Biodiversity knowledge extraction techniques (BioKET)

Somsack Inthasone

To cite this version:

Somsack Inthasone. Biodiversity knowledge extraction techniques (BioKET). Other [cs.OH]. Université Nice Sophia Antipolis, 2015. English. NNT : 2015NICE4013 . tel-01166027

HAL Id: tel-01166027

https://theses.hal.science/tel-01166027

Submitted on 22 Jun 2015

UNIVERSITÉ DE NICE - SOPHIA ANTIPOLIS

ÉCOLE DOCTORALE STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION
ET DE LA COMMUNICATION

THÈSE

pour obtenir le titre de

Docteur en Sciences

de l'Université de Nice - Sophia Antipolis

Mention : Computer Science

Présentée et soutenue par

Somsack INTHASONE

Biodiversity Knowledge Extraction Techniques (BioKET)

Thése dirigée par

Nicolas PASQUIER et Andrea G. B. TETTAMANZI

soutenue le 2 avril 2015

Jury :

Rapporteurs :	Dario MALCHIODI	-	University of Milan, Italy
	Engelbert MEPHU NGUIFO	-	Blaise Pascal University, France
	Patrick COQUILARD	-	INRA-PACA, UNS CNRS
Examinateurs :	Frederic PRECIOSO	-	I3S, UNS CNRS
Directeurs :	Nicolas PASQUIER	-	I3S, UNS CNRS
	Andrea G. B. TETTAMANZI	-	I3S, UNS CNRS
Invité :	Célia DA COSTA PEREIRA	-	I3S, UNS CNRS

ACKNOWLEDGMENTS

First of all, I would like to take this opportunity to express my deepest gratitude to the adviser and co-advisers, Nicolas PASQUIER, Andrea G. B. TETTAMANZI and Célia DA COSTA PEREIRA, with their great advice and contributions were able to conduct this research succeeded.

I am especially grateful to Prof. Dario MALCHIODI, Engelbert MEPHU NGUIFO and Patrick COQUILARD who spent the time to review my work.

I would like to acknowledge appreciatively the help and supports of the following people during my PhD research:

⊙ Prof. Marc and Francine DIENER, the EMMA coordinator and host supervisors.

⊙ Lao botanists/home supervisors: Somchanh BOUNPHANMY and Vichith LAMXAY.

⊙ Regine SAELENS the secretary of Ecole Doctorale STIC and Prof. Bernot GILLES the I3S director.

⊙ MinD team leader and secretary: Prof. Frederic PRECIOSO and Magali RICHIR.

⊙ My family for all their encouragement.

Special thanks to:

⊙ Erasmus Mundus Mobility with Asia (EMMA) Programme.

⊙ The Institute of Traditional Medicine at Ministry of Public Health, Faculty of Natural Sciences and Faculty of Forestry at National University of Laos.

⊙ Université de Nice - Sophia Antipolis, Ecole Doctorale STIC, Laboratoire d'Informatique, Signaux et Systèmes de Sophia-Antipolis (I3S) - UMR7271 - UNS CNRS.

Somsack INTHASONE
Université de Nice - Sophia Antipolis
April 2, 2015

Contents

List of Figures

List of Tables

Listings

Techniques d'Extraction de Connaissances en Biodiversité

Résumé : Les données sur la biodiversité sont généralement représentées et stockées dans différents formats. Cela rend difficile pour les biologistes leur agrégation et leur intégration afin d'identifier et découvrir des connaissances pertinentes dans le but, par exemple, de classer efficacement des spécimens. Nous présentons ici l'entrepôt de données BioKET issu de la consolidation de données hétérogènes de différentes sources. Actuellement, le champ d'application de BioKET concerne la botanique. Sa construction a nécessité, notamment, d'identifier et analyser les ontologies et bases botaniques existantes afin de standardiser et lier les descripteurs utilisés dans BioKET. Nous avons également développé une méthodologie pour la construction de terminologies taxonomiques, ou thésaurus, à partir d'ontologies de plantes et d'informations géo-spatiales faisant autorité. Les données de biodiversité et botanique de quatre fournisseurs majeurs et de deux systèmes d'informations géo-spatiales ont été intégrées dans BioKET. L'utilité d'un tel entrepôt de données a été démontrée par l'application de méthodes d'extraction de modèles de connaissances, basées sur les approches classiques Apriori et de la fermeture de Galois, à des ensembles de données générées à partir de BioKET. En utilisant ces méthodes, des règles d'association et des clusters conceptuels ont été extraits pour l'analyse des statuts de risque de plantes endémiques au Laos et en Asie du Sud-Est. En outre, BioKET est interfacé avec d'autres applications et ressources, tel que l'outil GeoCAT pour l'évaluation géo-spatiale des facteurs de risques, afin de fournir un outil d'analyse performant pour les données de biodiversité.

