17

Natural Selection: Finding Specimens in a Natural History Collection

Marieke van Erp[1], Antal van den Bosch[2], Steve Hunt[2], Marian van der Meij[3], René Dekker[3] and Piroska Lendvai[4]

[1]*VU University Amsterdam, Department of Computer Sciences*
[2]*Tilburg center for Cognition and Communication, Tilburg University*
[3]*Netherlands Centre for Biodiversity Naturalis*
[4]*Research Institute for Linguistics, Hungarian Academy of Sciences*
[1,2,3]*The Netherlands*
[4]*Hungary*

1. Introduction

The natural history domain is rich in information. For hundreds of years, biodiversity researchers have collected specimens and samples, and meticulously recorded the how, what, and where of these objects of research. To retrace this information, however, deep knowledge of the collection and patience is necessary. Whereas traditional access methods (e.g., analysing paper logs of specimen finds) can be used for smaller collections, the sheer size of most current natural history collections prohibits this. At the same time, information technology has advanced to the point where it is able to capture the intricacies of biodiversity collection information and provide the first steps towards full digital access.

The need for collection information access is dire, as lack of access impairs our ability to answer questions about species biodiversity, diversity and change through time (Scoble, 2010). Examples from the young field of biodiversity informatics stress that in order to assess and tackle problems such as predicting a species' reaction to changing environment or prioritisation of preservation policies, digitisation of and access to (large) collection databases is imperative (Guralnick & Hill, 2009; Johnson, 2007; Raes, 2009; Soberón & Peterson, 2004). Although much progress has been made, for example with the Global Biodiversity Data Portal[1] (Berendsohn et al., 2010), many collections have not yet been (fully) digitised.

In this contribution, we first present a new approach to collection digitisation, as well as a novel collection registration management system (CRS) as implemented at the Netherlands Centre for Biodiversity (NCB Naturalis). The new approach to digitisation at NCB Naturalis implements a cascaded digitisation approach: in parts of the collection that have not yet been digitised, first a shelf or drawer is assigned a unique ID in the CRS, along with a description of the specimens contained within it. Whenever the shelf or drawer is revisited, the new policy dictates that specimens that are taken and used from this set be recorded in the CRS. Furthermore, the CRS is linked to taxonomic resources, which enable integration with reference sources. We present two use cases that illustrate the benefits for smarter collection

[1] `http://data.gbif.org`

information management systems, employing natural language processing techniques. The first use case focuses on data cleaning (Section 5), the second on data retrieval (Section 6). Prior to the use cases, we first explain the background of the NCB Naturalis (Section 2), followed by an overview of the key features of the collection registration system (Section 3) and the collection data used in our studies (Section 4).

2. NCB Naturalis and its collections

The Netherlands Centre for Biodiversity Naturalis[2] is a collaboration between the University of Amsterdam[3], Leiden University[4], Wageningen University[5], and the Dutch National Museum of Natural History Naturalis[6]. They form the combined institute that collaborates with the academic partners to foster the expertise in biodiversity in the Netherlands. The institute will harbour the largest natural history collection in the Netherlands, consisting of over 37 million objects, currently the fifth largest collection worldwide.
NCB Naturalis collections contain fossils, vertebrates, invertebrates, insects, botanical and geological specimens. The majority of the specimens are collected in former colonies of the Netherlands in tropical America, South-East Asia, and Africa, but the collection also provides a broad account of Dutch biodiversity.
In order to manage such a collection properly and make parts of it available to researchers, for example via the Internet, a sound management system is needed. Like those of many other natural history institutions in the world, the collections at NCB Naturalis go back a long way in history. Part of the collection dates back to the 18th century and ranges from specimens collected during the voyages of Captain Cook in the South Pacific and Von Siebold and Bürger in Japan to recent marine and terrestrial collections from expeditions to South-East Asia. With the growth of the collections, curation and collection management practices evolved, but only in the past two decades has technology come into play in collection management systems. As with any innovation, use and best practices have needed time to develop and take root. Over the past few years, NCB Naturalis has been taking stock of the various ways each department have organised their collection information and have started to develop an institution-wide collection management system, taking into account the lessons learnt from each department. In the following Section we detail how this has influenced our design choices for the new NCB Naturalis Collection Registration System (CRS).

