceforth referred to as *Birds database*) describes the cir-

[2]NWO Groot project number: 175.010.2003.010

cumstances under which the bird specimens were found and how they are stored in the collection. Twenty columns are used by all three databases. Then, each database has some additional columns that contain information specific to that part of the bird collection. The birds databases are cleaner than the reptiles and amphibians, presumably, as they were created over a shorter time span and fewer people were involved in the data entry process. The main language is Dutch, but in addition to the taxonomic names in Latin, it also contains English, Indonesian, German, and some French and Portuguese.

In Table 6.1, some general statistics about the data in the Birds database are summarised. A data sample is presented in Table 6.2. In Appendix B, a more detailed account of the birds database is given.

# Columns	32
# Records	220,018
Collection dates	1779-2006
Collection Coverage	100%
Geographical Coverage	Europe, Africa, Asia, North-America, South-America, Oceania

Table 6.1: Statistics on birds database

Column	Value
Registration #	103146
Family	Accipitridae
Genus	*Buteo*
Species	*buteo*
Author	(Linnaeus)
Country	Netherlands
Locality	Limburg
Remarks	Found dead in nest after storm

Table 6.2: Data sample from birds database

6.3 Queries

External researchers regularly request access to Naturalis' extensive specimen collection or to the meta-data that is found in the databases describing the col-

lections. As the databases are not publicly available, requests regarding the collection and collection information are usually received by collection managers at Naturalis. For the work done in this thesis, collection managers provided questions from researchers regarding the bird and herpetology collections. These questions give a good idea of what kind of information researchers are initially looking for and are at the basis of the evaluation of Mira.

Both the birds and herpetology questions were extracted from longer (often email) messages. The questions have been summarised manually into only the information request and not the introduction for why the information is requested. A full information request may look as follows.

> *I recently learned that there is a specimen of Cuculus poliocephalus at Leiden Museum found dead at Anse aux Pins, Mahe in October 1979 (Dr. Dent). Mr. Prefect kindly provided me with your contact details to follow up. Would it be possible to inspect the specimen, confirm the identity and provide any other details if possible (age of bird, exact date found)?*

Which is summarised manually as:

> *Cuculus poliocephalus Anse aux Pins Mahe 1979*

for use in Mira.

6.3.1 Reptile and Amphibians

The 100 reptile and amphibians queries were gathered from requests about the herpetology collection at Naturalis sent between September 2003 and December 2008.

Some example queries are given below.

- What type specimens of New Guinee skink do you have in your collection?
- Do you have male specimens of Hypsilurus godeffroyi?
- Are there Dipsas species other than D. catesbyi and D. variegata from the Guianas and Eastern Venezuela in the collection?
- How many species of Rana palmipes as defined by Spix in 1824 are in the collection?

The types of queries can roughly be divided into (1) queries that only contain a genus and a species name (37), (2) queries that contain a genus and a species name, plus a geographical name (18), (3) queries that contain a genus and a species name, plus require one or more other types of information (33), and (4) queries that do not contain a genus and a species name (12).

Of the 33 the queries that enquire after specimens of a particular genus and species with one or more restrictions, three queries enquire after a genus and species plus a particular size, two queries after a specimen of a particular genus and species and sex, one specimen of a particular genus, species, sex found in a particular location. Two queries only mention the literature reference in addition to the genus and species, five queries mention, besides genus, species, and literature, a geographical location and four queries are even more specific asking for a particular genus, species, literature reference, geographical location and registration number. Six queries pertain to a genus and species name and a registration number, one query asks, in addition to genus, species and registration number for a specific location, and another query asks for a type specimen of a particular genus, species and registration number. There are four queries that enquire after a type specimen of a particular genus and species, and in one additional case the literature reference is also given. Finally, one query enquires after one or more specimens of a particular genus and species that was found in a particular geographical location during a particular year.

Of the 12 queries that do not contain a genus and a species name, nine inquire after a genus and either no restrictions (1), a restriction on the location where it was collected (6) or on the type (4). Two queries only contain a registration number and one query is made up of a registration number and a geographical restriction.

The cases in which a query contains a registration number (17 in total), would facilitate the most precise querying, if it were not for the fact that registration numbers are not unique.

For each of these queries, the relevant database records were identified manually. For 16 queries no specimens were present in the database. For the remaining 84 questions the number of returned records varies greatly. The number of relevant records per query is given in Table 6.3 in which the numbers of results relevant to a query are binned in bins that increase in size on base-10 logarithmic scale.

# Relevant Records	# Queries
1	21
2–10	37
11–100	20
101–1000	6
0	16

Table 6.3: Number of relevant records for reptiles and amphibians queries

6.3.2 Birds

The 100 queries for the birds experiments were gathered from requests about the bird collection that were received between 1992 and 2006.

Some example queries are given below.

- Are there any specimens Nipponia nippon in the Naturalis collection?
- Are there any striped crakes (Aenigmatolimnas marginalis) from Africa collected by Andersson in 1867 in the collection?
- Is there a juvenile female specimen of Hypotaenidia celebensis (now Gallirallus torquatus celebensis) in the collection?
- Are there any skins of Leclanchers Bunting (P. leclancherii) in the collection?

The birds queries can be divided into queries that (1) contain only a genus and a species name (49), (2) contain a genus and a species name plus a type of material indication (25), (3) contain a genus and a species name plus one or more other restrictions (18), and (4) do not contain a genus and a species name (8).

Contrary to the reptiles and amphibians questions, there are no mentions of registration numbers; this is due to the fact that in ornithological publications it is not customary to mention the registration number, whereas in herpetological publications it is. The requests for information on the birds collection do specify what type of material is needed for the research, as 30 the queries mention skin or skeleton. The reason that this is not relevant for most herpetological research is that those specimens are most often kept in alcohol, whereas birds are mostly kept partly or completely dried and mounted.

Of the 18 queries that contain a genus and a species name plus one or more restrictions other than only type of material, four queries pertain to a particular

genus, species and give a literature reference, four queries enquires after specimens of a particular genus and species that were found at a particular location, three queries enquire after a type specimen of a particular genus and species, and three queries after a particular sex. Furthermore, three queries enquire after specimens of a particular genus, species, sex, and age and one query is directed at specimens of a particular genus, species, size, and material, and found at a particular location.

The cases in which the queries that do not contain a genus and a species name (8), three enquire after a genus and a particular material, one inquires after all specimens from a particular location, three only a genus, and one of all specimens of a particular type of material from a particular location.

As with the Reptiles and Amphibians, the gold standard against which MIRA was tested, was compiled by manually identifying the relevant records to each query in the database. Again, for 16 questions no relevant specimens were available in the databases. The number of relevant records present in the database for the other 84 queries are given in Table 6.4.

# Relevant Records	# Queries
1	9
2–10	25
11–100	42
101–1000	8
0	16

Table 6.4: Number of relevant records for birds queries in bins increasing in size

6.4 MIRA Modules

In this section, the eight modules that provide more enhanced access to the specimen database are described. MIRA contains eight modules, split into three types of modules (query interpretation, query expansion, and result ranking) that aim to improve performance for retrieval of natural history data through using domain knowledge. MIRA is built on eXist[3], an open source XML database management system. It features efficient, index-based XQuery processing[4].

[3] http://exist.sourceforge.net/ Last visited: 3 July 2009
[4] http://www.w3.org/TR/xquery/ Last visited 29 December 2009

6.4.1 Query Interpretation

Most of the queries in the test sets require more precise formulation than simply and/or queries. Consider for example the query *Are there any specimens of species Dendrophis pictus (=Dendrelaphis inornatus) in the collection?*. Here, the user is looking for database records that contain either the terms "Dendrophis" and "pictus" or records that contain the terms "Dendrelaphis" and "inornatus".

To be able to handle such queries a query language that can encode that for part of the query **any** query term should match (which can also be expressed by the boolean OR operator) and for part of the query **all** query terms should match (which can also be expressed by the boolean AND operator). The query terms that are extracted from the example query are: *dendrophis*, *pictus*, *dendrelaphis*, and *inornatus*. To express that specimens of genus *Dendrophis* and species *pictus* or of genus *Dendrelaphis* and species *inornatus* are to be retrieved, the query is rewritten to *any(all(dendrophis,pictus),all(dendrelaphis,inornatus))*.

Users can be taught this query format, but due to the availability of taxonomic resources MIRA can also automatically translate basic query term enumerations such as *dendrophis*, *pictus*, *dendrelaphis*, and *inornatus* into the interpreted genus and species pairs for the reptiles and amphibians. A substantial body of work on query interpretation and reformulation exists. Tata and Lohman (2008), for example, developed a system that translates keyword queries into SQL queries to query databases. Their approach utilises a parser to match query terms to database schema elements and a set of rules to generate and rank possible query trees. A similar approach is taken in Zhou *et al.* (2007) in order to construct SPARQL queries from keyword queries. Background knowledge from ontologies are used to translate keyword queries to description logic conjunctive queries in Tran *et al.* (2007). Tran *et al.* match query terms to knowledge base entities, after which connections between the selected knowledge base entities are identified and listed in a graph. This graph is translated to the final query.

The MIRA query interpretation module is rule-based and uses domain knowledge from external resources. First, the MIRA query interpretation module looks up each query term in the taxonomic and geographic resources to classify it as either a genus, species, or geographic name. The module can also recognise registration numbers as terms that contain numbers and possibly letters. After each term is classified, the module builds up the query according to the following

procedure that restricts possible combinations of types of terms.

Initially, every query is translated to an AND query as *all(...,...)*, denoting that all terms between the brackets should occur in a document for it to be retrieved, as one wants as many terms as possible to match. Then, the system tries to classify each term as either indicating a genus, species or geographic or registration number value. If two terms of the same type are identified within the same query this could indicate a synonym and this part of the query is then relaxed to facilitate for records to be retrieved that include either one of the synonymous terms. The synonymous terms are thus embedded in an OR query-segment by *any(...,...)*. Cases in which two genus/species pairs are encountered are translated to *any(all* $(genus_1, species_1)$*,all(*$genus_2, species_2$*))* and are expanded if more than two pairs are found. The automatic translation module is checked against a gold-standard of manual rewriting of each query. For the reptiles and amphibians, it translates 77% of the questions correctly. The cause for translation module failure is, in all cases, due to a term not matching in the external taxonomy.

As was mentioned in Subsection 4.5.3, the accepted taxonomic resources are not complete, due to the continuous debates in the taxonomy domain, which severely limits automatic query interpretation. Also, for the birds domain, no accepted taxonomy was freely available for querying. Therefore, in the experiments presented in Section 6.6, the manually rewritten queries are used.

6.4.2 Query expansion

Query expansion is a popular topic in information retrieval. It is often used to add synonymous terms to a query to make sure pages or documents using a different but synonymous term are also retrieved (e.g., a query containing the word "forest" could be expanded with the term "wood"). Query expansion to aid retrieval was first introduced in Spärck-Jones (1971). In this work, a clustering technique was used to discover related terms based on co-occurrence in documents to expand queries with semantically similar terms. In Guarino *et al.* (1999), query expansion is performed through searching for occurrences of synonyms for a query term found in WordNet. Milne *et al.* (2007) use Wikipedia instead of WordNet to identify synonymous terms to expand queries with. In addition to this, results are clustered by topics which are browsable by users, thus providing an extra means to filter out possibly negative results.

Domain-specific resources to identify terms to expand queries with are used by Bodner and Song (1996), and Büttcher *et al.* (2004). Bodner and Song use domain specific knowledge bases (as well as WordNet) to search for synonymous, as well as hypernymous and hyponymous, terms to expand queries with. In (Büttcher *et al.*, 2004), queries are automatically expanded with terms taken from domain specific databases. The authors achieve mixed results, some of the databases they used proving more useful than others. The reader is referred to (Andreou, 2005) for an overview of previous work on query expansion.

In Mira, three query expansion modules are implemented that are aimed at increasing the recall by providing additional synonymous keywords or to remedy the influence of language variation on the retrieval of relevant results. The first expansion module expands to taxonomic synonyms, the second expansion module expands to geographic terms, and the third module attempts to expand both taxonomic and geographic terms.

Taxonomic term expansion

In Chapter 2, certain peculiarities of the natural history domain were mentioned, such as the continuous changes to the taxonomy. When inspecting an official taxonomic resource such as Amphibian Species of the World for amphibians and The Reptile Database for reptiles, one will often find synonyms for particular species names. If one would for example want to retrieve all snakes present in the collection, one could query for all records describing a specimen of suborder 'Serpentes'. However, the suborder 'Serpentes' is also known as 'Ophidiae'. Another problem with this query is that the reptiles and amphibians database does not contain a suborder column (although sometimes the suborder value is entered in the order field), hence in order to retrieve all snakes in the collection one would have to query the database for all 18 snake families, which each may be known by synonyms as well. To relieve users from needing to formulate a query that contains each of the 18 snake families along with their possible synonyms, a query expansion approach will be employed. To perform automatic query expansion, Mira will use the taxonomic resources described in Chapter 2. For each query not only the name that occurs in the query was searched for in the database, but also occurrences of its synonymous terms (Latin and vernacular).

Geographic term expansion

Similar to the taxonomic resource, but slightly different in operation is the enrichment of the knowledge base with a geographic resource. In order to be able to process the geographical component of a query such as "What specimens of Rana catesbeiana collected in Holland are present in the collection?" it is necessary to be able to search the database for synonyms of the term 'Holland', as the database contains information in different languages. Also, by 'Holland', most likely 'the Netherlands' is meant. Three other flavours of geographical expansion modules were investigated in the MIRA experiments; (1) hypernym expansion, (2) hyponym expansion, and (3) both hypernym and hyponym expansion. Expansion to hypernyms or hyponyms follow the idea of Voorhees (1994). In a hypernym module, if the query contained the term 'Nebraska', the query was expanded to 'United States of America' to remedy the negative influence of missing values in the 'province/state' column. Although hypernym and hyponym expansion are popular approaches that work for other systems (see Navigli and Velardi (2003) for an overview) it did not aid object retrieval for the herpetological and birds collections in these experiments. This is probably due to the fact that the retrieval process suffers most from the language variation in the database and not so much from missing values. Therefore the geographical expansion was only limited to expanding to synonymous terms found in GeoNames.

Expand via both taxonomic and geographic resources

In MIRA, besides investigating the influence of taxonomic and geographic query expansion separately, also the influence of expanding queries both taxonomically and geographically is investigated.

6.4.3 Result Ranking

As strict querying in which all query terms ought to match (the ALL query mode) often leads to low recall, two less strict search options are employed to boost the recall, namely the ANY and COMPLEX query modes. As the results in Section 6.6 will show, the broadest, ANY, search works best to achieve the highest recall. However, as by default eXist retrieves results in the order in which they were indexed, this may mean that the more relevant results may not be presented first. Therefore, it pays off to rank the results according to relevance so users

will be presented with the more relevant results first. In MIRA, four ranking methods are investigated. The first two measures are the simplest and most straightforward. The third and fourth ranking methods utilise the ontologies presented in Chapters 3 and 5.

Order by number of matches

This method of ranking is the most straightforward and is not relevant for querying in the ALL query mode. However, as by default eXist does not rank results and returns records in the order in which they were added to the database, records that match with only one query term may be presented before records that match several query terms in the ANY query mode. For these cases, a ranking that presents the records that match with the highest number of query terms first is investigated.

Order by knowledge of column importance

Analysis of the queries has shown that queries usually do not pertain to information in some of the longer database cells such as special remarks. Hence, when giving each column equal importance, a query such as *Bufo marinus* will return results such as:

> *RMNH 34003* ***Bufo marinus*** *Lely Range, airstrip, distr. Marowijne, Surinam, 11-05-1975, 15.50h, on airstrip, near tall forest, 650m, l + d. X.X. Xxxxxxxx. RMNH 34003*

as well as:

> *RMNH 20761 TANK NO Slide 1980-10- 37 (fell) Paleosuchus trigonatus 1 ex. km 110, 19-09-1980, 20.45 h, in swamp, flooded part of forest with many dead trees and low bushes, near jeep trail through tall forest, 100 m. length 1.445 m, skin and carcass to create skeleton. Stomach contents kept separately: crab +* ***Bufo marinus*** *+ grit. Observed this specimen already on 16-09-1980 (see p.89).*

After analysis of the queries, it was clear that most queries pertain to the request for information from the genus and species columns and never from the special remarks column in which one might, for example, find information on a specimen's

stomach contents. Matches found in the 'genus' and 'species' columns, as well as in the 'registration number' column are thus ranked higher than matches found in other columns.