Mots clés : Biodiversity, Data Mining, Knowledge Integration, Data Warehouse, Information Technology, Ontologies.

Biodiversity Knowledge Extraction Techniques (BioKET)

Abstract: Biodiversity data are generally stored in different formats. This makes it difficult for biologists to combine and integrate them in order to retrieve useful information and discover novel knowledge for the purpose of, for example, efficiently classifying specimens. In this work, we present the BioKET data warehouse which is a consolidation of heterogeneous data stored in different formats and originating from different sources. For the time being, the scope of BioKET is botanical. Its construction required, among others things, to identify and analyze existing botanical ontologies, to standardize and relate terms in BioKET. We also developed a methodology for mapping and defining taxonomic terminologies, that are controlled vocabularies with hierarchical structures from authoritative plant ontologies, Google Maps, and OpenStreetMap geospatial information system. Data from four major biodiversity and botanical data providers and from the two previously mentioned geospatial information systems were then integrated in BioKET. The usefulness of such a data warehouse was demonstrated by applying classical knowledge pattern extraction methods, based on the classical Apriori and Galois closure based approaches, to several datasets generated from BioKET extracts. Using these methods, association rules and conceptual bi-clusters were extracted to analyze the risk status of plants endemic to Laos and Southeast Asia. Besides, BioKET is interfaced with other applications and resources, like the GeoCAT Geospatial Conservation Assessment Tool, to provide a powerful analysis tool for biodiversity data.

Keywords: Biodiversity, Data Mining, Knowledge Integration, Data Warehouse, Information Technology, Ontologies.

Chapter 1

Introduction

This chapter addresses problems statement, objectives, contributions and outlines of entire research work.

Contents

1.1 Problem and Motivation

Biological diversity, or biodiversity, refers to the natural variety and diversity of living organisms [2013q]. Biodiversity is assessed by considering the diversity of ecosystems, species, populations and genes in their geographical locations and their evolution over time. Biodiversity is of paramount importance for a healthy environment and society, as it ensures the availability of natural resources and the sustainability of ecosystems [Eldredge 2002, Grillo 2011, MA 2005, Midgley 2012, Shah 2011, Talent 2012]. The effects of biodiversity loss on the environment, caused by habitat loss and fragmentation, pollution, climate change, invasive alien species, human population, and over-exploitation can affect all life forms and lead to serious consequences [2013k]. Understanding biodiversity is an essential prerequisite for sustainable development.

> The issues on ASEAN Important Plant Area (IPA) Meeting in Hanoi, Vietnam: How to distribute information of threatened plant to community and how to encourage local people to join plant protection.
>
> - To list important plant protection areas in each country.
> - To list which species are threatened, rare and etc by IUCN category plant species in each IPA.
> - To find the most appropriate approach to protect plant species.

For many years, biodiversity datasets have been stored in different formats, ranging from highly structured (databases) to plain text files, containing plant descriptions (vocabularies and terms). Numerous data and knowledge repositories containing biodiversity and environmental information are available on the Internet as on-line and off-line resources nowadays. Data repositories store large amounts of information depicting facts on concrete objects related to a specific domain of application, e.g., results of environmental studies or inventories of species in a geographic location. This makes it difficult for botanists or zoologists to combine and integrate them to retrieve useful information for the purpose of identifying and describing new species.