3. NCB Naturalis collection registration system

The collection registration system (CRS) at NCB Naturalis is novel in the sense that it is specifically designed for natural history collections, by researchers and collection managers at NCB Naturalis in collaboration with a database company. The CRS differs from other collection management systems in that it is not only a collection management tool for a wide range of users that allows retrieving objects in the collection, and inspect what is their condition or whether they are on loan, but also a tool for researchers. Most systems currently used in natural history institutions are developed for only one of these goals. In the CRS, different user roles are defined, that give users rights to see only general data, or or all data.

[2] http://www.ncbnaturalis.nl/
[3] http://www.uva.nl/
[4] http://www.leidenuniv.nl
[5] http://www.wur.nl
[6] http://www.naturalis.nl

Some data fields are restricted (such as the monetary value of an object), and are not made public.

Although the CRS employs its own, custom-made underlying data model, it is based on the Access to Biological Collection Data (ABCD) standard, 'Extension for GeoSciences' (EFG)[7]. It is furthermore compatible with existing protocols as CIDOC-CRM[8], Spectrum[9], and various technical standards.

To overcome the overwhelming backlog in collection registration, the CRS implements a cascaded registration approach; first the drawers containing boxes filled with specimens are registered, then the boxes contained in the drawers, and finally the individual specimens. This ensures that at least series of specimens are registered and can be located, which is an important consideration in a collection of 37 million objects. In particular, the entomology collection contains millions of specimens; the cascaded approach moves the recording of individual specimens to the future. In other sub-collections with relatively fewer specimens, for instance those of birds and mammals, each specimen will be recorded.

Whereas in the past users of the collections has the choice to enter a specimen they inspected into the database, new policy enforces that the specimen be entered in the CRS if it does not have an individual record yet. The most basic set of metadata information that can be entered about a specimen or collection unit is the information that is on the labels attached to it. This information can further be enriched by records from existing registration or acquisition books, some of which may already have been digitised and are available as databases in the CRS, or from research data such as field books or scientific publications on the unit. Objects are to be registered by copying information 'as is' from the label or paper register. It is considered important to retain the raw information to avoid information loss that may occur when some of the original paper record is incorrectly deemed unimportant. This is in line with the growing awareness of the importance of always keeping the original data and as much of its provenance information, such as a trace of the permutations on the data (i.e., who did what to the data) (Chapman, 2005).

4. Data used

Reptiles and amphibians

The Reptiles and Amphibians (R&A) database is a resource compiled from a manually created database containing 16,870 records (used in Section 5 and Section 6) and an additional automatically populated database containing 39,688 records (used in Section 6). Together, the manual and automatically created databases cover the entire reptiles and amphibians collection at NCB Naturalis.

Each record describes where, when and under what circumstances a reptile or amphibian specimen in the NCB Naturalis collection was found and how it is preserved. The manually created database was compiled by researchers at the institution. It contains 37 columns, of which twelve contain taxonomic information, and eight contain geographic information. The remaining columns describe additional features of the specimen and administrative information. The automatically populated database was created by automatically segmenting and labelling the field notes and registers (this process is described in (Lendvai & Hunt, 2008)). The database is mostly composed in Dutch and English, but also contains some information in German and Portuguese.

[7] http://www.geocase.eu/efg.asp
[8] http://www.cidoc.ics.forth.gr
[9] http://www.mda.org.uk/schema

Taxonomic resources

For the amphibians, the Frost taxonomy is used, as published online (Frost, 2009). The version used in this work (version 5.3) contains descriptions of 6,433 amphibian specimens with references to the literature and synonyms.
For the reptiles, the TIGR Reptile Database (Uetz et al., 2008) is used. It 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. 8,600 reptile species are described.

GeoNames

GeoNames[10] is an aggregated geographical data base that is available through a Creative Commons attribution license and accessible through various Web services. The GeoNames database is compiled from a collection of smaller geographic resources. In June 2009, GeoNames contained over eight million geographical names, of which 6.5 million unique entities. It is an attractive resource to pair our taxonomic data with, as it contains alternative names for geographic entities in numerous languages. In Section 5, we describe the utilisation of GeoNames for the automated detection of inconsistencies in geographical fields in collection databases. In Section 6, we show how GeoNames can be employed to increase recall while querying a multilingual database.

5. Knowledge-driven data cleaning

While data typists and curators do their utmost to create database records meticulously, errors are impossible to avoid. It is estimated that about 5% or more of all data entered by humans contains errors (Maletic & Marcus, 2000; Orr, 1998; Redman, 1997). Most errors that are reported in natural history data occur in the taxonomic, geographic and 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 or circumscript recordings 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, particularly those 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 locations of their findings.
Errors in person names are less frequent than errors in the taxonomic or geographic information about a specimen. The main error encountered here is 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 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.