RecordRank

RecordRank is a simplified version of the original PageRank algorithm to rank results by relevancy (Brin and Page, 1998). The main assumption behind PageRank is that some webpages are more authoritative than others and those should rank higher than pages that are deemed less authoritative. The idea to rank the MIRA retrieval results by some measure of authority is given by the hypothesis that researchers might pose questions about the specimens or species Naturalis is known for more often. For example, Naturalis has a big collection of reptiles and amphibian specimens from the Amazon. Therefore, Naturalis may be more of an authority on that field than on reptiles and amphibians from other regions and researchers might ask more often about specimens from the Amazon, thus perhaps these should be presented first.

Authority in PageRank is measured by the number of incoming links to a page. Furthermore, PageRank does not consider links from all pages equally; links from pages with a higher PageRank are considered more important than links from pages with a lower PageRank. The PageRank of a set of webpages is calculated using an iterative calculation, and the PageRank scores for the total set of webpages forms a probability distribution expressing how likely it is that a random user will visit a page. PageRank is summarised below:

> *Consider a set of pages W consisting of pages $w_1 \ldots w_n$. For every link from page w_i to w_j the rank of page w_j increases. The amount with which the rank of w_j increases is dependent on the rank of w_i.*

The PageRank algorithm has sparked interest in applications other than search engines as ranking results for entity relation graphs (Chakrabarti, 2007) and Word Sense Disambiguation (Agirre and Soroa, 2009). Similar to the aim in MIRA, the PageRank algorithm has also been translated to a relational database setting by Balmin *et al.* (2004). In Balmin *et al.*'s work, databases are translated to modelled graphs in which objects are nodes and their semantic connections the edges. Although the databases used for MIRA were originally flat, the ontologies

Record	Genus	Species	Country
1	*Bufo*	*Marinus*	*Suriname*
2	*Hyla*	*Triangulum*	*Equador*
3	*Hyla*	*Minuscula*	*Venezuela*
4	*Bufo*	*Marinus*	*Venezuela*

Object		Score	Object		Score
Bufo	:	4	*Minuscula*	:	1
Hyla	:	4	*Suriname*	:	1
Marinus	:	2	*Equador*	:	1
Triangulum	:	1	*Venezuela*	:	2

presented in Chapters 3 and 5 provide the databases with the necessary structure to consider them as a relational data resource.

Once the databases are relational, it would seem straightforward to follow (Balmin *et al.*, 2004) and convert the database to a graph and apply their ObjectRank algorithm to it. Due to the fact that all relations in the ontology are bi-directional the approach is even simpler because the PageRank or ObjectRank scores can be approximated by taking a shortcut of ranking the objects by degree. The notion of degree comes from Graph Theory and denotes the number of edges (relations) linked to a node (object) (Diestel, 2005). In order to compute the rank-score of an object, the number of relations is counted and objects with a high number of relations receive a higher score and thus higher rank.

All unique database values are objects in the Mira RecordRank module. In order to go from a ranking of objects in the domain to a ranking of records in the database the scores of all objects that occur in a database record are added up and normalised over the number of objects present in the database record (as database cells can be empty).

Consider for example the following 4 fragments of database records.

In the Twibio ontology, there are relations between the 'genus' and 'species' and 'genus' and 'country' classes, thus in these records, there are also relations between *Bufo* and *Marinus* (twice, in Record 1 and Record 4), *Bufo* and *Suriname*, *Hyla* and *Triangulum* and *Hyla* and *Equador*, *Hyla* and *Minuscula*, *Hyla* and *Venezuela*, and *Bufo* and *Venezuela*. Then, the scores of the objects are computed by counting the number of times they occur in the data in some relation to another object. This results in the following scores.

In order to rank the records, the scores of the objects occurring in each record

Record	Genus				Species				Country				Score
1	*Bufo*	:	4	+	*Marinus*	:	2	+	*Suriname*	:	1	=	7
2	*Hyla*	:	4	+	*Triangulum*	:	1	+	*Equador*	:	1	=	6
3	*Hyla*	:	4	+	*Minuscula*	:	1	+	*Venezuela*	:	3	=	7
4	*Bufo*	:	4	+	*Marinus*	:	2	+	*Venezuela*	:	3	=	8

are added up.
This results in Record 4 achieving the highest score and thus rank 1, followed by Record 1 and Record 3, and finally Record 2.

Two flavours of the RecordRank result ranking module are tested in MIRA: (1) one that is driven by the manually constructed ontology as presented in Chapter 3 and (2) one that is driven by the TWIBIO ontology as presented in Chapter 5.

Query-sensitive RecordRank

As PageRank was designed for very large-scale web search it is biased to the most important webpages, or in this case, database records. The advantage for the natural history domain is that it reflects the distribution of topics the domain is concerned with, ranking database entries that pertain to the core business first. A drawback is that for broader queries in a smaller domain the same set of database entries is always ranked on top. It may therefore be more useful to present a ranking of importance relative to a query. This idea was explored in Haveliwala (2002), in which a topic-sensitive PageRank approach is presented. The idea of only computing the rank over the retrieved result is also used in the HITS algorithm, another link analysis algorithm that is used to rank web pages according to authority (Kleinberg, 1999), which was developed around the same time as PageRank. In Haveliwala's approach, a set of topic-specific PageRank vectors is computed only from pages relevant to the query, which are then used to retrieve results for a query on a particular subject. Since the natural history databases provide a smaller domain which cannot be easily broken up in more subdomains, the MIRA query-sensitive RecordRank module does not use precomputed vectors. Instead, for each query the RecordRank scores are computed at run-time, but only for the retrieved results.

As with the general RecordRank module in MIRA, there are two flavours of the query-sensitive RecordRank module: (1) driven by the manually constructed ontology (presented in Chapter 3) and (2) one driven by the TWIBIO ontology (presented in Chapter 5).

6.5 Evaluation Metrics

There is a vast amount of evaluation metrics available for IR, that each address a different dimension of the results. The basic precision and recall measures (van Rijsbergen, 1979) do not evaluate ranking of results, only how many of the retrieved records were indeed relevant and how many of the total number of relevant records were retrieved. In order to measure whether relevant results are ranked on top, mean average precision and the rank of the first relevant result for each series of queries is also reported.

Mean average precision (MAP) is a standard metric in IR evaluation which besides precision, also takes ranking and relevance into account (Manning *et al.*, 2008). It is computed over all queries where the average precision of the results returned for a query are computed after truncating the list of returned results after each of the relevant results in turn. The formula is shown in Equation 6.1 in which r is the rank, N the number of retrieved results, $rel(r)$ a binary function on the relevance of a given rank and $P(r)$ the precision at a given rank.

$$AvgPrec = \frac{\Sigma_{r=1}^{N}(P(r) \times rel(r))}{number\ of\ relevant\ entries} \tag{6.1}$$

A disadvantage of MAP is that it provides a single score for one run of experiments (100 queries in this work), weighing queries for which one record is to be retrieved equally to queries for which more records are to be retrieved. Therefore for each of the series of Mira experiments, a finer grained analysis of the results is also presented. All measures were computed using the evaluation script used in the Text REtrieval Conferences (TREC)[5].

6.6 Experiments and Results

Separate series of experiments were run with the reptiles and amphibians (Subsection 6.3.1) and birds queries (Subsection 6.3.2) for all Mira modules as well as for combinations of modules.

First a keyword search of terms from all queries is performed on the initial database of 16,870 entries. Precision for a query in which every keyword ought to match (All query mode) is 28.06% and where any keyword ought

[5] http://trec.nist.gov/ Last visited: 3 July 2009

to match (ANY query mode) is 22.27%. Recall is 18.33% and 55.45% respectively. Both results for the ALL and ANY query modes are reported as baseline as some queries require that all terms match (e.g., *Is there a **syntype** of **Megapodius rubripes** present in the collection?*), whereas other require that some terms match (e.g., *Are there any specimens of genus Centrolenella (=Cochranella) present in the collection?*). For the birds, a keyword search on the databases yields a precision and recall of 13.18% and 77.01% respectively for an ANY search and a precision and recall of 46.33% and 46.83% for a restrictive ALL search. As can be seen from the results in Table 6.5 the precision and mean average precision are higher for both data sets when the COMPLEX query mode (e.g., *all(all(Hyla,minuta),any(Guyana,Brazil,Trinidad, Columbia,Ecuador,Peru)))*) is used, but recall still lags behind.

When the non-overlapping field logbooks and registers are added for the reptiles and amphibians, recall is boosted to 31.67% for all and 84.37% for any. Precision stays low at 33.07% and 21.62%, due to the imprecise nature of simple keyword search. Although the field logbooks and registers make up for 2/3 of the data on the reptiles and amphibians it does not help as much proportionally in these experiments. The reason for this is probably that the collection managers entered the information on the most important specimens into the database first, for example those specimens that are cited or enquired after most often. The results of the unexpanded and unranked ALL and ANY experiments on the full reptiles and amphibians data set (i.e., manually constructed database and automatically segmented and labelled field logbooks and registers) are taken as the baseline results over which the MIRA modules should provide improvements.

The overall baselines are summarised in Table 6.5.

In the remainder of this section, the results for the reptiles and amphibians will be presented in Subsection 6.6.1, followed by the results for the birds in Subsection 6.6.2. In all tables, UnExp denotes the unexpanded query mode (also presented in Table 6.5), ALL denotes the most restricted AND search mode, ANY denotes the most unrestricted OR search mode and CX denotes the interpreted complex query mode. Furthermore, the MIRA expansion modules are abbreviated to 'TaxExp' for taxonomic expansion, 'GeoExp' for geographical expansion, and the combination of the two as 'TaxGeoExp'. The ranking modules are denoted by 'NumMatch' for ranking via number of matches, 'GenSpec' for ranking according to genus and species columns first, 'GlobRRMan' and 'GlobRRWiki' denote

	All	(DB)	(FB)	Any	(DB)	(FB)	Cx	(DB)	(FB)
R&A									
Precision	33.07	(28.06)	(16.53)	21.62	(22.71)	(9.38)	40.13	(29.22)	(21.47)
Recall	31.67	(18.33)	(12.51)	84.37	(55.45)	(36.65)	37.59	(18.59)	(13.99)
MAP	30.04	(18.10)	(11.39)	28.87	(21.47)	(8.49)	35.87	(18.35)	(12.78)
Birds									
Precision	46.33			13.18			46.97		
Recall	46.83			77.01			46.70		
MAP	46.17			14.15			46.61		

Table 6.5: Baseline scores for unranked All, Any and Complex (Cx) queries on reptiles and amphibians and birds data. Results for reptiles and amphibians are split out for database (DB) and field logs and registers (FB)

ranking via global RecordRank through the manual and wikipedia ontologies, respectively, and 'QSRRMan' and 'QSRRWiki' denote ranking via query-sensitive RecordRank through the manual and wikipedia ontologies, respectively. All significance scores are computed at the p=0.05 level using a paired t-test (Box *et al.*, 1978) in R version 2.8.0.

6.6.1 Results for Reptiles and Amphibians

In this subsection, the results of the experiments on the reptiles and amphibians queries are presented, starting with the results of the query expansion modules.

Query Expansion

In Table 6.6, the overall precision, recall and mean average precision scores of the expansion modules are presented. The expansion modules aid retrieval significantly; in particular the geographical expansion module accounts for a doubling of the recall in comparison to the baseline in the All and Complex query modes (from 31.67 to 83.30 and from 37.59 to 85.85 respectively). However, precision goes down, as also more irrelevant records are retrieved.

The bold scores denote that the result is statistically significant with respect to the results for the UnExpanded experiments. For the All query mode this means that all expansion modules provide significant improvements. For the Any query mode this is not the case because the recall for these in the baseline system is already quite high and the expansion modules broaden the search even more resulting in more results that are less precise. Although the precision decreases

overall, for the COMPLEX query mode the mean average precision improves, as the recall is improved by the expansion modules (the MAP goes from 35.87 to 44.29 for the taxonomic expansion module and even to 51.61 for the geographical expansion module). Both expansion modules separately improve the results, this effect is not observed when both are used at the same time. This is likely due to the fact that it expands the queries so much that the results become too broad. The expansion modules are also not precise in the sense that in advance they do not know which query terms are taxonomic and which ones are geographic of nature, therefore it could be that the a query term such as *marinus* (as part of the species name *Chaunus marinus*) is expanded in the geographical module where it finds a match with *Marinus Canyon* and *Kavakli* (which has *Marinus* as an alternative name). Thus all records that contain either of these terms will also be returned. When the modules are used separately, this effect is cancelled out by the benevolent effects of the expansion.

ALL	UnExp	TaxExp	GeoExp	TaxGeoExp
Precision	33.07	**22.84** ▽	**20.92** ▽	32.88 ▽
Recall	31.67	**68.66** ▲	**83.30** ▲	**61.82** ▲
MAP	30.04	**41.45** ▲	**47.61** ▲	**44.78** ▲
ANY	UnExp	TaxExp	GeoExp	TaxGeoExp
Precision	21.62	**15.88** ▽	21.56 ▽	21.62 •
Recall	84.37	84.37 •	84.37 •	84.37 •
MAP	28.28	28.87 ▲	28.87 ▲	28.87 ▲
CX	UnExp	TaxExp	GeoExp	TaxGeoExp
Precision	40.13	**22.86** ▽	**20.95** ▽	30.38 ▽
Recall	37.59	**69.18** ▲	**85.85** ▲	**54.18** ▲
MAP	35.87	**44.29** ▲	**51.61** ▲	41.14 ▲

Table 6.6: Precision, recall and mean average precision scores for baseline and expansion modules. Numbers in bold face denote significant increase or decrease of results compared to the UnExp column

In Table 6.7, the number of queries for which a relevant record is returned are presented. As Table 6.7 shows, the expansion modules help smooth out the discrepancies between the queries and the data (e.g., taxonomic and language variations) as for the majority of the queries relevant results are returned. What is striking here, is that through the geographical expansion module for the ALL query mode, 78 queries can be answered, leaving only 22 queries remain unanswered (of which 16 are meant to be unanswered, thus MIRA only fails for 6

queries for which a relevant record is to be found). Although the ANY query mode already achieves this number without query expansion, the precision scores for the ANY query mode are considerably lower.

	UnExp	TaxExp	GeoExp	TaxGeoExp
ALL	32	66 ▲	78 ▲	63 ▲
ANY	78	78 ●	78 ●	78 ●
COMPLEX	38	66 ▲	78 ▲	54 ▲

Table 6.7: Number of reptile and amphibian queries for which a relevant result is returned

The expansion modules ensure that for the majority of the queries relevant results are returned, but as the expansion modules also cause more results to be returned, it is important to present more relevant results first. The experiments with the various ranking modules in MIRA are used to investigate what the best approach is to do this. In the majority of the queries that we received from the collection managers, the request is for "any specimens of type X". Hence, in many cases, simply finding one relevant record answers the question. Although the ANY query mode also finds at least one relevant result to 78 of the R&A queries, the COMPLEX query mode provides a higher precision, thus less irrelevant returned results. The results shown in Tables 6.6 and 6.7 indicate that the MIRA system provides a significant improvement in access to their data for herpetologists.

Result Ranking

In Table 6.8, the mean average precision scores of the different ranking modules are presented. Table 6.8 shows that only the NumMatch and GenSpec ranking approaches provide a noteworthy improvement in the mean average precision scores and this effect is only observed for the ANY query mode (from 28.28 to 42.57 for NumMatch and to 42.38 for GenSpec). This is not surprising as fewer irrelevant entries are retrieved for the COMPLEX and ALL query modes due to their strictness. Although the ontology-based ranking approaches do not provide significant improvements, the manual ontology and the wikipedia ontology obtain similar scores. This indicates that although the ontologies are quite different from each other, they both express at least the relations that are important to the ranking modules.