The ever increasing availability of data relevant to biodiversity makes the idea of applying data mining techniques to the study of biodiversity tempting [Hochachka 2007].

Data mining, also known as knowledge discovery from data (KDD), is a set of concepts, methods and tools for the rapid and efficient discovery of previously unknown information, represented as knowledge patterns and models, hidden inside massive information repositories [Han 2011].

One important obstacle to the application of data mining techniques to the study of biodiversity is that the data that might be used to this aim are somewhat scattered and heterogeneous [Spehn 2009]. Different datasets cover different aspects of the problem or focus on some geographical areas only. None of them is complete and there is no standard format.

To overcome these limitations, we have designed and implemented BioKET, a data warehouse whose purpose is to consolidate heterogeneous data sources on biodiversity in a logically organized, coherent, and comprehensive resource that can be used by the scientific community as a basis for data-intensive studies.

1.2 Objective and Scope

The main aim of this research is to help botanists in their studies on plants and their works on plant protection and conservation. These tasks involve the study of plant structure, growth, developement, biochemistry, diseases, evolutionary relationships, differenciation and taxonomy (Biological classification). One important step regarding these objectives is the analysis of morphological and environmental properties of plants. For this, the present work was divided in the following main steps.

- Define an ontology mapping including methodology and process for the integration and normalization of plant data.

- Construct BioKET Data Warehouse integrating (data and knowledge) from the different resource repositories on plants and their conservation status. The four following data repositories were integrated in BioKET: BIOTIK [2011b] (Western Ghats of India and National University of Laos), BRAHMS repository [2013e] (National University of Laos, Faculty of Forestry), NAPIS repository [2013n] (Lao Ministry of Public Health, Institute of Traditional Medicine) and the IUCN Red List Data [2014s]. All sources are linked to an institution, except for the last one.

- Interface BioKET geolocation information on plant specimens with standard geospatial information systems. For this step, Google geospatial data (Google Maps Geocoding Service) were integrated in BioKET and an interface with the GeoCAT visualization platform was developed.

- Generate a BioKET dataset from the BioKET data warehouse and perform an analysis validation with classical knowledge pattern extraction methods based on the Apriori and Galois closure approaches for association rule extraction and biclustering.

1.3 Contribution

The main contribution is to provide details of how the BioKET data warehouse has been designed and populated, by consolidating and integrating multiple and heterogeneous sources of data. The reader should not underestimate the methodological challenges and the practical problems that had to be overcome in order to achieve that result. As all data mining practitioners agree, pre-processing, which includes data cleaning, integration, and transformation is the most time-consuming and critical phase of the data mining process [Marbán 2009, Mariscal 2010] illustrated in Figure 1.1.

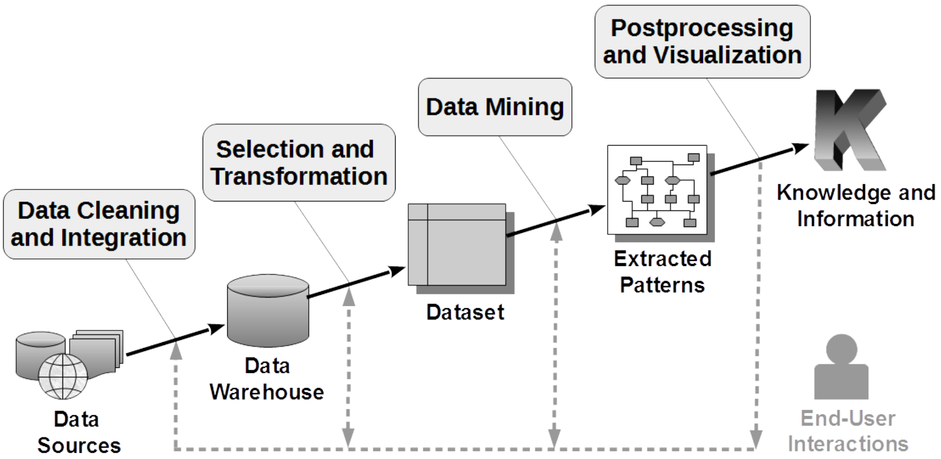

Postprocessing and Visualization
Data Mining
Selection and Transformation
Data Cleaning and Integration
Knowledge and Information
Extracted Patterns
Dataset
Data Warehouse
Data Sources
End-User Interactions

Figure 1.1: Main phases of a data mining process

We demonstrate the use of such resource by applying TFIST and Weka, a combined biclustering and conceptual association rule extraction method already described in the literature [Mondal 2012], on a dataset extracted from the BioKET data warehouse in order to analyze the risk status of plants endemic to Laos.