[10] `http://www.geonames.org`, Last queried 15 July, 2009

As it is unfeasible to manually correct all records, there is a need for the automatic checking of information in databases, so that experts can be guided towards prioritised lists of potential errors. Another argument for developing a computer-supported means of data correction for taxonomic databases, is that the information in these databases is subject to change as the taxonomy continues to be debated, revised, and expanded. We therefore developed an automatic approach that uses knowledge from existing taxonomic and geographic resources, as well as a set of rules to decide which database values are suspicious.

5.1 Name and date normalisation

Any first step in data cleaning should consist of making sure all data fields are formatted consistently. To show that this is a non-trivial step, we have normalised (1) diacritics in person names (e.g., removal of umlauts), (2) date formatting (i.e., converting dates to yyyy-mm-dd), and (3) name formatting (i.e., converting person names to *lastname, firstname* or *lastname, initials*) in the reptiles and amphibians database. Table 1 illustrates the amount of data affected by these three types of inconsistencies.

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)
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)
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)

Table 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)

The amount of formatting consistency varies greatly; for some person name columns such as *Determinator* and *Donator*, only 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 is a necessary step in data cleanup.

5.2 Content cleaning

Our knowledge-based database cleaning approach utilises knowledge about the domain from taxonomies and other resources to infer whether a value is correct or suspicious. It works by combining pieces of information from the collection information system and an external resource or rule, to decide whether a value is correct or not. We give a schematic example in Figure 1. The fictitious domain ontology with main concepts A, B, and C is represented on the left-hand side. In this figure, the operators $>$, $<$, ==, and != are used to express possible relations in the domain. According to the ontology there should be a $>$ relation between

co-occurring values of concepts A and B, and an == relation between co-occurring values of concepts 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.

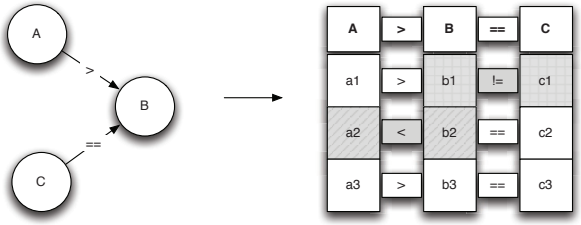

Fig. 1. Schematic overview of the ontology-based error correction approach

In the remainder of this section, we present three types of error detection experiments. The first experiment shows how knowledge about the collection process captured in rules can aid in identifying incorrect database values. In the second and third experiments we link the database to external knowledge sources to find such inconsistencies. In each experiment, we take a database record and compare the values of two different database columns against the knowledge resource (rules or external resource). If one of the values is not consistent with the rule or the resource, it is considered suspicious, and flagged to be checked by a user. In some cases the resource can suggest a correction, but as it is not possible to always know what caused the inconsistency, we choose to keep the human in the loop to determine whether the flagged value is indeed an error, and if so, if the correction is the right value that should replace the incorrect one.

Cleanup of temporal data

To identify inconsistencies in the temporally related information in the database, the database columns containing date information were selected for inspection. The temporal information regarding *Collection* is described by column 'collection 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 four columns, 'collection date', 'entry date', 'determination date', and 'recorder date', are interrelated by *Occurs before* relations. The chronological order of the events related to these dates are summarised in Table 2. Inferred relations are also listed; such a relation, for instance, is 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. The *Occurs before* relation is transitive: if A occurs before B, and B occurs before C, then A occurs before C, and thus it can be inferred that the *Collection* takes place before all other events, up to the *Creation of Database Record* event. This is indicated by the inferred relations in Table 2.

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 2. Summary of chronological relations present in specimen data

The results of the consistency check on dates are presented in Table 3. When a constraint violation is detected by the rule, it is not possible to determine which date is the one containing an erroneous value: the error can be in either value, or in both.