	All	Any	Cx
NumMatch	29.54 ▽	**42.57** ▲	35.42 ▽
GenSpec	30.40 ▲	**42.38** ▲	36.23 ▲
GlobRRMan	30.27 ▲	29.47 ▲	36.15 ▲
GlobRRWiki	30.81 ▲	28.70 ▲	36.68 ▲
QSRRMan	30.24 ▲	29.17 ▽	36.11 ▲
QSRRWiki	30.26 ▲	28.17 ▽	36.13 ▲
Unranked	30.04	28.28	35.87

Table 6.8: Mean average precision results of the ranked results without query expansion for the reptiles and amphibian queries

Result Ranking and Query Expansion

As was already shown in Table 6.8, ranking only improves the results for the unexpanded Any query mode. For the All and Complex query modes ranking even decreases the performance in most cases. To investigate whether this effect persists when the queries are expanded, the best performing ranking methods are also tested in combination with the expansion modules. The results of these experiments are shown in Table 6.9. Again, only the Any query mode benefits from ranking.

All	UnExp	TaxExp	GeoExp
NumMatch	29.54 ▽	45.12 ▲	46.88 ▽
GenSpec	30.40 ▲	39.77 ▽	41.68 ▽
Unranked	30.04	41.45	47.61
Any	**UnExp**	**TaxExp**	**GeoExp**
NumMatch	**42.57** ▲	**42.57** ▲	**42.56** ▲
GenSpec	**42.38** ▲	**39.89** ▲	**39.86** ▲
Unranked	28.28	28.87	28.87
Cx	**UnExp**	**TaxExp**	**GeoExp**
NumMatch	35.42 ▽	44.48 ▲	45.98 ▽
GenSpec	36.23 ▲	39.75 ▽	**41.60** ▽
Unranked	35.87	44.29	51.61

Table 6.9: Mean average precision results expanded ranked reptile and amphibian queries

However, as most of the information requests are aimed at only one relevant result, the overall mean average precision scores do not tell the whole story. Table 6.7 already showed that for as many queries relevant results are found in

either query mode. Although ranking may not work for all relevant results it is still possible that at least one or two relevant results show up at the top of the list. Whether this is the case is investigated through the results presented in Table 6.10.

In Table 6.10, queries are grouped per occurrence at a particular position in the list of returned results. The number of results are binned in bins of size increasing on a base-10 logarithmic scale. Since users typically do not look further than the first 10 results after an information request (Silverstein *et al.*, 1999), the focus is on whether a relevant record is presented within the first 10 results. For the ANY query mode one can see that the ranking helps increase the number of queries for which a relevant entry is presented at rank one goes from 23 up to 45 for the ranking by a match of the query terms found in the genus or species column (the 'GenSpec' ranking module). For the ALL query mode, both the 'NumMatch' and 'GenSpec' ranking modules ensure that at least one relevant result is present in the first 10 results in 50% or more of the cases when used in combination with query expansion. For the COMPLEX query mode the number of queries for which a relevant result is returned within the first 10 results is slightly higher when the ranking modules are used but this effect is minimal.

6.6.2 Results for Birds

In this subsection, the results for the birds experiments are presented. As 30% of the questions mentions a type of material for the birds, such as skin or skeleton, the ANY query mode caused results to explode drastically as the mention of 'skin' causes already 184,983 results to be returned (more than 6/7 of all records in the birds database) making the number of results unmanageable. Therefore, for the birds only the ALL and COMPLEX search modes are investigated.

For the Birds experiments, there was no taxonomic expansion module available as taxonomic resources for birds are not freely available and usable. The ontology-based ranking method could only be run with the manually constructed ontology as the construction of a Wikipedia-based ontology was impaired due to the incompleteness of Wikipedia articles on the different bird species. These are both limitations that in the near future should be resolved for the advancement of data access.

All	Unranked			NumMatch			GenSpec		
	UnExp	TaxExp	GeoExp	UnExp	TaxExp	GeoExp	UnExp	TaxExp	GeoExp
1	29	37	41	28	39	35	30	41	41
2–10	3	7	10	4	15	23	2	9	12
11–100	0	13	16	0	8	16	0	11	15
101–1000	0	9	11	0	4	7	0	5	13
0	68	34	22	68	34	19	68	34	19
Any	Unranked			NumMatch			GenSpec		
	UnExp	TaxExp	GeoExp	UnExp	TaxExp	GeoExp	UnExp	TaxExp	GeoExp
1	23	23	23	32	32	32	45	44	44
2–10	15	15	15	22	22	22	10	10	10
11–100	25	25	25	15	15	15	17	15	15
101–1000	15	15	15	9	9	9	6	9	9
0	22	22	22	22	22	22	22	22	22
Complex	Unranked			NumMatch			GenSpec		
	UnExp	TaxExp	GeoExp	UnExp	TaxExp	GeoExp	UnExp	TaxExp	GeoExp
1	35	40	45	34	39	35	36	42	42
2–10	3	6	8	4	17	25	2	9	12
11–100	0	12	17	0	6	14	0	10	14
101–1000	0	8	8	0	5	7	0	5	13
0	62	34	22	62	34	19	62	34	19

Table 6.10: Rank first relevant result reptile and amphibian queries

Query Expansion

In Table 6.11, the precision, recall and mean average precision of the experiments on the birds queries without ranking are presented. Similar to the experiments for the reptiles and amphibians, the geographical expansion module provides for a major increase in recall. The drops in precision and mean average precision however are significant.

	All		Cx	
	UnExp	GeoExp	UnExp	GeoExp
Precision	46.33	**08.66** ▽	46.97	**12.05** ▽
Recall	46.83	**65.42** ▲	47.70	**87.22** ▲
MAP	46.17	**12.06** ▽	46.61	**16.36** ▽

Table 6.11: Precision, recall and mean average precision for unranked expanded bird queries

As the results in Table 6.12 show, query expansion in combination with the Complex query mode nearly doubles the number of queries for which a relevant record is returned, leaving only 3 of the queries for which a relevant record is

present in the database unanswered as there are 16 queries that are meant to be unanswered.

	UnExp	GeoExp
ALL	34	61 ▲
COMPLEX	35	81 ▲

Table 6.12: Number of birds queries for which one or more relevant results are retrieved

Result Ranking and Query Expansion

The results for the ranking modules are presented in Table 6.13. The ranking modules do not significantly improve the mean average precision scores, but this was also not the case for the ALL and COMPLEX queries modes for the reptiles and amphibians. Both query modes already perform quite well without query expansion or ranking, for the queries for which an answer is found, therefore, neither query expansion, nor ranking aids the mean average precision scores. This is probably due to the fact that the birds queries are often simpler than the reptiles and amphibians queries, as there are fewer queries that enquire after more than a genus and a species name.

	ALL		CX	
	UnExp	GeoExp	UnExp	GeoExp
NumMatch	46.17 •	14.07 ▲	46.61 •	16.36 •
GenSpec	46.17 •	**19.94** ▲	46.61 •	21.16 ▲
GlobRRMan	46.17 •	13.37 ▲	46.71 ▲	14.94 ▽
QSRRMan	46.17 •	14.15 ▲	46.61 •	20.03 ▲
Unranked	46.17	12.06	46.61	16.36

Table 6.13: Mean average precision results ranked results unexpanded and expanded birds queries

As Table 6.14 shows, there is some variation in the ranking results. Although the unexpanded queries leave many queries unanswered, for the ones MIRA does find an answer, it places it on top in the majority of the cases. However, when the query expansion module is employed, only in a minority of the cases is a relevant record found within the top 10. This is not surprising, as for the reptiles and amphibians, the ranking modules also did not improve results of the ALL and COMPLEX query modes, so it is not surprising that this is also not the case for the birds.

All	Unranked		NumMatch		GenSpec		GlobalRR		LocalRR	
	UnExp	Geo	UnExp	Geo	UnExp	Geo	UnExp	Geo	UnExp	Geo
1	33	9	32	9	32	15	33	13	31	11
2–10	1	7	2	7	2	5	1	7	3	4
11–100	0	21	0	15	0	13	0	11	0	8
101–1000	0	24	0	25	0	31	0	21	0	24
0	66	39	66	44	66	36	66	48	66	53
CX	Unranked		NumMatch		GenSpec		GlobalRR		LocalRR	
	UnExp	Geo	UnExp	Geo	UnExp	Geo	UnExp	Geo	UnExp	Geo
1	34	12	33	12	33	15	33	13	32	15
2–10	1	9	2	9	2	7	2	10	3	6
11–100	0	24	0	24	0	19	0	14	0	17
101–1000	0	36	0	36	0	40	0	40	0	38
0	65	19	65	19	65	19	65	23	65	24

Table 6.14: Rank first relevant result birds queries

6.7 Discussion and Conclusions

In this chapter, a retrieval system was presented that improves retrieval results through query interpretation, query expansion and result ranking. The novelty of the approach presented in the Mira system is the fact that domain knowledge is utilised in three stages of the retrieval process. While the mean average precision scores stay just below 50, in the majority of the queries that we received from the collection managers, the request is for "any specimens of type X". Hence, in many cases, simply finding one relevant record answers the question. When this is taken into account, the fact that for the reptiles and amphibians only 6 queries remain unanswered by using the Mira interpretation and expansion modules, compared to 54 in the baseline system and for the birds 3 queries remain unanswered compared to 52 in the baseline system, is a very useful result. Therefore, it may be concluded that Mira provides a significant improvement in access to natural history data. Also, the modules that significantly improve the retrieval results on a set of 100 test queries for the R&A domain and 100 test queries for the birds domain utilise some sort of domain knowledge. Hence, research question 3 can be answered:

RQ3: Can access to information be aided through a retrieval system enriched with domain knowledge?

The answer to **RQ3** is that access to information is indeed aided by a retrieval system that is based on domain knowledge. However, not all means of incorporating domain knowledge yield the same results. The experiments presented in the chapter show that query interpretation and expansion provide the greatest increases in performance. The ranking modules tested in MIRA do not provide significant improvements in result ranking, in particular for the experiments on the birds data.

When queries are not interpreted, the ANY query mode can attain similar performance but only when combined with ranking, in addition to the geographical expansion. However, the ANY query mode is highly imprecise and for some queries retrieves too many results to process and present to a user as the experiments on the birds database have shown, thus query interpretation is a better option.

The ontology-based ranking modules did not perform as well as expected, but an interesting outcome is that the manually constructed ontology and the TWIBIO ontology did yield similar performances. This may indicate that both ontological perspectives on the data provide at least the relations that are most important for the domain.

Complete automatic query interpretation was not possible, due to the unavailability of taxonomic resources for the birds and incompleteness of taxonomic resources for the reptiles and amphibians. This is a limitation to any system that utilises external knowledge. However, if users are willing to invest a couple of minutes in familiarising with the query interpretation formatting the conversion to the COMPLEX query mode can easily be done manually.

There is room for improvement. In some cases, the geographical expansion module also expands taxonomic query terms. Pre-classifying which terms are to be expanded (for example, by a language filter as taxonomic names are mostly in Latin) or have the user specify this could resolve this problem. The former will possibly slow down the retrieval process, and the latter is more demanding on the user. The implications of the chosen solution, and which option would suit users better should be investigated in tests with users. This is beyond the scope of this work.

7

Conclusions

The machine does not isolate us from the great problems of nature but plunges us more deeply into them.

Antoine de Saint Exupéry, Terre des Hommes, 1939

The work presented in the preceding chapters has provided discoveries and observations regarding the improvement of access to natural history collection information. In this chapter, the thesis contributions, research questions, and problem statement are revisited, after which directions for future work are presented. The chapter is set up as follows. In Section 7.1, the thesis contributions are presented. Section 7.2 provides the answers to the research questions, followed by Section 7.3 in which the problem statement is addressed. The chapter is concluded by Section 7.4 which provides recommendations for future work.

7.1 Thesis Contributions

In this thesis, studies were presented that are aimed at improving access to natural history collection information. The three main difficulties in access to collection information are (a) data quality, (b) data structure, and (c) data access. To solve these problems, hard and soft reasoning approaches were applied to data provided by the Dutch National Museum for Natural History Naturalis.

In Chapter 3, two minor thesis contributions were presented that laid the

groundwork for the main thesis contributions. The first contribution presented in Chapter 3 addresses the backlog in digitisation for the R&A collection by automatically populating the R&A database with nearly 40,000 entries from the R&A field logbooks and registers. No novel approach yielded a significant improvement for the segmenting and labelling of the field logbooks over the results achieved by Lendvai and Hunt (2008). However, analysis of the data in the field logbooks and the registers uncovered large discrepancies between the two resources. Therefore, for the registers a separate training data set was created following the example of (Lendvai and Hunt, 2008) for the field logbooks. MBT was retrained on new annotated data set for the registers. Retraining MBT on a dataset of the same type resulted in a significant improvement of the results for the labelling task on the registers (from F=34.03 to F=75.20). This result shows again that although both the field logbooks and the registers are texts from the same domain describing the same objects, the differences between the two resources are too big to be treated as a single resource. MBT can be successfully trained to deal with the resources separately by annotating a small number of examples of each.

The second contribution, presented in Chapter 3, is a manually constructed ontology that describes the relations between the concepts in the natural history domain. The ontology is based on the CIDOC-CRM reference model, which facilitates sharing collection information with other institutions and integration with other resources. When the ontology is linked to a natural history database a knowledge base is created that provides more structure and thus an enriched representation of the domain than a flat database does.

Three technical thesis contributions were presented that allow the research questions to be answered.

TC1: an automatic ontology-driven error detection and correction method for structured data

In Chapter 4, **TC1**, an ontology-driven data cleaning method named VALIDATO was presented. VALIDATO provides a means to clean up data through making explicit constraints on the data that are imposed by an ontological structure. The ontological structure used in the VALIDATO experiments, is the manually constructed ontology that was presented in Chapter 3. In addition to constraints derived from the ontology, VALIDATO utilises domain knowledge from external resources such as zoological taxonomies and geographical resources. One drawback of this

approach is that the domain knowledge needs to be available. An advantage is that VALIDATO can easily be plugged in to an updated version of a domain specific resource for fast cleanup and updating of a database.

TC2: an instance-driven ontology construction method

In Chapter 5, **TC2**, an ontology construction method named TWIBIO, was presented. TWIBIO is a system that makes relations present in the R&A database explicit through linking it to an external resource that contains explicit information about the domain. The external resource used here is Wikipedia, the online community-generated encyclopaedia that contains rich information about a large variety of topics, including the R&A domain. TWIBIO utilises the fact that Wikipedia articles are linked to each other, and assumes that the existence of a wikilink between two Wikipedia articles indicates a semantic relation between the topics of the two Wikipedia articles. If two Wikipedia articles are linked in one sentence, TWIBIO attempts to extract a relation label for the link from the text, indicating the label of the relation. For each concept in the domain (i.e., each database column) all label candidates are aggregated and the highest rated label as evaluated by four human judges is included in the TWIBIO ontology. Inherent to approaches that utilise external knowledge, is the fact that the approach is limited by the availability of information. Therefore, an extension to TWIBIO that not only uses Wikipedia but also institute intrinsic knowledge (e.g., biographies of persons mentioned as actors in the specimen collection, preservation and description in registration and publications) is to be investigated.

TC3: an ontology-driven information retrieval system that automatically expands queries and ranks the results to improve access

In Chapter 6, **TC3**, a retrieval system named MIRA, was presented. MIRA is a system that utilises domain knowledge to aid retrieval of relevant objects and rank more relevant objects above those that are less relevant. The domain knowledge used comes from three different sources; (1) knowledge from external resources, (2) knowledge from the manual and TWIBIO ontologies, and (3) knowledge about the domain from analyses of typical queries in the domain. The external knowledge and knowledge from the query analysis are used to interpret and expand queries. The ontologies are used to guide MIRA's ranking modules. The knowledge-driven

access that MIRA provides to the R&A and birds databases performs significantly better than simple database access in the experiments that were run.