The outcomes of this research will be a guideline for Lao biologists/botanists and a knowledge base for Lao government's country development plans as well as Lao PDR and countries in Southeast Asia.

1.4 Structure of thesis

The thesis consists of five chapters. Chapter 1: Introduction presents from the points of problem and motivation, objective and scope, contribution, and structure of thesis. Chapter 2: Background and Related Work presents definitions, issues, challenges, resources in biodiversity fieldwork, literature on state-of-the art technologies and applications. Chapter 3: BioKET Data Warehouse states in depth how to build, design, and implement the BioKET Data Warehouse, and how to integrate and visualize the BioKET data with geolocation data (geospatial data), and how to prepare patterns for generating the BioKET dataset and presenting the BioKET plant ontology mapping concept. Chapter 4: Experiments and Results presents the BioKET dataset schema, patterns extraction, patterns evaluation, and experimental results. Chapter 5: Conclusions and Further Work summarizes what has been done during this research work and presents perspectives from the viewpoint of the extension of this work and related applications.

Chapter 2

Background and Related Work

This chapter is stated as follows. Getting started from biodiversity definitions, issues, challenges, resources, and literature on state-of-the art technologies and applications.

Contents

2.1 Biodiversity

There is a general agreement that biodiversity consists of three main types, or levels, of diversity as depicted in Figure 2.1. The first type is *Genetic diversity*, the second one is *Species diversity*, and the last one is *Ecosystem diversity*, or Ecological diversity [2013u, Groombridge 2002, Popy 2009, Sala 2003, Swingland 2001, Talent 2012, 2013v]. These are the three levels at which biological variety has been identified.

Genetic diversity refers to the genetic variation and heritable traits within organisms. All species are related with other species through a genetic network, but the variety of genetic properties and features makes creatures different in their morphologic characteristics. Genetic diversity applies to all living organisms having inheritance of genes, including the amount of DNA per cell and chromosome structures. Genetic diversity is an important factor for the adaptation of populations to changing environments and the resistance to certain types of diseases. For species, a higher genetic variation implies less risk. It is also essential for species evolution.

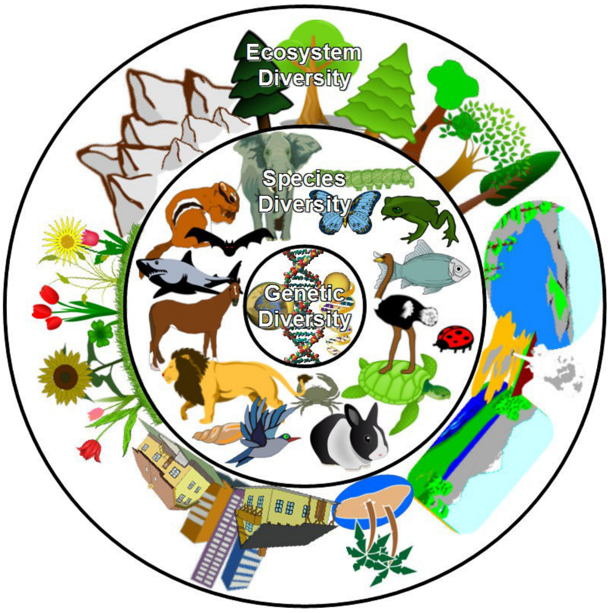


Figure 2.1: Types of diversity: Genetic (inner), Species (middle), and Ecosystem (outer)

Species diversity refers to the variety of living organisms within an ecosystem, an habitat or a region. It is evaluated by considering two factors: species richness and species evenness. The first corresponds to the number of different species present in a community per unit area, and the second to the relative abundance of each species in the geographical area. Both factors are evaluated according to the size of populations or biomass of each species in the area. Recent studies have shown relationships between diversity within species and diversity among species. Species diversity is the most visible part of biodiversity.