Columns	Flagged by VALIDATO
Collection - Entry	64
Entry - Determination	7
Determination - Recorder	26
Collection - Determination	5
Collection - Recorder	5
Entry - Recorder	2

Table 3. Results of ontology-based error detection experiments on temporal data

The approach has two important limitations. First, the approach cannot suggest a correct date if a constraint violation is encountered, as the domain offers no rules about how much time there should be between the different events. Second, only cases are flagged in which a value is violating constraints, not when the value is incorrect while no constraints are violated, such as when a recorded collection date is off by a few days, but all other dates pertaining to the database record are much later. Information that would be needed to detect and correct errors of this type could come from resources that describe the expedition, such as a logbook, or from employment records at the deteminator's lab. In certain cases this type of detective work may be warranted, and automatic techniques may assist in this type of expert work, but this lies beyond the scope of the current contribution.

Cleanup of geographical data

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; if there is no containment relation found in the returned records the database entry is flagged as containing possibly inconsistent geographic information. The relations that hold between selected geographic classes in the specimen database are summarised in Table 4.
Due to the multilingual nature of the data the rules need to leave room for considerable variation. If for instance the city-country pair 'swamp ca . 10 km E . of Parga'-'Griekenland' is encountered, the value in the city cell is first stripped of all non-capitalised and numeric tokens and tokens shorter than 2 letters. This results in the value 'Parga', which is then queried against the GeoNames database, returning twenty records. Along with every returned result all possible alternatives for the country name in the languages present in the R&A database are

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

Table 4. Summary of relations holding between the different geographic classes in the specimen data

looked up and compared to the original country value 'Griekenland'. In this case, a positive match occurs between a GeoNames match of 'Parga' – 'Greece', 'Griekenland' being the Dutch word for 'Greece'; thus, the record is not flagged as containing inconsistent geographic information.

In cases where no match between the city and country values is found in GeoNames, the database entry is flagged as containing a potential 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.

The results of the experiments are presented per pair of columns in Table 5. The disagreements flagged by the data cleaning system 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 5. 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. of airstrip Tafelberg' and 'Right kabalebo river, kamp keyzer, voet K. valle'. It is indeed a dilemma for the person entering the nearest city name, as specimens are often found well outside habited areas, and the nearest city may not at all be the most obvious anchor point to describe the geographical coordinates of the finding. Yet, entering circumscript phrases such as '5km NW of' or 'near' only obfuscates the precise location. Modern technology can aid in such cases as a location could unambiguously defined by the usage of a GPS device. For older data it would be better to redefine the column as 'city or nearest city', and a separate column 'other localisation information' could be devised in which additional information such as '4km W. of airstrip' could be entered.

Most errors in the city-province test are a systematic mix-up of the two Surinam districts *Nickerie* and *Sipaliwini*, which is a frequency effect of the many expeditions that took place in these districts, and the erroneous values proliferating. In 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. This particular case illustrates the limitations of GeoNames as *Departement Guyane*[11] 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 are 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 because of 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.

Cleanup of taxonomic data

Taxonomic inconsistencies in the data are detected through a process similar to the detection of geographical inconsistencies. The taxonomic hierarchy can be defined through a *Has broader term* relation. This transitive relation applies to 'species', 'genus', 'order', 'family' and 'class' consecutively as shown in Table 6. The 'subspecies' level could not be queried as the data formatting of the resources prevented reliable identification of the 'subspecies' values.

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 6. Summary of hierarchical taxonomic relations holding between the different taxonomic levels

To investigate why the system flagged an instances, 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 a content error (CE). The results are presented in Table 7.
The most peculiar result from the taxonomic 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.
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[12] and the Global

[11] The full official name is Département de la Guyane.
[12] `http://www.eol.org/`

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 7. Results of ontology-based error detection experiments on taxonomic information

Biodiversity Information Facility[13]). 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* were removed from the *Dendrobatidae* family to form 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 genuine errors. As the suborder vs. order error is overly frequent, the addition of a suborder column in the database might be considered.
The same types of knowledge that helped clean up the database can also help increase access to it. Due to the complexity of the data, simple queries are often not enough. As we want to preserve as much information as possible, it is important that for example synonyms of taxons are linked, so that a single query can retrieve all specimens of a species, regardless of the name they are registered by. To show the benefits of this, we carried out experiments with and without simultaneous synonym search.

6. Knowledge-driven specimen access

To improve the accessibility of specimen information in natural history data collections through search engines, we developed a knowledge-driven database access method that utilises domain knowledge at three different stages in the retrieval process. The domain knowledge is employed to (1) aid query formulation, (2) expand queries with relevant synonyms, and (3) rank results. We compared the knowledge-driven access method with the original collection database system. Our results show that the domain knowledge markedly improves recall results on the reptiles and amphibians domain that we tested the approach on: from 32% to 86%.