MIRA concludes the pipeline of digitisation, data cleaning, data structuring, and retrieval presented in this thesis. The research questions are revisited in the next section.

7.2 Answers to Research Questions

In this section, the research questions as presented in Chapter 1 are answered.

RQ1a: Can data-driven and knowledge-driven methods provide improvements to the data quality of structured textual resources describing collection objects?

The data cleaning methods TIMPUTE and VALIDATO have shown to detect a large number of errors in the R&A database. Experiments to estimate the percentage of errors caught with TIMPUTE showed that TIMPUTE is able to capture 80–100% of the artificially introduced spelling errors, 98–100% of the artificially introduced lexical errors and 94–100% of the artificially introduced content errors. Recall was not measured separately for VALIDATO as the nature of the approach prevented a similar experiment. However, as VALIDATO detects and corrects a large number of errors, it can be concluded that VALIDATO also contributes significantly to the improvement of the data quality. Hence, **RQ1a** is answered positively.

b: To what extent are the data-driven and knowledge-driven methods complementary?

TIMPUTE and VALIDATO are complementary in the results they yield. The coverage of the rules used in VALIDATO is small, but precise. VALIDATO's small coverage has the disadvantage that it does not generalise easily to other domains. This in contrast with TIMPUTE, which provides a one-size-fits-all approach that is identical for each database cell and even for each new database. Therefore, if many similar databases are to be checked, adapting VALIDATO to the domain is an option. When there is not much overlap between the databases, TIMPUTE provides an easier option. The results of TIMPUTE and VALIDATO are also different. TIMPUTE is limited to detecting inconsistent cell values in data. VALIDATO

however, besides detecting individual inconsistencies in the database records, also shows inconsistencies or problems in the database schema. For example, for some fields the same inconsistency (with respect to the accepted resource) occurs many times. As it is consistent within the database, TIMPUTE will not identify this as an error or problem, whereas VALIDATO will. These two complementary characteristics of TIMPUTE and VALIDATO support a positive answer to **RQ1b**.

RQ2 Do automatic methods for building ontologies provide different structure for data not achieved by manual ontology building?

The ontology constructed automatically by TWIBIO provides a different structure for the R&A domain than the manually constructed ontology does. This result supports a positive answer to **RQ2**. Further research should clarify whether the insights from the TWIBIO ontology are as useful as those from the manually constructed ontology (or vice versa), which depends on the application the ontology is to be used in. The different ontologies do both reveal information about the perspectives from which they are built; the manually constructed ontology is created from an organisational point of view and is thus more hierarchical, the TWIBIO ontology shows off the aim of an encyclopaedia, namely expressing all relevant relations, leading to a different structure than the one provided by the manually constructed ontology. This insight is valuable in itself, as it illustrates that it is possible to have two different ontological views of one domain.

RQ3: Can access to information be aided through a retrieval system enriched with domain knowledge?

As results of retrieval experiments with MIRA show, access to information can be aided through domain knowledge in the retrieval system, thus **RQ3** can be answered positively. However, not all MIRA modules that utilise domain knowledge provide the same increase in performance of the retrieval results. The experiments presented in Chapter 6 show that query interpretation and expansion provide the greatest increases in performance.

7.3 Answer to Problem Statement

With the answers to the research questions, the problem statement can now be answered.

Problem Statement To what extent can manual and automatic soft- and hard-reasoning approaches improve the data quality, structure, and access to information in an analogue cultural heritage collection of natural history?

In this thesis, three different aspects of collection information accessibility are addressed; (1) data quality, (2) data structure, (3) data access. The hard and soft reasoning techniques presented in this thesis have shown to greatly diminish the problem of data accessibility by turning analogue data resources with limited access into more accessible digital data resources. Each of the main topics is discussed in turn.

Firstly, TIMPUTE and VALIDATO were able to significantly reduce the errors as experiments in Chapter 4 have shown. Moreover, it was possible to detect individual errors as well as errors resulting from a problem with the database schema, providing users with pointers to improve on different aspects of the database.

Secondly, to provide structure to data, two orthogonal ontology construction methods were presented. The first is a manual approach to ontology construction, the second an automatic (TWIBIO). The resulting ontologies provide different, yet worthwhile perspectives on data. The manual ontology provides a hierarchical and object-centred perspective, the TWIBIO ontology provides a semantic relatedness perspective. Both ontologies can provide structure for the data that provides new information for applications or users of the data.

Finally, MIRA, a novel knowledge-driven information retrieval system was presented that enhances the retrieval process through the usage of domain knowledge from ontologies and external resources. MIRA provides significant improvements in retrieval performance through query interpretation, query expansion and ranking of results. Through these modules MIRA provides users with significantly better access to the collection information than before.

To conclude, before the application and development of techniques and systems discussed in the thesis, researchers at Naturalis had to deal with partly corrupt, unorganised, and inaccessible data resources. While the extent of improvement of the techniques and systems presented in this thesis was measured using objective measures for each technique and system separately, the total change in the way researchers can access data goes beyond increases in recall or precision scores. The work presented in this thesis brings the reality of transforming existing analogue information collections to large-scale high-quality and enriched information resources for informatics supported natural history research one step closer.

7.4 Future Work

The future work that can be done based on the work presented in this thesis divides into two categories: improvements on the presented approaches and broader applications.

First, future work should address the open issues of the approaches presented in this thesis and test its generalisability to other domains. Experiments with TIMPUTE on databases from other domains have shown that this approach can be applied successfully there as well (Van den Bosch *et al.*, 2009a). However, the exact applications and limitations of TIMPUTE need to be investigated in further research. A second avenue of research for data cleaning could be dealing with missing values. VALIDATO could be applied to the taxonomic and geographic columns to suggest missing values. For the correction of temporal information, a more sophisticated approach should be developed that, for example, can deduce dates from external resources on the objects described in the data.

Second, the main issue that is not solved by TWIBIO is its dependence on the availability of a high-quality external resource to extract relations from. Currently, it is not possible to relate concepts that do not occur in the external resource to other concepts and this could mean (a) they are not meant to be related to other concepts or (more likely) (b) the resource is incomplete.

Third, research on improving MIRA could investigate more precise query interpretation and query expansion, for example by pre-selecting terms that ought to be included in the expansion. In order to do this, the automatic interpretation module could be used, or an interface could be designed where users could easily do this themselves.

Finally, although some collection managers at Naturalis are already using some of the prototypes developed during the work done for this thesis, such as the reptiles and amphibians database that also contains the automatically segmented and labelled field logs and registers, much work remains to be done. The approaches presented in Chapters 4–6 were limited on some accounts because such resources were not or only partly available. Thus, perhaps the greatest challenge that lies ahead for the biodiversity community is data sharing. With improved data sharing, the approaches presented in this work will contribute to large-scale, high quality resources that provide researchers with easy access to biodiversity data.

References

Aggarwal, C. C. and Yu, P. S. (2001). Outlier detection for high dimensional data. *ACM SIGMOD Record*, **31**(2), 37–46.

Agirre, E. and Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)*.

Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. Technical report, IBM Research Division.

Amar, A. D. (2002). *Managing knowledge workers: unleashing innovation and productivity*. Greenwood Publishing Group, Santa Barbara, CA, USA.

Andreou, A. (2005). *Ontologies and Query expansion*. Master's thesis, School of Informatics, University of Edinburgh.

Auer, S. and Lehmann, J. (2007). What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. In F. et al., editor, *Proceedings of European Semantic Web Conference (ESWC'07)*, volume 4519 of *Lecture Notes in Computer Science*, pages 503–517, Innsbruck, Austria. Springer.

Balmin, A., Hristidis, V., and Papakonstantinou, Y. (2004). ObjectRank: Authority-based keyword search in databases. In M. A. Nascimento, M. T. Özsu, D. Kossmann, R. J. Miller, J. A. Blakeley, and K. B. Schiefer, editors,

Proceedings of the Thirtieth International Conference on Very Large Data Bases (VLDB 2004), pages 564–575, Toronto, Canada. Morgan Kaufmann.

Beeferman, D., Berger, A., and Lafferty, J. (1999). Statistical models for text segmentation. *Machine Learning*, **34**(1-3), 177–210.

Bellare, K. and McCallum, A. (2007). Learning extractors from unlabeled text using relevant databases. In *Proceedings of Sixth International Workshop on Information Integration on the Web (IIWeb-07), in conjunction with AAAI-07*, pages 10–16, Vancouver, Canada. AAAI Press.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, **284**(5), 28–37.

Bitton, D. and DeWitt, D. J. (1983). Duplicate record elimination in large data files. *ACM Transactions on Database Systems*, **8**(2), 255–265.

Blohm, S. and Cimiano, P. (2007). Using the web to reduce data sparseness in pattern-based information extraction. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 18–29, Warsaw, Poland. Springer.

Bobbarjung, D. R., Jagannathan, S., and Dubnicki, C. (2006). Improving duplicate elimination in storage systems. *ACM Transactions on Storage (TOS)*, **2**, 424–448.

Bodner, R. C. and Song, F. (1996). Knowledge-based approaches to query expansion in information retrieval. In *Proceedings of the 11th Biennial Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence*, pages 146–158. Springer-Verlag, London, UK.

Borgo, S., Guarino, N., and Masolo, C. (1996). Stratified ontologies: The case of physical objects. In *Workshop on Ontological Engineering (ECAI'96)*, pages 5–16, Budapest, Hungary.

Borkar, V., Deshmukh, K., and Sarawagi, S. (2001). Automatic segmentation of text into structured records. In *Proceedings of ACM SIGMOD 2001*, pages 175 – 186, Santa Barbara, CA, USA. ACM Press.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*. John Wiley & Sons, Hoboken, NJ, USA.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engines. *Computer Networks and ISDN Systems*, **30**(1-7), 107–117.

Bruni, R. and Sassano, A. (2001). Errors detection and correction in large scale data collecting. In F. Hoffmann, D. J. Hand, N. Adams, D. Fisher, and G. Guimaraes, editors, *Advances in Intelligent Data Analysis, 4th International Conference (IDA 2001)*, pages 84–94, Cascais, Portugal.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, **32**(1), 13–47.

Buitelaar, P., Cimiano, P., and Magnini, B., editors (2005). *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press, Amsterdam, The Netherlands.

Büttcher, S., Clarke, C. L. A., and Cormack, G. V. (2004). Domain-specific synonym expansion and validation for biomedical information extraction (MultiText Experiments for TREC 2004). In *Proceedings of the 13th Text Retrieval Conference*.

Canisius, S. and Sporleder, C. (2007). Bootstrapping information extraction from field books. In *Proceedings of the 2007 Joint Meeting of the Conference on Empirical Methods on Natural Language Processing (EMNLP) and the Conference on Natural Language Learning (CoNLL)*, pages 827–836, Prague, Czech Republic. ACL.

Cannatella, D. C. (1985). *A phylogeny of primitive frogs (archaeobatrachians)*. Ph.D. thesis, The University of Kansas, Lawrence, USA.

Chakrabarti, S. (2007). Dynamic personalized pagerank in entity-relation graphs. In *Proceedings of the 16th international conference on World Wide Web*, pages 571–580, Banff, Alberta, Canada. ACM Press.

Chambers, R., Hentges, A., and Zhao, X. (2004). Robust automatic methods for outlier and error detection. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **167**(2), 323–339.

Chandrasekaran, B., Josephson, J. R., and Benjamins, V. R. (1999). What are ontologies and why do we need them? *IEEE Intelligent Systems*, **12**(5), 20–26.

Chapman, A. D. (2005). Principles and methods of data cleaning. Technical report, Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark.

Chernov, S., Iofciu, T., Nejdl, W., and Zhou, X. (2006). Extracting Semantic Relationships between Wikipedia Categories. In *Proceedings of the First Workshop on Semantic Wikis - From Wiki to Semantics [SemWiki2006] - at the third European Semantic Web Conference (ESWC 2006)*, pages 153 –163, Karlsruhe, Germany. Springer Science+Business Media.

Chi, Y.-L. (2007). Elicitation synergy of extracting conceptual tags and hierarchies in textual document. *Expert Systems with Applications*, **32**(2), 349–357.

Ciaramita, M., Rangemi, A., Ratsch, A., Šarić, E., and Rojas, I. (2005). Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005)*, pages 659–664, Edinburgh, Scotland, U.

Cimiano, P. (2006). *Ontology Learning and Population from Text Algorithms, Evaluation and Applications*. Springer Science+Business Media, New York, NY, USA.

Cole, E., Pisano, E. D., Clary, G. J., Zeng, D., Koomen, M., Kuzmiak, C. M., Seo, B. K., Lee, Y., and Pavic, D. (2006). A comparative study of mobile electronic data entry systems for clinical trials data collection. *International Journal of Medical Informatics*, **75**(10-11), 722–729.

Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, **13**(1), 21–27.

Crofts, N., Doerr, M., Gill, T., Stead, S., and Stiff, M. (2008). Definition of the CIDOC Conceptual Reference Model. Technical report, ICOM/CIDOC CRM Special Interest Group. version 4.2.5.

Daelemans, W., Buchholz, S., and Veenstra, J. (1999). Memory-based shallow parsing. In *Proceedings of CoNLL'99*, pages 53–60, Bergen, Norway.

Daelemans, W., Zavrel, J., Van der Sloot, K., and Van den Bosch, A. (2004). TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide. ILK

Research Group Technical Report Series 04-02, Induction of Linguistic Knowledge, Tilburg University, Tilburg, The Netherlands.

Daelemans, W., Zavrel, J., Van den Bosch, A., and Van der Sloot, K. (2007). MBT: Memory-Based Tagger, version 3.1, Reference Guide. ILK Research Group Technical Report Series 07-08, Induction of Linguistic Knowledge, Tilburg University, Tilburg, The Netherlands.

Dasu, T. and Johnson, T. (2003). *Exploratory data mining and data cleaning*. Wiley-IEEE, Hoboken, NJ, USA.

Diestel, R. (2005). *Graph Theory*. Graduate texts in mathematics. Springer-Verlag, New York, NY, USA, 3rd edition.

Faure, D. and Nedéllec, C. (1998). A corpus-based conceptual clustering method for verb frames and ontology acquisition. In P. Velardi, editor, *Proceedings of LREC workshop on Adapting lexical and corpus resources to sublanguages and applications*, pages 5–12, Granada, Spain.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, **17**(3), 37–54.

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

Frost, D. R. (2009). Amphibian species of the world: an online reference. version 5.3. Electronic Database accessible at http://research.amnh.org/herpetology/amphibia/. American Museum of Natural History, New York, NY, USA.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, **438**(7070), 900–901.

Gómez-Pérez, A. (1998). *The Handbook of Applied Expert Systems*, chapter Knowledge Sharing and Reuse, pages 10–1 –10–36. CRC Press LLC, Boca Raton, FL, USA.

Gong, P. and Mu, L. (2000). Error detection through consistency checking. *Journal of Geographic Information Sciences*, **6**(2), 188–193.

Grant, T., Frost, D. R., Caldwell, J. P., Gagliardo, R., Haddad, C. F. B., Kok, P. J. R., Means, D. B., Noonan, B. P., Schargel, W. E., and Wheeler, W. (2006). Phylogenetic systematics of dart-poison frogs and their relatives (amphibia, athesphatanura, dendrobatidae). *Bulletin of the American Museum of Natural History*, **299**, 1–262.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, **5**(2), 199–220.

Guarino, N., Masolo, C., and Vetere, G. (1999). OntoSeek: Content-based access to the web. *IEEE Intelligent Systems*, **14**(3), 70–80.

Gupta, A. and Jain, R. (1997). Visual information retrieval. *Communications of the ACM*, **40**(5), 70–79.

Guralnick, R. and Hill, A. (2009). Biodiversity informatics: automated approaches for documenting global biodiversity pattern and processes. *Bioinformatics*, **25**(4), 421–428.

Halb, W., Raimond, Y., and Hausenblas, M. (2008). Building linked data for both humans and machines. In *WWW 2008 Workshop: Linked Data on the Web (LDOW2008)*, Bejing, China. ACM Press.