Ecosystem diversity refers to the variety of landscape of ecosystems in each region of the world. An ecosystem is a combination of communities – associations of species – of living organisms with the physical environment in which they live (e.g., air, water, mineral soil, topography, climate). Ecosystems vary in size and in every geographic region there is a complex mosaic of interconnected ecosystems. Ecosystems are environments with a balanced state of natural elements (water, plants, animals, fungi, microbes, molecules, climate, etc.). Ecosystem diversity embraces the variety of habitats and environmental parameters that occur within a region. To preserve biodiversity, the conservation and protection of a representative array of interacting ecosystems, and their associated genetic and species diversities, is decisive.

Biodiversity applies wherever there is life, that is all around the world, from the earth's surface to marine ecosystems. Biologists most often define biodiversity as the "totality of genes, species, and ecosystems of a region". Ecologists consider biodiversity according to the three following interdependent primary characteristics: Ecosystems composition, i.e., the variety and richness of inhabiting species, ecosystems structure, i.e., the physical and three dimensional patterns of life forms, and ecosystems function, i.e., biogeochemical cycles and evolving environmental conditions. Even though many application tools have been developed, evaluating biodiversity still faces difficulties due to the complexity of precise evaluations of these parameters. Hence, the overall number of species that can be measured and officially identified all around the world is only 1.7 to 2 millions and 5 to 30 millions respectively [2013w, Magurran 2013, Magurran 2011].

2.1.1 Environmental Issues

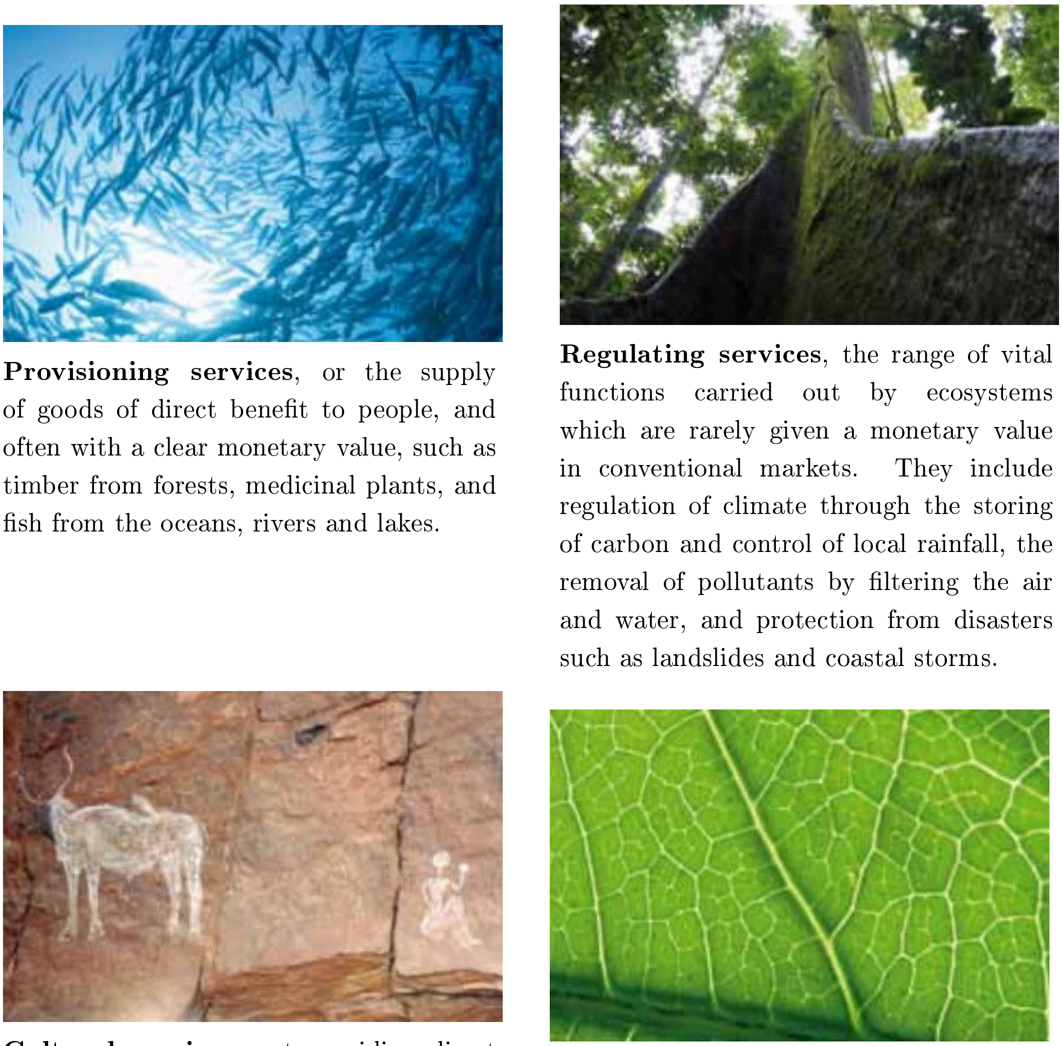
In nature, biodiversity is the key to keep natural balance in changing environmental conditions. It functions as services, such as consumption service, that is to serve the natural resources to human (e.g., food, clothing, housing, medicines), industrial production service, that is to serve productivity of forest to be used either directly or indirectly (e.g., extracting chemicals from plants in the forest), and others (non-consumptive uses) including values of maintenance of ecosystems to be sustainable (e.g., soil maintenance, nitrogen to the soil, synthesis of plant power, humidity control) [Eldredge 2002, Grillo 2011, Shah 2011, Talent 2012].