6.1 Queries

External researchers often request access to NCB Naturalis' extensive specimen collection or to the meta-data that is found in the databases describing the collections. As the databases are

[13] http://www.gbif.org/

not (yet) publicly available, these questions are usually directed to the collection managers at NCB Naturalis. To test the system, collection managers have saved these questions they received regarding the reptiles and amphibians collection. These queries give a good idea of the type of information researchers are looking for.
The questions were extracted from longer (often email) messages. The questions have been summarised into only the information request and not the introduction for why the information is requested. For each of the queries the relevant records in the databases were identified manually to create a gold standard.

Reptile and amphibian queries

The 100 reptile and amphibians queries were gathered from requests to the reptile and amphibian collection managers and researchers at NCB Naturalis that were received between September 2003 and December 2008.
Some example queries are:

- 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 Venezuela in the collection?
- How many species of *Rana palmipes* as defined by Spix in 1824 are in the collection?

12% of the questions enquire after a genus, 86% after a genus and a species, and in 41% the request poses a restriction on the geographical location of where the specimen was collected. Additionally, in 15% of the questions a registration number is given, which should make it easier to retrieve correct database record, but as registration numbers are not unique this is not always the case.
For 16 queries no relevant records were present in the database. For the remaining 84 queries the number of returned records varies greatly. For example, for 21 queries only 1 relevant result is present in the database whereas there are 4 queries for which over 500 relevant results are present in the database.

6.2 System architecture

In this section, the system setup is presented. An overview of the system is presented in Figure 2. The domain knowledge comes from taxonomic and geographic resources (see Subsections 4 and 4), a domain ontology, domain-specific rules and analysis of typical queries in the domain. Below, each of the system modules is described.

6.2.1 Query interpretation

Most of the queries in the test sets require more precise formulation than queries using the operators 'and' and 'or'. Consider for example the query *Are there Dipsas species other than D. catesbyi and D. variegata from the Guianas and Venezuela in the collection?*. Here, the user is looking for database records that describe specimens of genus Dipsas, but not those records of species *Dipsas catesbyi* and *Dipsas variegata*. The second constraint is that the user wants the relevant records about specimens collected in the Guianas or Venezuela.
To be able to handle such queries, we devised a query language that can encode that for part of the query any query term should match and for part of the query all query terms should match. The query language can also exclude terms on the basis of a negation. The query terms that we extract from the example query are: *dipsas*, *-catesbyi*, *-variegata*, *guianas* and *venezuela*. To express that specimens of genus *Dipsas* found in the *Guianas* or in

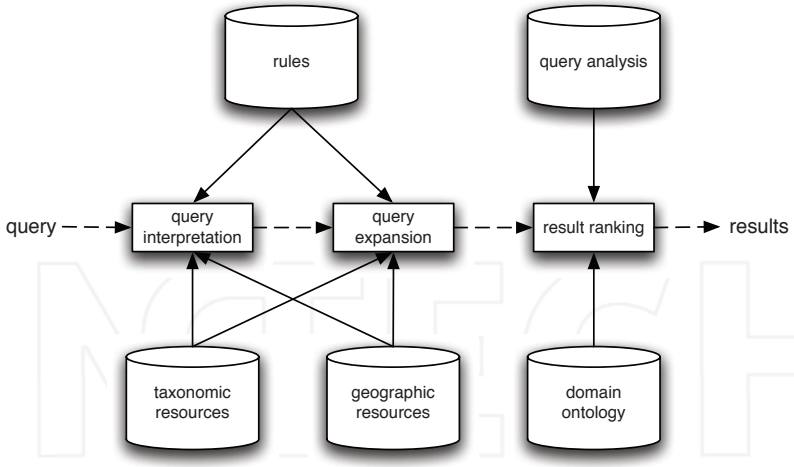

Fig. 2. Overview of the knowledge-driven specimen information retrieval system

Venezuela are to be retrieved, the query is rewritten to *all(dipsas,any(guianas,venezuela))*. To exclude the records on specimens of species *catesbyi* and *variegata* the query is written as *all(dipsas,-catesbyi,-variegata),any(guianas,venezuela))*.

Users can be taught this query format, but due to the availability of taxonomic resources, the system can also automatically translate basic query term enumerations such as *dipsas*, *-catesbyi*, *-variegata*, *guianas*, and *venezuela* into the desired complex query for the reptiles and amphibians. In order to do so, it 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 two or three capital letters and 3 to 6 numbers. After each term is classified, the module constructs the query according to rules that restrict possible combinations of types of terms.