Han, H., Giles, C. L., Manavogly, E., Zha, H., Zhang, Z., and Fox, E. A. (2003). Automatic document metadata extraction using support vector machines. In *In proceedings of Joint Conference on Digital Libraries (JCDL 2003)*, pages 37–48, Houston, TX, USA. ACM Press.

Han, J. and Kamber, M. (2001). *Data Mining Concepts and Techniques*. Morgan Kaufmann, San Francisco, CA, USA.

Hartmann, J., Sure, Y., Giboin, A., Maynard, D., del Carmen Suárez-Figueroa, M., and Cuel, R. (2004). Methods for ontology evaluation. KWeb Deliverable D1.2.3, University of Karlsruhe, Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB).

Haveliwala, T. H. (2002). Topic-Sensitive PageRank. In *Proceedings of WWW2002*, Honolulu, Hawaii, USA. ACM Press.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING-92*, pages 539–545, Nantes, France. ACL.

Hearst, M. A. (1997). TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, **23**(1), 33–64.

Hernández, M. A. and Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, **2**(1), 9–37.

Horridge, M., Knublauch, H., Rector, A., Stevens, R., and Wroe, C. (2004). *A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools Edition 1.0*. The University Of Manchester, Manchester, England, UK, 1.0 edition.

Jaffri, A. (2007). Knowledge enhanced searching on the web. In *Proceedings 6th International Semantic Web Conference (ISWC 2007)*, volume 4825 of *Lecture Notes in Computer Science*, pages 921–925, Busan, Korea. Springer-Verlag.

Jiang, M. F., Tseng, S. S., and Susan, C. M. (2001). Two-phase clustering process for outliers detection. *Pattern Recognition Letters*, **22**(6-7), 691–700.

Kalashnikov, D. V. and Mehrotra, S. (2006). Domain-independent data cleaning via analysis of entity-relationship graph. *ACM Transactions on Database Systems*, **31**(2), 716–767.

Kamps, J. and Koolen, M. (2008). The importance of link evidence in Wikipedia. In C. Macdonald, I. Ounis, V. Plachouras, I. Rutven, and R. W. White, editors, *Advances in Information Retrieval: 30th European Conference on IR Research (ECIR 2008)*, volume 4956 of *Lecture Notes in Computer Science*, pages 270–282, Glasgow, Scotland.

Karr, A. F., Sanil, A. P., and Banks, D. L. (2006). Data quality: A statistical perspective. *Statistical Methodology*, **3**(2), 137–173.

Kavalec, M. and Svátek, V. (2005). *Ontology Learning from Text*, chapter A Study on Automated Relation Labelling in Ontology Learning, pages 44–58. IOS Press, Amsterdam, The Netherlands.

Kedad, Z. and Métais, E. (2002). *Natural Language Processing and Information Systems*, volume 2553 of *Lecture Notes in Computer Science*, chapter Ontology-based Data Cleaning, pages 137–149. Springer-Verlag, Berling/Heidelberg, Germany.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, **46**(5), 604–632.

Knorr, E. M. and Ng, R. T. (1998). Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB'98)*, New York, NY, USA.

Krishtalka, L. and Humphrey, P. S. (2000). Can natural history museums capture the future? *BioScience*, **50**(7), 611–617.

Kubica, J. and Moore, A. (2003). Probabilistic noise identification and data cleaning. In *Proceedings of 3rd IEEE International Conference on Data Mining (ECDM-03)*, pages 131–138, Melbourne, FL, USA. IEEE Press.

Lafferty, J., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289, Williamstown, MA, USA. Morgan Kaufmann.

Lampe, K.-H., Riede, K., and Doerr, M. (2008). Research between natural and cultural history information: Benets and it-requirements for transdisciplinarity. *ACM Journal on Computing and Cultural Heritage*, **1**(1), 1–22.

Lee, M. L., Ling, T. W., and Low, W. L. (2000). Intelliclean: A knowledge-based intelligent data cleaner. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 290 – 294, Boston, MA, USA. ACM Press.

Lendvai, P. and Hunt, S. (2008). From field notes towards a knowledge base. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 644–649, Marrakech, Morocco. European Language Resources Association (ELRA).

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, **10**(8), 707–710. Translated from Doklady Akademii Nauk SSSR, 163(4):845-848, 1965 (Russian).

Linnaeus, C. (1735). Systema naturae.

Loshin, D. (2001). *Enterprise Knowledge Management: The Data Quality Approach*. Morgan Kaufmann, San Francisco, CA, USA.

Maedche, A. and Staab, S. (2000). Discovering conceptual relations from text. In *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI 2000)*, pages 321–325, Berlin, Germany. IOS Press.

Maedche, A. and Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems*, **16**(2), 72–79.

Maedche, A. and Staab, S. (2002). Measuring similarity between ontologies. In *Proceedings of the 13th European Conference on Knowledge Acquisition and Management (EKAW 2002)*, pages 251–263, Siguenza, Spain. Springer.

Maletic, J. and Marcus, A. (2000). Data cleansing: Beyond integrity analysis. In *Proceedings of the Conference on Information Quality (IQ 2000)*, pages 200–209.

Maletic, J. I. and Marcus, A. (2006). *Data Mining and Knowledge Discovery Handbook*, chapter Data Cleansing, pages 21–36. Springer, New York, NY, USA.

Manning, C. D., Raghavan, P., and Schütze., H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England, UK.

McCallum, A. and Wellner, B. (2003). Object consolodation by graph partitioning with a conditionally-trained distance metric. In *Proceedings of ACM SIGKDD Workshop on Data Cleaning, Record Linkage and Object Consolidation*, pages 19–24, Washington, DC, USA. ACM Press.

McCallum, A., Nigam, K., Rennie, J., and K. Seymore, K. (2000). Automating the construction of the internet portals with machine learning. *Information Retrieval Journal*, **3**(2), 127–163.

Medelyan, O., Milne, D., Legg, C., and Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, **67**(9), 716–754.

Michener, C. D., Corliss, J. O., Cowan, R. S., Raven, P. H., Sabrosky, C. W., Squires, D. S., and Wharton, G. W. (1970). Systematics in support of biological research. Technical report, Division of Biology and Agriculture, National Research Council, Washington, DC, USA.

Milano, D., Scannapieco, M., and Catarci, T. (2005). Using Ontologies for XML Data Cleaning. In R. Meersman, Z. Tari, and P. Herrero, editors, *On the Move to Meaningful Internet Systems 2005: OTM Workshops*, pages 562–571. Springer Berlin/Heidelberg, Germany.

Milne, D., Witten, I. H., and Nichols, D. M. (2007). A Knowledge-Based Search Engine Powered by Wikipedia. In *Proceedings of CIKM'07*, pages 445–454, Lisbon, Portugal. ACM Press.

Mooers, C. N. (1948). *Application of Random Codes to the Gathering of Statistical Information*. Master's thesis, Massachusetts Institute of Technology.

Müller, H. and Freytag, J.-C. (2005). Problems, methods, and challenges in comprehensive data cleansing. Technical report, Humboldt-Universität zu Berlin, Germany.

Nakayama, K., Hara, T., and Nishio, S. (2008). Wikipedia Link Structure and Text Mining for Semantic Relation Extraction Towards a Huge Scale Global Web Ontology. In *Proceedings of SemSearch 2008 CEUR Workshop*, pages 59–73, Tenerife, Spain.

Navigli, R. and Velardi, P. (2003). An analysis of ontology-based query expansion strategies. In *Proceedings of 2003 Workshop on Adaptive Text Extraction and Mining (ATEM'03)*, pages 42–49, Cavtat-Dubrovnik, Croatia.

Nguyen, D. P. T., Matsuo, Y., and Ishizuka, M. (2007). Exploiting Syntactic and Semantic Information for Relation Extraction from Wikipedia. In *Proceedings of Workshop on Text-Mining & Link-Analysis (TextLink 2007) at IJCAI 2007*, pages 1414–1420, Hyderabad, India.

Noreen, E. W. (1989). *Computer-Intensive Methods for Testing Hypotheses*. John Wiley & Sons, Hoboken, NJ, USA.

Nothman, J., Murphy, T. R., and Curran, J. R. (2009). Analysing Wikipedia and Gold Standard Corpora for NER Training. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*, pages 612–620, Athens, Greece. ACL.

Noy, N. F. and Hafner, C. D. (1997). The state of the art in ontology design: A survey and comparative review. *AI Magazine*, **18**(3), 53–74.

Noy, N. F. and McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. Technical Report SMI-2001-0880, Stanford University, Stanford, CA, USA.

Orr, K. (1998). Data quality and systems. *Communications of the ACM*, **41**(2), 66–71.

Paijmans, H. (1999). *Explorations in the Document Vector Mordel of Information Retrieval*. Ph.D. thesis, Tilburg University, Tilburg, The Netherlands.

Pantel, P. and Pennacchiotti, M. (2008). *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, chapter Automatically Harvesting and Ontologizing Semantic Relations., pages 171–198. Frontiers in Artificial Intelligence and Applications. IOS Press, Amsterdam, The Netherlands.

Peterson, A. T. and Vieglais, D. (2001). Predicting species invasions using ecological niche modeling: New approaches from bioinformatics attack a pressing problems. *BioScience*, **51**, 363–371.

Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann, San Francisco, CA, USA.

Rabiner, L. R. and Juang, B. H. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2), 257–286.

Radev, D. R., Q, H., Wu, H., and Fan, W. (2002). Evaluating web-based question answering systems. In *Proceedings of the Third International Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain.

Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, **23**(4), 3–13.

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. *ACM SIGMOD Record*, **29**(2), 427–438.

Ravichandran, D. and Hovy, E. (2002). Learning surface text patterns for a quation answering system. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, PA, USA. ACL.

Redman, T. C. (1997). *Data Quality For The Information Age*. Artech House Publishers, Boston, MA, USA.

Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, **41**(2), 79–82.

Reynaert, M. (2005). *Text-induced spelling correction*. Ph.D. thesis, Tilburg University, Tilburg, The Netherlands.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw Hill, New York, NY, USA.

Sanderson, M. and Croft, B. (1999). Deriving concept hierarchies from text. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213, Berkeley, CA. ACM.

Schomaker, L. (1998). From handwriting analysis to pen-computer applications. *IEEE Electronics Communication Engineering Journal*, **10**(3), 93–102.

Schuh, R. T. (2000). *Biological Systematics: principles and applications*. Cornell University Press, Ithaca, NY, USA.

Schulte im Walde, S. (2000). Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 747–753, Saarbrücken, Germany. DFKI, Morgan Kaufmann.

Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. (1999). Analysis of a very large web search engine query log. *ACM SIGIR Forum*, **33**(1), 6–12.

Snyder, B. and Barzilay, R. (2007). Database-Text Alignment via Structured Multilabel Classification. In *Proceedings of the Twentieth International Joint Conferences on Artificial Intelligence (IJCAI-07)*, pages 1713–1718, Hyderabad, India.

Soberón, J. and Peterson, A. T. (2004). Biodiversity informatics: managing and applying primary biodiversity data. *The Philosophical Transactions of the Royal Society*, **359**, 689–698. Published online 18 March 2004.

Spärck-Jones, K. (1971). *Automatic Keyword Classification for Information Retrieval*. Buttersworth, Oxford, England, UK.

Sporleder, C., Van Erp, M., Porcelijn, T., Van den Bosch, A., Arntzen, P., and Van Nieukerken, E. (2006a). Cleaning and enriching research data on reptiles and amphibians. the MITCH pilot project and "nulmeting". Technical Report ILK 06-01, Tilburg University.

Sporleder, C., Van Erp, M., Porcelijn, T., and Van den Bosch, A. (2006b). Correcting 'wrong-column' errors in text databases. In *Proceedings of the Annual Machine Learning Conference of Belgium and The Netherlands (Benelearn-06)*, pages 49–56, Ghent, Belgium.

Sporleder, C., Van Erp, M., Porcelijn, T., Van den Bosch, A., and Arntzen, P. (2006c). Identifying named entities in text databases from the natural history domain. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-06)*, pages 1742–1745, Genoa, Italy.

Sporleder, C., Van Erp, M., Porcelijn, T., and Van den Bosch, A. (2006d). Spotting the 'odd-one-out': Data-driven error detection and correction in textual databases. In *Proceedings of the EACL 2006 Workshop on Adaptive Text Extraction and Mining (ATEM-06)*, pages 41–48, Trento, Italy. ACL.

Stoeckle, M. (2003). Taxonomy, DNA, and the Bar Code of Life. *BioScience*, **53**(9), 2–3.

Suchanek, F. M., Ifrim, G., and Wiekum, G. (2006). Leila: Learning to extract information by linguistic analysis. In *Proceedings of the ACL-06 Workshop on Ontology Learning and Population*, pages 18–25, Sydney, Australia. ACL.

Syed, Z. S., Finin, T., and Joshi, A. (2008). Wikitology: Using Wikipedia as an Ontology. Technical report, University of Maryland, Baltimore County, Baltimore, MD, USA.

Tata, S. and Lohman, G. M. (2008). SQAK: doing more with keywords. In *Proceedings of SIGMOD 2008*, pages 889–902, Vancouver, BC, Canada. ACM.

Tran, T., Cimiano, P., Rudolph, S., and Studer, R. (2007). Ontology-based interpretation of keywords for semantic search. In *Proceedings 6th International*

Semantic Web Conference (ISWC 2007), volume 4825 of *Lecture Notes in Computer Science*, pages 523–536, Busan, Korea. Springer.

Uetz, P., Goll, J., and Hallermann, J. (2008). The reptile database. http://www.reptile-database.org. Last visited: June 4, 2009.

Uschold, M. and Gruninger, M. (1996). Ontologies: Principles, methods and applications. *Knowledge Engineering Review*, **11**(2), 93–136.

Van den Bosch, A., Van Erp, M., and Sporleder, C. (2009a). Making a clean sweep of cultural heritage. *IEEE Intelligent Systems*, **24**(2), 54–63. Special Issue on Cultural Heritage.

Van den Bosch, A., Lendvai, P., Van Erp, M., Hunt, S., Van der Meij, M., and Dekker, R. (2009b). Weaving a new fabric of natural history. *Interdisciplinary Science Reviews*, **34**(2-3), 206–223.

Van Erp, M. (2006). Bootstrapping multilingual geographical gazetteers from corpora. In *Proceedings of the 11th ESSLLI Student Session*, pages 192–202, Málaga, Spain.

Van Erp, M. (2007). Retrieving lost information from textual databases: Rediscovering expeditions from an animal specimen database. In *Proceedings of the ACL 2007 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pages 17–24, Prague, Czech Republic. ACL.

Van Erp, M. and Hunt, S. (2010). Knowledge-driven information retrieval for natural history. In *Proceedings of the 10th Dutch-Belgian Information Retrieval Workshop (DIR 2010)*, pages 31–38, Nijmegen, The Netherlands.

Van Erp, M., Lendvai, P., and van den Bosch, A. (2009a). Comparing alternative data-driven ontological vistas of natural history. In *Proceedings of the eighth International Conference on Computational Semantics (IWCS-8)*, pages 282–285, Tilburg, The Netherlands.

Van Erp, M., Van den Bosch, A., Wubben, S., and Hunt, S. (2009b). Instance-driven discovery of ontological relation labels. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH - SHELT&R)*, pages 60–68, Athens, Greece. ACL.

van Rijsbergen, K. (1979). *Information Retrieval*. Buttersworth, Oxford, England, UK.

Veaux, R. D. D. and Hand, D. J. (2005). How to lie with bad data. *Statistical Science*, **20**(3), 231–238.

Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69, Dublin, Ireland. ACM Press.

Voss, J. (2005). Measuring Wikipedia. In *Proceedings 10th International Conference of the International Society for Scientometrics and Informetrics*, pages 221–231, Stockholm, Sweden.

Wang, X., Hamilton, H. J., and Bither, Y. (2005). An ontology-based approach to data cleaning. Technical report, Department of Computer Science, University of Regina, Regina, SK, Canada.