All life on the planet needs nutrients and oxygen, which are the main factors for survival. Especially, species depend on biodiversity resources produced by ecosystem services. The ecosystem services (Figure 2.2) can regulate climate changes, dispose of wastes, recycle nutrients, filter and purify water, purify air, buffer against flooding, and maintain soil fertility [Eldredge 2002, MA 2005, Midgley 2012, Talent 2012]. Changes in environmental factors and ecosystems can thus endanger life forms as reported in several scientific studies (Figure 2.3).

The report of Global Biodiversity Outlook 3 [2013k] of the Convention on Biological Diversity (CBD) highlights Ban Ki-moon's speech, United Nations General Secretary, on the fact that "the consequences of this collective failure, if it is not quickly corrected, will be severe for us. Biodiversity underpins the functioning of the ecosystems on which we depend for food and fresh water, health and recreation, and protection from natural disasters. Its loss also affects us culturally and spiritually. This may be more difficult to quantify, but is nonetheless integral to our well-being". The loss of biodiversity becomes a serious issue for the twenty-first century. Its loss has direct and indirect negative effects on many factors (Figure 2.4) connecting the elements of biodiversity as well as ecosystems [Cardinale 2012, Midgley 2012].

In *Environmental factors*, the loss of biodiversity means that the natural balance between environmental conditions and the different types of diversities cannot be conserved, which will affect stability of ecosystems. This can lead to climate changes, such as the global warming reported in several scientific studies, and consequently to natural disasters (landslides, floods, typhoons, cyclones, hurricanes, tsunamis, etc.) [Cardinale 2012, EBI 2013, Khallaf 2011, Midgley 2012, 2013g].

Tourism factors are impacted by environment factors: If they are affected, by natural disasters or pollution for instance, tourism structure systems, such as aesthetic natural landscapes and historical places, can be affected, and even destroyed. For example, the effect of pollution on the structure and environment of



Provisioning services, or the supply of goods of direct benefit to people, and often with a clear monetary value, such as timber from forests, medicinal plants, and fish from the oceans, rivers and lakes.

Regulating services, the range of vital functions carried out by ecosystems which are rarely given a monetary value in conventional markets. They include regulation of climate through the storing of carbon and control of local rainfall, the removal of pollutants by filtering the air and water, and protection from disasters such as landslides and coastal storms.