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 the translation module to fail is, in all cases, due to a term not matching in the resource.

6.2.2 Query expansion

The query expansion modules in the system are aimed at increasing the recall by providing additional keywords or to remedy the influence of language variation on the retrieval of relevant results.

Taxonomic term expansion

As a consequence of changes in species classifications due to new insights, accepted taxonomic lists that describe the classification of a taxonomic class contain many synonyms and outdated names for each taxon. For example, if one wants to retrieve all snakes present in the collection, one could query for all records describing a specimen of suborder 'Serpentes', but this suborder is also known as 'Ophidiae'. An additional problem with this query is that as we noted earlier, 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 having to formulate a query that contains each of the 18 snake families with their possible synonyms, the system applies a knowledge-based query expansion approach that expands query terms with their taxonomic synonyms.

Geographic term expansion

Similar to the taxonomic term expansion, but slightly different in operation is the enrichment with a geographic resource. If we reconsider the example given in Subsection 6.2.1, *Are there Dipsas species other than D. catesbyi and D. variegata from the Guianas and Venezuela in the collection?*, we notice that *the Guianas* does not denote one country, instead it denotes Guyana (formerly British Guiana), Suriname (formerly Dutch Guiana) and French Guiana. Furthermore, for each of these names alternate spellings exist, and the fact that our database contains data in several languages may also impair relevant records from being retrieved. Fortunately, GeoNames contains many of the synonyms to automatically expand our query with.

Several flavours of a geographical expansion module were investigated, such as in addition to expanding to synonymous terms (for example in different languages), to expand to hypernyms or hyponyms, following the idea of (Voorhees, 1994). Hypernym expansion operates in such a way that if the query contains the term 'Nebraska', the query is expanded to 'United States of America', to remedy the negative influence of missing values in the 'province/state' column. Hyponym expansion works the other way around; a broad term such as 'United States of America' is expanded to all of its known hyponyms in the next level of the geographical ontology. Although hypernym and hyponym expansion are popular approaches that have been known to work for other systems (see Navigli & Velardi, 2003 for an overview) it does not aid object retrieval for the herpetological collection in these experiments. Therefore the geographical expansion was limited to expanding only to synonymous terms and location names in different languages.

6.2.3 Ranking

In order to present the user with the more relevant records first, two ranking methods were investigated.

RecordRank

RecordRank is a simplified version of the basic PageRank algorithm developed by the founders of Google in 1998 to rank results by relevancy (Brin & 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 retrieval results by some measure of authority is given by the hypothesis that researchers might pose more questions about the specimens or species NCB Naturalis is known for (e.g., the reptiles and amphibians collections contain many specimens from the Amazon, therefore researchers might ask more about that part of the collection than about specimens collected in Africa as there are fewer of those).

Authority in PageRank is measured by the number of incoming links to a page. Also, links from pages with a higher PageRank are considered more important than links from pages with a lower PageRank.

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 & Soroa, 2009). Similar to our aim, the PageRank algorithm has also been translated to a relational database setting in (Balmin et al., 2004). In this work, databases

are translated to modelled graphs in which objects are nodes and their semantic connections the edges. Although the database we used was originally a flat table, the domain ontology that was developed for the natural history domain can enrich the databases with the necessary structure to consider them as a relational data resource.
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). For every database record the scores of every value are added up resulting in a RecordRank score by which the database records can be ranked.
A drawback of RecordRank 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 by Haveliwala in 2002, who presents a topic-sensitive PageRank approach. 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). In Haveliwala (2002)'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 reptiles and amphibians database provides a smaller domain that cannot be easily broken up in more subdomains, the 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. We distinguish the two flavours of RecordRank as Global RecordRank, in which database records are ranked by authority regardless of the query, and Local RecordRank, in which database records are ranked after records are retrieved.

Column order by importance

Analysis of the queries has shown that queries do not usually pertain to information in some of the longer database columns 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. Xxxxxxxxx. 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 a large majority of the queries pertain to the request for information from the genus and species columns and never from the special remarks column in which one might find information on a specimen's stomach contents. Records with matches found in these columns, as well as in the registration number column are thus presented before records with matches found in other columns.