Wilson, E. O. (2000). A global biodiversity map. *Science*, **289**(5488), 2279.

Wubben, S. (2008). Using free link structure to calculate semantic relatedness. ILK Research Group Technical Report Series 08-01, Induction of Linguistic Knowledge, Tilburg University, Tilburg, The Netherlands.

Zhou, Q., Wang, C., Xiong, M., Wang, H., and Yu, Y. (2007). SPARK: Adapting keyword query to semantic search. In *Proceedings 6th International Semantic Web Conference (ISWC 2007)*, volume 4825 of *Lecture Notes in Computer Science*, pages 694–707, Busan, Korea. Springer.

Zhu, X., Wu, X., and Yang, Y. (2004). Error detection and impact sensitive instance ranking in noisy datasets. In *Proceedings of 19th National Conference of Artificial Intelligence (AAAI;04)*, pages 278–383, San Jose, CA, USA. AAAI Press.

A

The Reptiles and Amphibians Database

Taxonomic information

Class Taxonomic name of class, 93.38% filled; 4 values, but only Amphibia and Reptilia are valid

Order Taxonomic name of the order, 93.35% filled; 14 different values of which 4 appear to be duplicates due to spelling variations

Family Taxonomic name of family, 93.00% filled; 84 different values

Genus Taxonomic name of genus, 98.87% filled; 650 different values

Species Taxonomic name for species, 98.71% filled; 1351 different values with spelling variations and abbreviations possibly forced by database limitations

Sub species Taxonomic name for subspecies, 19.54% filled; 286 different values including spelling variations

Author Name(s) and year of publication in taxonomy, 89.17% filled; 1038 values

Determination Date Date of determination, 14.42% filled; 249 different values, mostly DD-MM-YYYY format

Determinator Name of determinator, sometimes also date of determination, 59.49% filled; 152 different values

Type-name Type exemplar name, 2.22% filled; 123 different values including spelling variations; may include name of collector and year of collection

Type Type characterisation of specimen, 3.04% filled; 29 different values including spelling variations

Information on the collection of the specimen

Country Name of country where specimen was found, 9.06%filled; 71 different values; different languages

Country id Unique number referring to country where specimen was found, 99.77%filled; 126 different values; also holds unknown or invalid values

Province Province or state of finding place, 53.44% filled; 507 different values including spelling variations

Town/city Nearest town or description of location related to nearest town of specimen find, 89.56% filled; 3554 different values including spelling variations

Location Free text eld elaborating onf inding place and circumstances, 9.21% filled; 653 different values including spelling variations; English and Dutch

Coordinates Coordinates of finding place, 3.37% filled; 118 different values

Altitude Altitude level of finding place, 14.39% filled; 334 different values; different units of measurement

Biotope Description of environmental conditions at finding place.,11.65% filled; 700 different values including spelling and mark-up variations; English and Dutch

Collector Name(s) of collector(s), 88.64% filled; 1056 different values including spelling variations

Collection Date Date the specimen was collected, 84.66% filled; 2997 different values including mark-up variations; mostly numeric and in DD-MM-YYYY format

Collection date (old format) Collection date in old format, 2.89% filled; 81 different values including spelling and mark-up variations

Collection # Non-unique identier used by collector, 48.77% filled; 4096 different values including mark-up variations

Information on acceptance of the specimen in the collection and its preservation

Number Number of specimens the record refers to, 90.93% filled; 41 different values, 86% of the records refer to 1 specimen

Preservation method Description of conservation method, 91.10% filled; 43 different values including spelling variations

Donator Name of donator, sometimes also year of donation, 26.05% filled; 508 different values including spelling and mark-up variations

Entry Date Date of acceptance by the museum, 54.20% filled; 772 different values

Label information Serial number on label, 52.13% filled; 1814 different values including spelling and mark-up variations; different formats

Printed Auxiliary field indicating whether the specimen label is printed or handwritten, 99.62% filled; 7 different values

Database information

Record id Unique identier automatically generated by Microsoft Access,100% filled; 16,870 different values

Registration # Numerical registry index, 100% lled; 16,769 different values

Recorder Name of person that entered the record, 100% lled; 10 different values including mark-up variations

Recorder date Date the record was entered, 99.89% filled; 535 different values

Recorder time Time the record was entered, 59.26% filled; 7555 different values.

Characteristics of the specimen

Publication Free text field referring to/elaborating on publication in English and Dutch, 13.38% filled; 87 values including spelling variations

Sex Description of sex, including juvenile, 17.47% filled; 48 different values including spelling variations

Special Remarks Free text field with comments on anything that does not fit into the other fields, 56.97% filled; 2538 different values including spelling variations; mostly Dutch.

B

The Birds Database

Taxonomic Information

Family Family value as indicated on the specimen label, 14.30% filled; 147 different values

Label Name Genus Genus value as originally indicated on the specimen label 86.84% filled; 3901 different values

Taxon Name.Genus Corrected genus value, 54,05% filled; 822 different values

Label name Species Species value as originally indicated on the specimen label 86.56% filled; 6,689 different values

Taxon Name Species Corrected species value, 54.34% filled; 1.972 different values

Label Name Subspecies Subspecies value as originally indicated on the specimen label 35.50% filled; 4,371 different values

Taxon Name Subspecies Corrected subspecies value, 47,92% filled; 3,002 different values

Author Reference to the work in which the species was first defined, 72.58% filled; 4,294 different values

Type-name Reference to the type specimen of the species, 0.01% filled; 18 different values

Information on the Collection of a Specimen

Country Country where the specimen was collected, 48.67% filled; 920 different values, including spelling and language variation

Provincie Province or state where the specimen was collected, 1.72% filled; 14 different values

Region Region where the specimen was collected, 60.49% filled, 5,895 different values

Locality Additional location information on where the specimen was found 72.79% filled; 15,074 different values

Collector Name(s) of collector(s), 63.76% filled; 4.187 different values

Collection Date Date the specimen was collected, 78.62% filled; 32,911 different values

Information on acceptance of the specimen in the collection and its preservation

Number Number of specimens the record refers to, 45.93% filled; 44 different values

Preservation method Description of conservation method, 68.29% filled; 7 different values

Donator Name of donator, sometimes also year of donation, 55.43% filled; 2,219 different values

Acquisitiondate Date of acceptance by the museum, 54.25% filled; 6750 different values

Printed Auxiliary field indicating whether the specimen label is printed or handwritten, 14.30% filled; 3 different values

Captured Auxiliary field indicating whether the specimen died in captivity, 31.63% filled; 22 different values

Collection Information on the collection the specimen belongs to, 16.66% filled; 72 different values

Catalogue number Information on the catalogue in which the specimen is described, 73.27% filled; 779 different values

Database Information

Record id Unique identier automatically generated by Microsoft Access, 100% filled; 215,119 different values

Registration number Numerical registry index. 84.66% filled; 11925 different values

Registration number (old format) Old numerical registry index. 31.63% filled; 66,124 different values

Recorder Name of person that entered the record, 99.35% filled; 131 different values

Record date Date the record was entered, 68.36% filled; 1,245 different values

Characteristics of the Specimen

Sexe Description of sex, 36.88% filled; 100 different values, including spelling, markup and language variations and information that belongs in a different database column

Age Age of the specimen when it was collected (e.g., adult or juvenile), 3.64% filled; 43 different values

Type of Material Part of the specimen that is preserved (e.g., full specimen, skin, skull, wing), 58.51% filled;66 different values

Remarks Free text fields with comments on anything that does not fit into the other fields, 33.03% filled; 9,878 different values

Summary

Cultural heritage institutions harbour a vast treasure of information. However, this treasure of information is often confined to the walls of the archive, museum, or library. This thesis is about improving access to cultural heritage collections through digitisation and enrichment.

In this thesis, three themes that improve information access in a digital information collection from the Dutch National Museum for Natural History Naturalis were investigated: data cleaning, information structuring, and object retrieval. The problem statement that guides the research of this thesis is as follows.

> **Problem Statement:** To what extent can manual and automatic soft- and hard-reasoning approaches improve the data quality, structure, and access to information in an analogue cultural heritage collection of natural history?

The novelty in the work done for this thesis is that techniques from the Natural Language Processing field are applied to data from the natural history domain, which had not been done so far. Also, the interaction between soft-reasoning, or data-driven, and hard-reasoning, or knowledge-driven, approaches is investigated.

In Chapter 2, the field of natural history is introduced and some necessary background is given. Moreover, the resources involved in this work are described.

In Chapter 3, experiments and results on the automatic population of a database from semi-structured text are presented, as well as a manually constructed ontology for the natural history domain.

In Chapter 4, the issue of data quality is addressed. The chapter starts with an overview of issues regarding data that contains errors and an analysis of errors in data from the natural history domain. Then, two methods for automatic cleanup of databases are presented: TIMPUTE and VALIDATO. TIMPUTE is a data-driven method that checks the database for inconsistent values by predicting database values on the basis of all other values in the database. VALIDATO is a hard-reasoning method that utilises domain knowledge from the ontology presented in Chapter 3, as well as from external resources to check database values. Both TIMPUTE and VALIDATO detect a large number of inconsistencies in the data. The two approaches yield complementary results, as they detect different types of errors.

In Chapter 5, an automatic ontology construction method is presented. The chapter starts with a discussion of automatic ontology construction approaches. Then the approach that is developed in MITCH called TWIBIO is described. TWIBIO makes the implicit domain information present in the R&A database explicit by linking it to the online encyclopaedia Wikipedia. From Wikipedia, TWIBIO extracts relations between different database cells, which are then aggregated to find relations between the different database columns. The ontology constructed by TWIBIO provides a different structure for the R&A domain than the manually constructed ontology, which is a reflection of the point of view from the underlying resources used in building the ontology. The manually constructed ontology is created from an organisational point of view and is thus more hierarchical, the TWIBIO ontology shows off the aim of an encyclopaedia, namely expressing all relevant information, leading to a more unorganised structure. This insight is valuable in itself, as it illustrates that it is possible to have two different ontological views of one domain.

In Chapter 6, improvements for data retrieval are presented. Here, the MITCH Information Retrieval Appliance, or MIRA, is presented. The chapter starts with a short introduction of the field of information retrieval, then the resources used for the MIRA experiments are discussed after which the MIRA system is presented. MIRA is novel in that it utilises three different types of domain knowledge in three different stages of the retrieval process. It utilises knowledge from external resources and rules to interpret the queries to formulate more precise queries. It utilises the same types of knowledge to expand queries with synonyms to increase recall. To rank results by relevance, MIRA utilises knowledge from the domain

ontologies and query analysis. MIRA provides a significant improvement in data access as it decreases the number of unanswered queries. However, not all Mira modules that utilise domain knowledge provide the same increase in performance of the retrieval results. The experiments show that query interpretation and query expansion provide the greatest increases in performance.

Chapter 7 summarises to what extent the problem statement and each of the research questions are answered and provides conclusions and recommendations for future work.

Samenvatting

Erfgoedinstellingen bezitten een rijke schat aan informatie. Deze informatieschat is echter vaak niet toegankelijk buiten de muren van het archief, het museum, of de bibliotheek. Dit proefschrift behandelt het verbeteren van de toegankelijkheid van cultureel erfgoedcollecties door digitalisatie en verrijking van de informatie.

In dit proefschrift worden drie thema's behandeld ten behoeve van de toegang tot een digitale informatiecollectie van het Nationaal Natuurhistorisch Museum Naturalis: data opschoning, informatie structurering, en object retrieval. De probleemstelling die het onderzoek leidt luidt als volgt.

> **Probleemstelling:** In hoeverre kunnen handmatige en automatische data- en kennis-gedreven technieken de kwaliteit en de structuur van data en toegang tot informatie in een analoge cultureel erfgoed collectie van natuurlijke historie verbeteren?

Het vernieuwende van dit onderzoek is dat technieken uit het vakgebied van de natuurlijke taalverwerking zijn toegepast op data uit het natuurhistorisch domein. Bovendien is de interactie tussen verscheidene *soft-reasoning*, of data-gedreven, en *hard-reasoning*, of kennis-gedreven, methoden onderzocht.

In Hoofdstuk 2 wordt het natuurhistorisch domein geïntroduceerd en worden de achtergronden van het onderzoek gegeven. Ook wordt de data die gebruikt is in het onderzoek beschreven.

In Hoofdstuk 3 worden experimenten en resultaten gepresenteerd met betrekking tot het automatisch vullen van een database met semi-gestructureerde teksten. Eveneens wordt een handmatig gemaakte ontologie voor het natuurhistorisch domein beschreven.

In Hoofdstuk 4 wordt het probleem van data kwaliteit behandeld. Het hoofdstuk begint met een beschrijving van problemen die kunnen ontstaan door het gebruik van data die fouten bevatten. Vervolgens wordt een analyse van fouten in natuurhistorische data gepresenteerd. Daarna worden twee methoden voor de automatische opschoning van databases gepresenteerd: TIMPUTE en VALIDATO. TIMPUTE is een data-gedreven methode die de database opschoont door de waarden voor de databasecellen te voorspellen aan de hand van waarden in andere databasecellen. VALIDATO is een kennisgedreven methode die gebruik maakt van domeinkennis uit de ontologie die beschreven is in Hoofdstuk 3 en uit externe kennisbronnen. Voor zowel TIMPUTE als VALIDATO geldt dat ze een groot aantal inconsistente waarden in de data opsporen. Bovendien zijn de resultaten van beide methoden complementair, omdat ze verschillende soorten fouten opsporen.

In Hoofdstuk 5 wordt een methode voor automatische ontologieconstructie gepresenteerd. Het hoofdstuk begint met een discussie over automatische ontologie constructie methoden. Daarna wordt de methode die is ontwikkeld binnen het MITCH project, genaamd TWIBIO beschreven. TWIBIO maakt de domeinkennis die impliciet aanwezig is in de R&A database expliciet door deze te verbinden met de online encyclopedie Wikipedia. Wikipedia wordt gebruikt door TWIBIO om relaties te vinden tussen twee databasecellen, waarna deze relaties geaggregeerd worden om relaties tussen de verschillende databasekolommen te beschrijven. De ontologie die op deze manier geconstrueerd wordt door TWIBIO verschaft een andere structuur voor het R&A domein dan de handmatig geconstrueerde ontologie. In deze verschillende structuren worden de verschillende perspectieven gereflecteerd die ten grondslag liggen aan de verschillende bronnen die gebruikt zijn bij het creëeren van deze ontologieën. De handmatig geconstrueerde ontologie is gecreëerd vanuit een organisatorisch perspectief, wat zich vertaalt in een meer hiërarchische structuur. De TWIBIO ontologie laat het doel van een encyclopedie zien, namelijk alle relevante informatie weergeven, wat leidt tot een vrijere structuur. Dit inzicht is op zichzelf al waardevol, omdat het illustreert dat het mogelijk is om twee ontologische perspectieven op een domein te hebben.

In Hoofdstuk 6 worden verbeteringen in het terugvinden van informatie ge-

presenteerd. In dit hoofdstuk wordt het MITCH zoeksysteem, genaamd MIRA, gepresenteerd. Het hoofdstuk begint met een korte introductie van het vakgebied van *information retrieval*, waarna de bronnen die gebruikt worden in de MIRA-experimenten worden beschreven. Daarna wordt het MIRA-systeem gepresenteerd. MIRA is vernieuwend omdat het drie verschillende soorten domeinkennis gebruikt tijdens de drie verschillende stappen in het retrieval-proces. MIRA gebruikt kennis van externe bronnen en regels om de zoekopdrachten preciezer te kunnen formuleren. Dezelfde soorten domeinkennis worden gebruikt om zoekopdrachten te expanderen met synoniemen om de *recall*te verbeteren. Om de resultaten te ordenen naar relevantie gebruikt MIRA kennis uit domeinontologieën en analyse van de zoekopdrachten. MIRA laat een significante verbetering zien in het terugvinden van relevante informatie: het aantal onbeantwoorde zoekopdrachten daalt van rond de 50% naar minder dan 10%. Niet alle vormen van domeinkennis zorgen er echter voor dat MIRA beter werkt dan een standaard zoeksysteem; de experimenten laten zien dat vooral interpretatie en het expanderen van de zoekopdracht verantwoordelijk zijn voor de verbeteringen.