Cultural services, not providing direct material benefits, but contributing to wider needs and desires of society, and therefore to people's willingness to pay for conservation. They include the spiritual value attached to particular ecosystems such as sacred groves, and the aesthetic beauty of landscapes or coastal formations that attract tourists.

Supporting services, not of direct benefit to people but essential to the functioning of ecosystems and therefore indirectly responsible for all other services. Examples are the formation of soils and the processes of plant growth.

Figure 2.2: A brief description of ecosystem services [2013k]

the Venice city, in Italy, is known to endanger buildings [del Monte 1985, Pipe 1995, Zannetti 1977]. This can impact tourism as natural landscapes and historical places are attraction sites for tourists, which consequently contribute to develop economical and social activities [Cardinale 2012, Khallaf 2011].

Human factors are linked to nutrients, oxygen, and other essential needs which

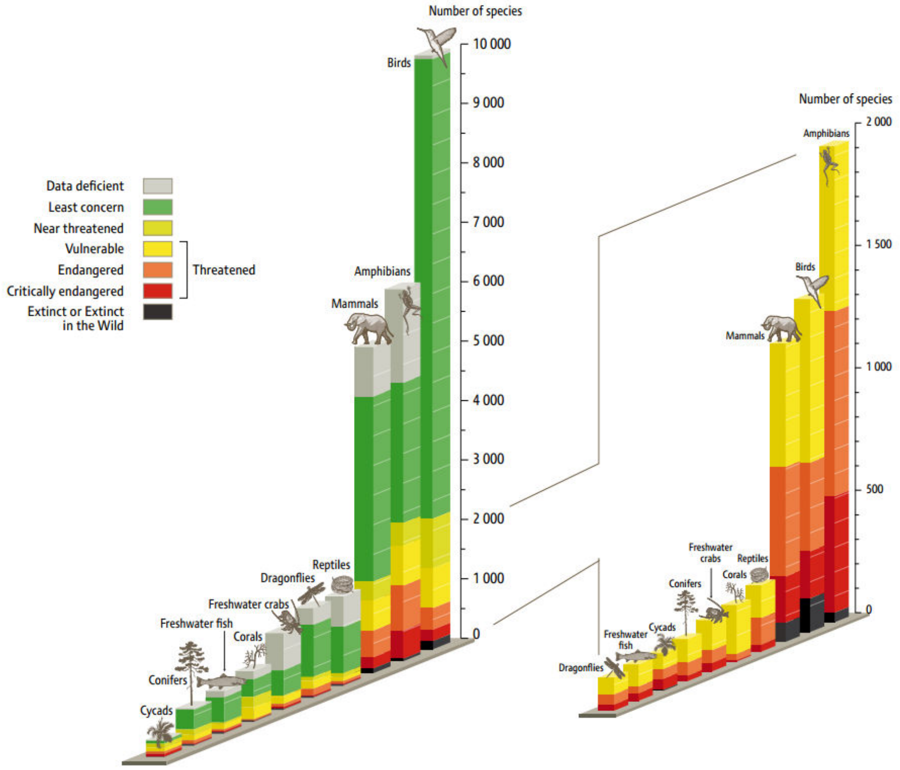


Figure 2.3: The risk status of species in each taxonomic group [2013k]

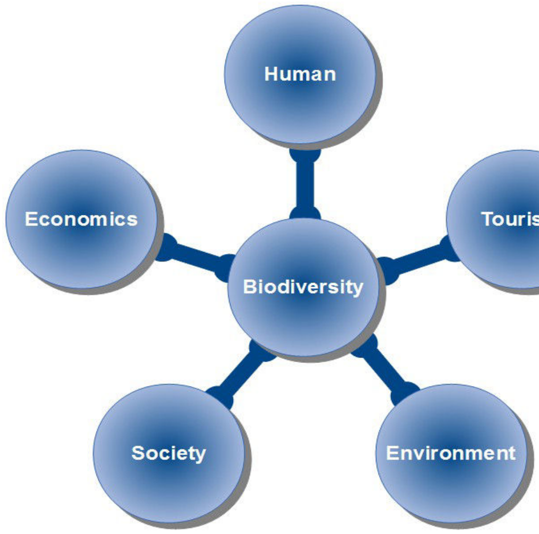


Figure 2.4: Relationships among biodiversity, human, society, environment, economics and tourism