6.3 Experiments and results

In this section, the results of the experiments of the retrieval of records from the reptiles and amphibians database are presented. Only the first 5000 results returned for each query are evaluated using the evaluation script used in the Text REtrieval Conferences (TREC)[14]. In

[14] http://trec.nist.gov/ Last visited: 27 April 2011

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 ▲
COMPLEX	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 8. Precision, recall and mean average precision scores for baseline and expansion modules

each of the tables presented below, the bold face results are significant with respect to the baseline results that the module is compared to. All significance scores are computed at the p=0.05 level using a paired t-test. The ALL query mode denotes a simple keyword search in which only records should be retrieved in which all query terms match. The ANY query mode is another simple query mode in which records should be retrieved in which any of the query terms match. The interpreted query mode (as described in Subsection 6.2.1) is denoted by COMPLEX in the tables.

The precision, recall and mean average precision (MAP) for the interpretation and expansion modules are presented in Table 8. As the results in Table 8 show, the ALL query mode benefits more than the ANY query mode of the query expansion. This is due to the fact that the ANY query mode already achieves high recall, simply because it retrieves records in which at least one of the query terms match. Separately, the expansion modules perform best (denoted by TAXEXP for taxonomic expansion and GEOEXP for geographic expansion). When combined, and thus when they expand both the geographic and the taxonomic queries (TAXGEOEXP), the achieved results are mixed. For the ALL query mode, the precision does not deteriorate significantly (whereas it does for the separate expansion modules), but recall does not improve as much as expected, therefore this module is not further investigated. This is probably due to an explosion of expanded terms for each query term and the subsequent retrieval of too many records.

The experiments carried out with the query interpretation module are found in the lower part of Table 8. The precision and mean average precision scores for the interpreted query mode are significantly higher than for the simple query modes. On its own, the COMPLEX query mode improves the mean average precision with 5.83% over the ALL query mode, and with 7.59% for the ANY query mode. Together with the query expansion modules, the COMPLEX query mode helps improve the scores even more, in particular the geographic expansion module. The difference in recall between the unexpanded ALL query mode experiments and the geographically expanded COMPLEX query mode experiments is even more than 50% (from 31.67% to 85.85%). Also the ALL query mode benefits from query expansion.

In Table 9 the mean average precision scores for the ranking modules are presented. Our assumption that the RecordRank modules would aid performance because the more authoritative records are presented first proved wrong. For the unexpanded queries, the mean average precision improves, but not significantly. For the expanded queries, the RecordRank modules even harm performance.

ALL	UnExp	TaxExp	GeoExp
GlobalRecordRank	30.27 ▲	**23.81** ▽	**18.25** ▽
LocalRecordRank	30.24 ▲	**27.79** ▽	**19.51** ▽
GenSpec	30.40 ▲	39.77 ▽	41.68 ▽
Unranked	30.04	41.45	47.61
ANY	UnExp	TaxExp	GeoExp
GlobalRecordRank	29.47 ▲	**23.81** ▽	**18.98** ▽
LocalRecordRank	29.17 ▲	23.42 ▽	**19.51** ▽
GenSpec	**42.38** ▲	**39.89** ▲	**39.86** ▲
Unranked	28.28	28.87	28.87
COMPLEX	UnExp	TaxExp	GeoExp
GlobalRecordRank	36.15 ▲	**23.83** ▽	**18.25** ▽
LocalRecordRank	36.11 ▲	**27.80** ▽	**19.49** ▽
GenSpec	36.23 ▲	39.75 ▽	**41.60** ▽
Unranked	35.87	44.29	51.61

Table 9. Mean average precision results expanded ranked reptile and amphibian queries

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

Table 10. Number of reptiles and amphibians queries for which one or more relevant results are retrieved

Due to the precise manner of querying provided by the COMPLEX query mode and the limitations imposed by the ALL query mode, result ranking only significantly aids the ANY query mode.
The GENSPEC module, that ranks records in which a match is found in the genus and species columns higher than the records in which a match is found in other columns does significantly improve results for the ALL and ANY query modes. For the interpreted query mode, results were already better and thus the ranking does not significantly aid performance.
If we look at the results in Tables 10, we see that, even though the precision drops when query expansion is used, the number of queries for which at least one relevant record is retrieved more than doubles. Thereby, it must also be noted that there are 16 queries for the reptiles and amphibians, for which there are no relevant records present in the databases. This means that for only six queries for which a relevant record should have been retrieved remain unanswered.