Hoofdstuk 7 vat samen in welke mate de probleemstelling en de onderzoeksvragen zijn beantwoord en presenteert conclusies en aanbevelingen voor toekomstig werk.

List of Publications

Journal

- Van den Bosch, A., Lendvai, P., Van Erp, M., Hunt, S., Van der Meij, M., and Dekker, R. (2009b). Weaving a new fabric of natural history. *Interdisciplinary Science Reviews*, **34**(2-3), 206–223

- Van den Bosch, A., Van Erp, M., and Sporleder, C. (2009a). Making a clean sweep of cultural heritage. *IEEE Intelligent Systems*, **24**(2), 54–63. Special Issue on Cultural Heritage

Conference & Workshop

- Van Erp, M. and Hunt, S. (2010). Knowledge-driven information retrieval for natural history. In *Proceedings of the 10th Dutch-Belgian Information Retrieval Workshop (DIR 2010)*, pages 31–38, Nijmegen, The Netherlands

- Van Erp, M., Van den Bosch, A., Wubben, S., and Hunt, S. (2009b). Instance-driven discovery of ontological relation labels. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH - SHELT&R)*, pages 60–68, Athens, Greece. ACL

- Van Erp, M., Lendvai, P., and van den Bosch, A. (2009a). Comparing alternative data-driven ontological vistas of natural history. In *Proceedings of the eighth International Conference on Computational Semantics (IWCS-8)*, pages 282–285, Tilburg, The Netherlands

- Van Erp, M. (2007). Retrieving lost information from textual databases: Rediscovering expeditions from an animal specimen database. In *Proceedings of the ACL 2007 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pages 17–24, Prague, Czech Republic. ACL

- Sporleder, C., Van Erp, M., Porcelijn, T., and Van den Bosch, A. (2006b). Correcting 'wrong-column' errors in text databases. In *Proceedings of the Annual Machine Learning Conference of Belgium and The Netherlands (Benelearn-06)*, pages 49–56, Ghent, Belgium

- Sporleder, C., Van Erp, M., Porcelijn, T., and Van den Bosch, A. (2006d). Spotting the 'odd-one-out': Data-driven error detection and correction in textual databases. In *Proceedings of the EACL 2006 Workshop on Adaptive Text Extraction and Mining (ATEM-06)*, pages 41–48, Trento, Italy. ACL

- Van Erp, M. (2006). Bootstrapping multilingual geographical gazetteers from corpora. In *Proceedings of the 11th ESSLLI Student Session*, pages 192–202, Málaga, Spain

- Sporleder, C., Van Erp, M., Porcelijn, T., Van den Bosch, A., and Arntzen, P. (2006c). Identifying named entities in text databases from the natural history domain. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-06)*, pages 1742–1745, Genoa, Italy

Technical Report

- Sporleder, C., Van Erp, M., Porcelijn, T., Van den Bosch, A., Arntzen, P., and Van Nieukerken, E. (2006a). Cleaning and enriching research data on reptiles and amphibians. the MITCH pilot project and "nulmeting". Technical Report ILK 06-01, Tilburg University

Curriculum Vitae

Marieke van Erp was born in Breda, the Netherlands, on 18 November 1982. She studied Language and Artificial Intelligence at Tilburg University, where she received her Bachelor's and Master's degrees in 2005 (both 2.1 Honours).

After her graduation, she stayed at Tilburg University as a Ph.D. student where she worked on the intersection of text analytics and the cultural heritage domain on the Mining for Information in Texts from the Cultural Heritage (MITCH) project. The MITCH project was funded by the Netherlands organization for Scientific Research and as it was a collaboration between Tilburg University and the Dutch National Museum of Natural History Naturalis, she split her time between both project partners. From October 2007 until May 2008, she worked at the USC/ISI Natural Language Processing group to deepen her knowledge of knowledge representation.

She published on her research in two international refereed journals and presented her work at numerous conferences and workshops (see list of publications).

As of October 2009, she is a researcher at the Vrije Universiteit Amsterdam, where she continues to improve access to cultural heritage collections.

SIKS Dissertation Series

1998[1]

1998-1 Johan van den Akker (CWI) DEGAS - An Active, Temporal Database of Autonomous Objects

1998-2 Floris Wiesman (UM) Information Retrieval by Graphically Browsing Meta-Information

1998-3 Ans Steuten (TUD) A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective

1998-4 Dennis Breuker (UM) Memory versus Search in Games

1998-5 Eduard Oskamp (RUL) Computerondersteuning bij Straftoemeting

1999

1999-1 Mark Sloof (VU) Physiology of Quality Change Modelling; Automated Modelling of Quality Change of Agricultural Products

1999-2 Rob Potharst (EUR) Classication using Decision Trees and Neural Nets

1999-3 Don Beal (UM) The Nature of Minimax Search

1999-4 Jacques Penders (UM) The Practical Art of Moving Physical Objects

1999-5 Aldo de Moor (KUB) Empowering Communities: A Method for the Legitimate User-Driven Specication of Network Information Systems

1999-6 Niek Wijngaards (VU) Re-Design of Compositional Systems

1999-7 David Spelt (UT) Verication Support for Object Database Design

1999-8 Jacques Lenting (UM) Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation

2000

2000-1 Frank Niessink (VU) Perspectives on Improving Software Maintenance

2000-2 Koen Holtman (TU/e) Prototyping of CMS Storage Management

2000-3 Carolien Metselaar (UvA) Sociaal-organisatorische Gevolgen van Kennistechnologie; een Procesbenadering en Actorperspectief

2000-4 Geert de Haan (VU) ETAG, A Formal Model of Competence Knowledge for User Interface Design

[1] 1 Abbreviations: SIKS Dutch Research School for Information and Knowledge Systems; CWI Centrum voor Wiskunde en Informatica, Amsterdam; EUR Erasmus Universiteit, Rotterdam; KUB Katholieke Universiteit Bra- bant, Tilburg; KUN Katholieke Universiteit Nijmegen; RUG Rijksuniversiteit Groningen; RUL Rijksuniversiteit Leiden; FONS Ferrologisch Onderzoeksinstituut Nederland/Sweden; RUN Radboud Universiteit Nijmegen; TUD Technische Universiteit Delft; TU/e Technische Universiteit Eindhoven; UL Universiteit Leiden; UM Universiteit Maastricht; UT Universiteit Twente, Enschede; UU Universiteit Utrecht; UvA Universiteit van Amsterdam; UvT Universiteit van Tilburg; VU Vrije Universiteit, Amsterdam.

2000-5 Ruud van der Pol (UM) Knowledge-Based Query Formulation in Information Retrieval

2000-6 Rogier van Eijk (UU) Programming Languages for Agent Communication

2000-7 Niels Peek (UU) Decision-Theoretic Planning of Clinical Patient Management

2000-8 Veerle Coupé (EUR) Sensitivity Analyis of Decision-Theoretic Networks

2000-9 Florian Waas (CWI) Principles of Probabilistic Query Optimization

2000-10 Niels Nes (CWI) Image Database Management System Design Considerations, Algorithms and Architecture

2000-11 Jonas Karlsson (CWI) Scalable Distributed Data Structures for Database Management

2001

2001-1 Silja Renooij (UU) Qualitative Approaches to Quantifying Probabilistic Networks

2001-2 Koen Hindriks (UU) Agent Programming Languages: Programming with Mental Models

2001-3 Maarten van Someren (UvA) Learning as Problem Solving

2001-4 Evgueni Smirnov (UM) Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets

2001-5 Jacco van Ossenbruggen (VU) Processing Structured Hypermedia: A Matter of Style

2001-6 Martijn van Welie (VU) Task-Based User Interface Design

2001-7 Bastiaan Schonhage (VU) Diva: Architectural Perspectives on Information Visualization

2001-8 Pascal van Eck (VU) A Compositional Semantic Structure for Multi-Agent Systems Dynamics

2001-9 Pieter Jan t Hoen (RUL) Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes

2001-10 Maarten Sierhuis (UvA) Modeling and Simulating Work Practice BRAHMS: a Multiagent Modeling and Simulation Language for Work Practice Analysis and Design

2001-11 Tom van Engers (VU) Knowledge Management: The Role of Mental Models in Business Systems Design

2002

2002-1 Nico Lassing (VU) Architecture-Level Modiability Analysis

2002-2 Roelof van Zwol (UT) Modelling and Searching Web-based Document Collections

2002-3 Henk Ernst Blok (UT) Database Optimization Aspects for Information Retrieval

2002-4 Juan Roberto Castelo Valdueza (UU) The Discrete Acyclic Digraph Markov Model in Data Mining

2002-5 Radu Serban (VU) The Private Cyberspace Modeling Electronic Environments Inhabited by Privacy-Concerned Agents

2002-6 Laurens Mommers (UL) Applied Legal Epistemology; Building a Knowledge-based Ontology of the Legal Domain

2002-7 Peter Boncz (CWI) Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications

2002-8 Jaap Gordijn (VU) Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas

2002-9 Willem-Jan van den Heuvel (KUB) Integrating Modern Business Applications with Objectied Legacy Systems

2002-10 Brian Sheppard (UM) Towards Perfect Play of Scrabble

2002-11 Wouter Wijngaards (VU) Agent Based Modelling of Dynamics: Biological and Organisational Applications

2002-12 Albrecht Schmidt (UvA) Processing XML in Database Systems

2002-13 Hongjing Wu (TU/e) A Reference Architecture for Adaptive Hypermedia Applications

2002-14 Wieke de Vries (UU) Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems

2002-15 Rik Eshuis (UT) Semantics and Verication of UML Activity Diagrams for Workow Modelling

2002-16 Pieter van Langen (VU) The Anatomy of Design: Foundations, Models and Applications

2002-17 Stefan Manegold (UvA) Understanding, Modeling, and Improving Main-Memory Database Performance

2003

2003-1 Heiner Stuckenschmidt (VU) Ontology-Based Information Sharing in Weakly Structured Environments

2003-2 Jan Broersen (VU) Modal Action Logics for Reasoning About Reactive Systems

2003-3 Martijn Schuemie (TUD) Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy

2003-4 Milan Petkovìc (UT) Content-Based Video Retrieval Supported by Database Technology

2003-5 Jos Lehmann (UvA) Causation in Articial Intelligence and Law A Modelling Approach

2003-6 Boris van Schooten (UT) Development and Specication of Virtual Environments

2003-7 Machiel Jansen (UvA) Formal Explorations of Knowledge Intensive Tasks

2003-8 Yong-Ping Ran (UM) Repair-Based Scheduling

2003-9 Rens Kortmann (UM) The Resolution of Visually Guided Behaviour

2003-10 Andreas Lincke (UT) Electronic Business Negotiation: Some Experimental Studies on the Interaction between Medium, Innovation Context and Cult

2003-11 Simon Keizer (UT) Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks

2003-12 Roeland Ordelman (UT) Dutch Speech Recognition in Multimedia Information Retrieval

2003-13 Jeroen Donkers (UM) Nosce Hostem Searching with Opponent Models

2003-14 Stijn Hoppenbrouwers (KUN) Freezing Language: Conceptualisation Processes across ICT-Supported Organisations

2003-15 Mathijs de Weerdt (TUD) Plan Merging in Multi-Agent Systems

2003-16 Menzo Windhouwer (CWI) Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouse

2003-17 David Jansen (UT) Extensions of Statecharts with Probability, Time, and Stochastic Timing

2003-18 Levente Kocsis (UM) Learning Search Decisions

2004

2004-1 Virginia Dignum (UU) A Model for Organizational Interaction: Based on Agents, Founded in Logic

2004-2 Lai Xu (UvT) Monitoring Multi-party Contracts for E-business

2004-3 Perry Groot (VU) A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving

2004-4 Chris van Aart (UvA) Organizational Principles for Multi-Agent Architectures

2004-5 Viara Popova (EUR) Knowledge Discovery and Monotonicity

2004-6 Bart-Jan Hommes (TUD) The Evaluation of Business Process Modeling Techniques

2004-7 Elise Boltjes (UM) VoorbeeldI G Onderwijs; Voorbeeldgestuurd Onderwijs, een Opstap naar Abstract Denken, vooral voor Meisjes

2004-8 Joop Verbeek (UM) Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale Politiële Gegevensuitwisseling en Digitale Expertise

2004-9 Martin Caminada (VU) For the Sake of the Argument; Explorations into Argument-based Reasoning

2004-10 Suzanne Kabel (UvA) Knowledge-rich Indexing of Learning-objects

2004-11 Michel Klein (VU) Change Management for Distributed Ontologies

2004-12 The Duy Bui (UT) Creating Emotions and Facial Expressions for Embodied Agents

2004-13 Wojciech Jamroga (UT) Using Multiple Models of Reality: On Agents who Know how to Play

2004-14 Paul Harrenstein (UU) Logic in Conict. Logical Explorations in Strategic Equilibrium

2004-15 Arno Knobbe (UU) Multi-Relational Data Mining

2004-16 Federico Divina (VU) Hybrid Genetic Relational Search for Inductive Learning

2004-17 Mark Winands (UM) Informed Search in Complex Games

2004-18 Vania Bessa Machado (UvA) Supporting the Construction of Qualitative Knowledge Models

2004-19 Thijs Westerveld (UT) Using generative probabilistic models for multimedia retrieval

2004-20 Madelon Evers (Nyenrode) Learning from Design: facilitating multidisciplinary design teams

2005

2005-1 Floor Verdenius (UvA) Methodological Aspects of Designing Induction-Based Applications

2005-2 Erik van der Werf (UM) AI techniques for the game of Go

2005-3 Franc Grootjen (RUN) A Pragmatic Approach to the Conceptualisation of Language

2005-4 Nirvana Meratnia (UT) Towards Database Support for Moving Object data

2005-5 Gabriel Infante-Lopez (UvA) Two-Level Probabilistic Grammars for Natural Language Parsing

2005-6 Pieter Spronck (UM) Adaptive Game AI

2005-7 Flavius Frasincar (TU/e) Hypermedia Presentation Generation for Semantic Web Information Systems

2005-8 Richard Vdovjak (TU/e) A Model-driven Approach for Building Distributed Ontology-based Web Applications

2005-9 Jeen Broekstra (VU) Storage, Querying and Inferencing for Semantic Web Languages

2005-10 Anders Bouwer (UvA) Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments

2005-11 Elth Ogston (VU) Agent Based Matchmaking and Clustering - A Decentralized Approach to Search

2005-12 Csaba Boer (EUR) Distributed Simulation in Industry

2005-13 Fred Hamburg (UL) Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen

2005-14 Borys Omelayenko (VU) Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics

2005-15 Tibor Bosse (VU) Analysis of the Dynamics of Cognitive Processes

2005-16 Joris Graaumans (UU) Usability of XML Query Languages

2005-17 Boris Shishkov (TUD) Software Specication Based on Re-usable Business Components

2005-18 Danielle Sent (UU) Test-selection strategies for probabilistic networks

2005-19 Michel van Dartel (UM) Situated Representation

2005-20 Cristina Coteanu (UL) Cyber Consumer Law, State of the Art and Perspectives

2005-21 Wijnand Derks (UT) Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics

2006

2006-1 Samuil Angelov (TU/e) Foundations of B2B Electronic Contracting

2006-2 Cristina Chisalita (VU) Contextual issues in the design and use of information technology in organizations

2006-3 Noor Christoph (UvA) The role of metacognitive skills in learning to solve problems

2006-4 Marta Sabou (VU) Building Web Service Ontologies

2006-5 Cees Pierik (UU) Validation Techniques for Object-Oriented Proof Outlines

2006-6 Ziv Baida (VU) Software-aided Service Bundling - Intelligent Methods & Tools for Graphical Service Modeling

2006-7 Marko Smiljanic (UT) XML schema matching balancing efciency and effectiveness by means of clustering

2006-8 Eelco Herder (UT) Forward, Back and Home Again - Analyzing User Behavior on the Web

2006-9 Mohamed Wahdan (UM) Automatic Formulation of the Auditors Opinion

2006-10 Ronny Siebes (VU) Semantic Routing in Peer-to-Peer Systems

2006-11 Joeri van Ruth (UT) Flattening Queries over Nested Data Types

2006-12 Bert Bongers (VU) Interactivation - Towards an e-cology of people, our technological environment, and the arts

2006-13 Henk-Jan Lebbink (UU) Dialogue and Decision Games for Information Exchanging Agents

2006-14 Johan Hoorn (VU) Software Requirements: Update, Upgrade, Redesign - towards a Theory of Requirements Change

2006-15 Rainer Malik (UU) CONAN: Text Mining in the Biomedical Domain

2006-16 Carsten Riggelsen (UU) Approximation Methods for Efcient Learning of Bayesian Networks

2006-17 Stacey Nagata (UU) User Assistance for Multitasking with Interruptions on a Mobile Device

2006-18 Valentin Zhizhkun (UvA) Graph transformation for Natural Language Processing

2006-19 Birna van Riemsdijk (UU) Cognitive Agent Programming: A Semantic Approach

2006-20 Marina Velikova (UvT) Monotone models for prediction in data mining

2006-21 Bas van Gils (RUN) Aptness on the Web

2006-22 Paul de Vrieze (RUN) Fundaments of Adaptive Personalisation

2006-23 Ion Juvina (UU) Development of a Cognitive Model for Navigating on the Web

2006-24 Laura Hollink (VU) Semantic Annotation for Retrieval of Visual Resources

2006-25 Madalina Drugan (UU) Conditional log-likelihood MDL and Evolutionary MCMC

2006-26 Vojkan Mihajlovic (UT) Score Region Algebra: A Flexible Framework for Structured Information Retrieval

2006-27 Stefano Bocconi (CWI) Vox Populi: generating video documentaries from semantically annotated media repositories

2006-28 Borkur Sigurbjornsson (UvA) Focused Information Access using XML Element Retrieval

2007

2007-1 Kees Leune (UvT) Access Control and Service-Oriented Architectures

2007-2 Wouter Teepe (RUG) Reconciling Information Exchange and Condentiality: A Formal Approach

2007-3 Peter Mika (VU) Social Networks and the Semantic Web

2007-4 Jurriaan van Diggelen (UU) Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach

2007-5 Bart Schermer (UL) Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance

2007-6 Gilad Mishne (UvA) Applied Text Analytics for Blogs

2007-7 Natasa Jovanovic (UT) To Whom It May Concern - Addressee Identication in Face-to-Face Meetings

2007-8 Mark Hoogendoorn (VU) Modeling of Change in Multi-Agent Organizations

2007-9 David Mobach (VU) Agent-Based Mediated Service Negotiation

2007-10 Huib Aldewereld (UU) Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols

2007-11 Natalia Stash (TU/e) Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System

2007-12 Marcel van Gerven (RUN) Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty

2007-13 Rutger Rienks (UT) Meetings in Smart Environments; Implications of Progressing Technology

2007-14 Niek Bergboer (UM) Context-Based Image Analysis

2007-15 Joyca Lacroix (UM) NIM: a Situated Computational Memory Model

2007-16 Davide Grossi (UU) Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems

2007-17 Theodore Charitos (UU) Reasoning with Dynamic Networks in Practice

2007-18 Bart Orriens (UvT) On the development and management of adaptive business collaborations

2007-19 David Levy (UM) Intimate relationships with articial partners

2007-20 Slinger Jansen (UU) Customer Conguration Updating in a Software Supply Network

2007-21 Karianne Vermaas (UU) Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005

2007-22 Zlatko Zlatev (UT) Goal-oriented design of value and process models from patterns

2007-23 Peter Barna (TU/e) Specication of Application Logic in Web Information Systems

2007-24 Georgina Ramírez Camps (CWI) Structural Features in XML Retrieval

2007-25 Joost Schalken (VU) Empirical Investigations in Software Process Improvement

2008

2008-1 Katalin Boer-Sorbán (EUR) Agent-Based Simulation of Financial Markets: A modular, continuous- time approach

2008-2 Alexei Sharpanskykh (VU) On Computer-Aided Methods for Modeling and Analysis of Organizations

2008-3 Vera Hollink (UvA) Optimizing hierarchical menus: a usage-based approach

2008-4 Ander de Keijzer (UT) Management of Uncertain Data - towards unattended integration

2008-5 Bela Mutschler (UT) Modeling and simulating causal dependencies on process-aware information systems from a cost perspective

2008-6 Arjen Hommersom (RUN) On the Application of Formal Methods to Clinical Guidelines, an Articial Intelligence Perspective

2008-7 Peter van Rosmalen (OU) Supporting the tutor in the design and support of adaptive e-learning

2008-8 Janneke Bolt (UU) Bayesian Networks: Aspects of Approximate Inference

2008-9 Christof van Nimwegen (UU) The paradox of the guided user: assistance can be counter-effective

2008-10 Wauter Bosma (UT) Discourse oriented Summarization

2008-11 Vera Kartseva (VU) Designing Controls for Network Organizations: a Value-Based Approach

2008-12 Jozsef Farkas (RUN) A Semiotically oriented Cognitive Model of Knowledge Representation

2008-13 Caterina Carraciolo (UvA) Topic Driven Access to Scientic Handbooks

2008-14 Arthur van Bunningen (UT) Context-Aware Querying; Better Answers with Less Effort

2008-15 Martijn van Otterlo (UT) The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Process Framework in First-Order Domains

2008-16 Henriette van Vugt (VU) Embodied Agents from a Users Perspective

2008-17 Martin Opt Land (TUD) Applying Architecture and Ontology to the Splitting and Allying of Enterprises

2008-18 Guido de Croon (UM) Adaptive Active Vision

2008-19 Henning Rode (UT) From document to entity retrieval: improving precision and performance of focused text search

2008-20 Rex Arendsen (UvA) Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer met een overheid op de administratieve lasten van bedrijven

2008-21 Krisztian Balog (UvA) People search in the enterprise

2008-22 Henk Koning (UU) Communication of IT-architecture

2008-23 Stefan Visscher (UU) Bayesian network models for the management of ventilator-associated pneumonia

2008-24 Zharko Aleksovski (VU) Using background knowledge in ontology matching

2008-25 Geert Jonker (UU) Efcient and Equitable exchange in air trafc management plan repair using spender-signed currency

2008-26 Marijn Huijbregts (UT) Segmentation, diarization and speech transcription: surprise data unraveled

2008-27 Hubert Vogten (OU) Design and implementation strategies for IMS learning design

2008-28 Ildikó Flesh (RUN) On the use of independence relations in Bayesian networks

2008-29 Dennis Reidsma (UT) Annotations and subjective machines - Of annotators, embodied agents, users, and other humans

2008-30 Wouter van Atteveldt (VU) Semantic network analysis: techniques for extracting, representing and querying media content

2008-31 Loes Braun (UM) Pro-active medical information retrieval

2008-32 Trung Hui (UT) Toward affective dialogue management using partially observable Markov decision processes

2008-33 Frank Terpstra (UvA) Scientic worflow design; theoretical and practical issues

2008-34 Jeroen De Knijf (UU) Studies in Frequent Tree Mining

2008-35 Benjamin Torben-Nielsen (UvT) Dendritic morphology: function shapes structure

2009

2009-1 Rasa Jurgenelaite (RUN) Symmetric Causal Independence Models

2009-2 Willem Robert van Hage (VU) Evaluating Ontology-Alignment Techniques

2009-3 Hans Stol (UvT) A Framework for Evidence-based Policy Making Using IT

2009-4 Josephine Nabukenya (RUN) Improving the Quality of Organisational Policy Making using Collaboration Engineering

2009-5 Sietse Overbeek (RUN) Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality

2009-6 Muhammad Subianto (UU) Understanding Classication

2009-7 Ronald Poppe (UT) Discriminative Vision-Based Recovery and Recognition of Human Motion

2009-8 Volker Nannen (VU) Evolutionary Agent-Based Policy Analysis in Dynamic Environments

2009-9 Benjamin Kanagwa (RUN) Design, Discovery and Construction of Service-oriented Systems

2009-10 Jan Wielemaker (UvA) Logic programming for knowledge-intensive interactive applications

2009-11 Alexander Boer (UvA) Legal Theory, Sources of Law and the Semantic Web

2009-12 Peter Massuthe (TU/e) Operating Guidelines for Services

2009-13 Steven de Jong (UM) Fairness in Multi-Agent Systems

2009-14 Maksym Korotkiy (VU) From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)

2009-15 Rinke Hoekstra (UvA) Ontology Representation - Design Patterns and Ontologies that Make Sense

2009-16 Fritz Reul (UvT) New Architectures in Computer Chess

2009-17 Laurens van der Maaten (UvT) Feature Extraction from Visual Data

2009-18 Fabian Groffen (CWI) Armada, An Evolving Database System

2009-19 Valentin Robu (CWI) Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets

2009-20 Bob van der Vecht (UU) Adjustable Autonomy: Controling Influences on Decision Making

2009-21 Stijn Vanderlooy (UM) Ranking and Reliable Classification

2009-22 Pavel Serdyukov (UT) Search For Expertise: Going beyond direct evidence

2009-23 Peter Hofgesang (VU) Modelling Web Usage in a Changing Environment

2009-24 Annerieke Heuvelink (VU) Cognitive Models for Training Simulations

2009-25 Alex van Ballegooij (CWI) RAM: Array Database Management through Relational Mapping"

2009-26 Fernando Koch (UU) An Agent-Based Model for the Development of Intelligent Mobile Services

2009-27 Christian Glahn (OU) Contextual Support of social Engagement and Reflection on the Web

2009-28 Sander Evers (UT) Sensor Data Management with Probabilistic Models

2009-29 Stanislav Pokraev (UT) Model-Driven Semantic Integration of Service-Oriented Applications

2009-30 Marcin Zukowski (CWI) Balancing vectorized query execution with bandwidth-optimized storage

2009-31 Sofiya Katrenko (UvA) A Closer Look at Learning Relations from Text

2009-32 Rik Farenhorst and Remco de Boer (VU) Architectural Knowledge Management: Supporting Architects and Auditors

2009-33 Khiet Truong (UT) How Does Real Affect Affect Affect Recognition In Speech?

2009-34 Inge van de Weerd (UU) Advancing in Software Product Management: An Incremental Method Engineering Approach

2009-35 Wouter Koelewijn (UL) Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling

2009-36 Marco Kalz (OUN) Placement Support for Learners in Learning Networks

2009-37 Hendrik Drachsler (OUN) Navigation Support for Learners in Informal Learning Networks

2009-38 Riina Vuorikari (OU) Tags and self-organisation: a metadata ecology for learning resources in a multi- lingual context

2009-39 Christian Stahl (TUE, Humboldt-Universität zu Berlin) Service Substitution A Behavioral Approach Based on Petri Nets

2009-40 Stephan Raaijmakers (UvT) Multinomial Language Learning: Investigations into the Geometry of Language

2009-41 Igor Berezhnyy (UvT) Digital Analysis of Paintings

2009-42 Toine Bogers (UvT) Recommender Systems for Social Bookmarking

2010

2010-1 Matthijs van Leeuwen (UU) Patterns that Matter

2010-2 Ingo Wassink (UT) Work flows in Life Science

2010-3 Joost Geurts (CWI) A Document Engineering Model and Processing Framework for Multimedia documents

2010-4 Olga Kulyk (UT) Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments

2010-5 Claudia Hauff (UT) Predicting the Effectiveness of Queries and Retrieval Systems

2010-6 Sander Bakkes (UvT) Rapid Adaptation of Video Game AI

2010-7 Wim Fikkert (UT) A Gesture interaction at a Distance

2010-8 Krzysztof Siewicz (UL) Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments

2010-9 Hugo Kielman (UL) A Politiële gegevensverwerking en Privacy, Naar een effectieve waarborging

2010-10 Rebecca Ong (UL) Mobile Communication and Protection of Children

2010-11 Adriaan Ter Mors (TUD) The world according to MARP: Multi-Agent Route Planning

2010-12 Susan van den Braak (UU) Sensemaking software for crime analysis

2010-13 Gianluigi Folino (RUN) High Performance Data Mining using Bio-inspired techniques

2010-14 Sander van Splunter (VU) Automated Web Service Reconfiguration

2010-15 Lianne Bodenstaff (UT) Managing Dependency Relations in Inter-Organizational Models

2010-16 Sicco Verwer (TUD) Efficient Identification of Timed Automata, theory and practice

2010-17 Spyros Kotoulas (VU) Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications

2010-18 Charlotte Gerritsen (VU) Caught in the Act: Investigating Crime by Agent-Based Simulation

2010-19 Henriette Cramer (UvA) People's Responses to Autonomous and Adaptive Systems

2010-20 Ivo Swartjes (UT) Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative

2010-21 Harold van Heerde (UT) Privacy-aware data management by means of data degradation

2010-22 Michiel Hildebrand (CWI) End-user Support for Access to Heterogeneous Linked Data

2010-23 Bas Steunebrink (UU) The Logical Structure of Emotions

2010-24 DmytroTykhonov Designing Generic and Efficient Negotiation Strategies

2010-25 Zulfiqar Ali Memon (VU) Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective

2010-26 Ying Zhang (CWI) XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines

2010-27 Marten Voulon (UL) Automatisch contracteren

2010-28 Arne Koopman (UU) Characteristic Relational Patterns

2010-29 Stratos Idreos(CWI) Database Cracking: Towards Auto-tuning Database Kernels

2010-30 Marieke van Erp (UvT) Accessing Natural History: Discoveries in Data Cleaning, Structuring, and Retrieval

TiCC Dissertation Series

1. Pashiera Barkhuysen
Audiovisual prosody in interaction
Promotor: M.G.J. Swerts, E.J. Krahmer
Tilburg, 3 October 2008

2. Ben Torben-Nielsen
Dendritic morphology: function shapes structure
Promotores: H.J. van den Herik, E.O. Postma
Co-promotor: K.P. Tuyls
Tilburg, 3 December 2008

3. Hans Stol
A framework for evidence-based policy making using IT
Promotor: H.J. van den Herik
Tilburg, 21 January 2009

4. Jeroen Geertzen
Act recognition and prediction. Explorations in computational dialogue modelling
Promotor: H.C. Bunt
Co-promotor: J.M.B. Terken
Tilburg, 11 February 2009

5. Sander Canisius
Structural prediction for natural language processing: a constraint satisfaction approach
Promotores: A.P.J. van den Bosch, W.M.P. Daelemans
Tilburg, 13 February 2009

6. Fritz Reul
New Architectures in Computer Chess
Promotor: H. J. van den Herik
Co-promotor: J. Uiterwijk
Tilburg, 17 June 2009

7. Laurens van der Maaten
Feature Extraction from Visual Data
Promotores: E.O. Postma, H.J. van den Herik
Co-promotor: A.G. Lange
Tilburg, 23 June 2009

8. Stephan Raaijmakers
Multinomial Language Learning: Investigations into the Geometry of Language
Promotores: W. Daelemans, A.P.J. van den Bosch Tilburg, 1 December 2009

9. Igor Berezhnyy
Digital Analysis of Paintings
Promotores: E.O. Postma, H.J. van den Herik
Tilburg, 7 December 2009

10. Toine Bogers
Recommender Systems for Social Bookmarking
Promotor: A. P. J. van den Bosch
Tilburg, 8 December2009

11. Sander Bakkes
Rapid Adaptation of Video Game AI
Promotor: H. J. van den Herik
Tilburg, 3 March 2010

12. Maria Mos
Complex Lexical Items
Promotor: A. P. J. van den Bosch
Co-promotores: A. Vermeer, A. Backus
Tilburg, 12 May 2010

13. Marieke van Erp
Accessing Natural History: Discoveries in Data Cleaning, Structuring and Retrieval
Promotor: A. P. J. van den Bosch
Co-promotor: P. Lendvai
Tilburg, 30 June 2010