7. Conclusions

In this contribution, we first presented a new approach to collection digitisation, and a novel collection registration system (CRS) as implemented at NCB Naturalis. The new CRS enables the researcher to search in unfiltered collection unit metadata, allowing for new interpretations. Previously, searching in collection databases produced filtered, interpreted data, further constrained by the fact that databases only covered specialised sub-collections, making it hard for the expert, and impossible for the non-expert, to assess the value of the search results. The new system, operating on enriched data, promises to not only aid

the expert better, but also provide means to visualise search results in ways suitable to the layperson, such as plotting findings on maps and timelines.
Our contribution then focused on two systems aimed at improving the accessibility of data in the CRS; the systems are semi-automatic, in the sense that they perform automated steps in the process of data cleaning and data retrieval, with the aim of supporting experts by saving time (as manual cleaning of all data is simply infeasible) and finding relevant information faster.
The computer-supported data cleaning system presented uses logical rules to detect clear violations of constraints in pairs of dates (a collection of a specimen always precedes all other actions), in geographical names (a Brazilian city needs to be located in Brazil), and in taxonomic names (a species name has to fit a path in the taxonomic tree). The fact that 'hard' domain constraints are used does not constrain the applicability of the system (although all kinds of variations and changes through time can cause certain relations be softer than the rule-based method assumes).
The knowledge-driven data retrieval system indeed boosted the usability of digital information considerably. We observed significantly better retrieval of specimen cases from the CRS when the queries were automatically improved, either by expansion of taxonomic or geographical names (e.g., by their synonyms), or by guiding the matching function to match on particular database fields rather than all fields. Authority-based re-ranking as used in web search engines did not prove to be useful, indicating that the collection database, when viewed as a graph (a relational database) does not have the typical small-world network properties with 'authority' nodes that the web has.

8. Acknowledgements

This chapter is partly based on (Van Erp, 2010; Van Erp & Hunt, 2010). The research described in this contribution was funded by the Netherlands Organisation for Scientific Research under the CATCH programme, as part of the MITCH project. The authors wish to thank Pim Arntzen, Erik van Nieukerken and Ronald de Ruiter for their expert feedback.

9. References

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

Balmin, A., Hristidis, V. & 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 & K. B. Schiefer (eds), *Proceedings of the Thirtieth International Conference on Very Large Data Bases (VLDB 2004)*, Morgan Kaufmann, Toronto, Canada, pp. 564–575.

Berendsohn, W. G., Chavan, V. & Macklin, J. A. (2010). Recommendations of the GBIF task group on the global strategy and action plan for the mobilization of natural history collections data, *Biodiversity Informatics* 7: 67–71.

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

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

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

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.

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. & 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.

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

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

Johnson, N. F. (2007). Biodiversity informatics, *Annual Review of Entomology* 52(2): 421–438.

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

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

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

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

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

Raes, N. (2009). *Borneo: A quantitative analysis of botanical richness, endemicity and floristic regions based on herbarium records*, PhD thesis, Leiden University.

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

Scoble, M. (2010). Natural history collections digitization: Rationale and value, *Biodiversity Informatics* 7: 77–80.

Soberón, J. & 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.

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

Van Erp, M. (2010). *Accessing Natural History: Discoveries in Data Cleaning, Structuring, and Retrieval*, PhD thesis, Tilburg University.

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

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

Changing Diversity in Changing Environment
Edited by PhD. Oscar Grillo

ISBN 978-953-307-796-3
Hard cover, 392 pages
Publisher InTech
Published online 14, November, 2011
Published in print edition November, 2011

As everybody knows, the dynamic interactions between biotic and abiotic factors, as well as the anthropic ones, considerably affect global climate changes and consequently biology, ecology and distribution of life forms of our planet. These important natural events affect all ecosystems, causing important changes on biodiversity. Systematic and phylogenetic studies, biogeographic distribution analysis and evaluations of diversity richness are focal topics of this book written by international experts, some even considering economical effects and future perspectives on the managing and conservation plans.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Marieke van Erp, Antal van den Bosch, Steve Hunt, Marian van der Meij, René Dekker and Piroska Lendvai (2011). Natural Selection: Finding Specimens in a Natural History Collection, Changing Diversity in Changing Environment, PhD. Oscar Grillo (Ed.), ISBN: 978-953-307-796-3, InTech, Available from: http://www.intechopen.com/books/changing-diversity-in-changing-environment/natural-selection-finding-specimens-in-a-natural-history-collection

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821