are produced from biodiversity resources. If the number of biodiversity resources is decreased, the volume of vital products, such as food, water, plants and animals, will also decrease. This can lead to human's consumption and survivability concerns. For example, the augmentation of the production costs can lead to difficulties for human populations to have access to vital resources, such as medicine, food, etc. [Blanco 2011, Cardinale 2012, Casalegno 2011, MA 2005, Khallaf 2011, COHAB 2010, 2013g].

Society factors are affected by biodiversity loss as most parts of society infrastructures and livelihood depend on the basic system and structure of nature. One factor is nature productivity, that depends on land structures for agriculture and irrigation, wood materials for building habitats, and natural and energy materials for other forms of consumption. For example, an important part of people living in rural societies depend mainly on productivity of agriculture and livestock for their livelihood, while people living in metropolitan areas need more biodiversity productivity as the demands for food, energy, materials and other resources are increased (transportation, construction, consumption products, etc.). Scarceness of resources can thus cause an augmentation in production costs leading to a reduction of the part of the population that has access to these resources [Blanco 2011, Cardinale 2012, EBI 2013, 2013p, Khallaf 2011].

Economics factors are impacted by direct benefits and added-values of natural resources (e.g., food, bio-fuels and renewable energies, animals and fibers, wood materials, bio-medical treatments). These resources contribute to the economic exchanges between countries around the world through internal and external commerces. Biodiversity loss, and scarceness of resources, can affect populations from an economical viewpoint. For instance, the important human population (more than 60 percent) that use bio-medication for main health cares [Gaston 2004]. It can also lead to higher production costs, implying more competition, financial crisis, and others economic related issues [Cardinale 2012, Casalegno 2011, 2013r, Karahalil 2005, Nijkamp 2008, Perrings 2010].

2.1.2 Topics and Challenges

Biodiversity loss is a major problem that bioscientists must take into account, considering and analyzing each parameter of loss. We describe below six main categories of causes and effects on biodiversity loss, as well as their impact on ecosystems and ecosystem services.

Habitat Loss and Fragmentation Habitat Loss are affected by many factors, for example, deforestation for agriculture, sawed timber, factories, etc. [Karahalil 2005]. According to the *Global forest land-use change 1990-2005* report by Food and Agriculture Organization (FAO) [Lindquist 2012], the percentage of decrease in global forest areas is 1.7 percent in 30 percent (3.8 billion hectares) of overall forest areas around the world between year 1990 and 2005. This decrease is due to deforestation for agriculture, land uses, and other purposes. In [2013k], the authors used data and knowledge on percent of deforestation to address and warn about the problem of habitat loss. In addition, Habitat loss can be caused by natural disasters such as flooding, earthquake, landslide, and so on. However, habitats remaining from destruction are fragmented to small parts and resulting fragments are not enough wide for local organisms to live and migrate within, and among, other organisms [Lameed 2012, 2013k, Thornton 2011].

Pollution Pollution of the air, land and water is caused mainly by human and natural factors, such as manufacture, transportation, construction, burned forest, electric power generation, and nuclear power generation [Khallaf 2011, 2013x, 2013k]. This cause a risk of poisoning for all living organisms both on land and in the water on the planet. In addition, vehicle emissions, industrial emissions, and drainage of waste are factors that can increase carbon dioxide in the atmosphere (Figure 2.5) and directly affect ecosystems. This can lead to climate changes, as well as global warming. According to *Global Health Observatory* report by the *World Health Organization* (WHO), the number of deaths by air pollution is about 4.6 million people in each year [2013a], and the worldwide percent of deaths by lung cancer about 9%, 5% of cardiopulmonary, and 1% of respiratory infection [Casalegno 2011, Silva 2013, 2013l].

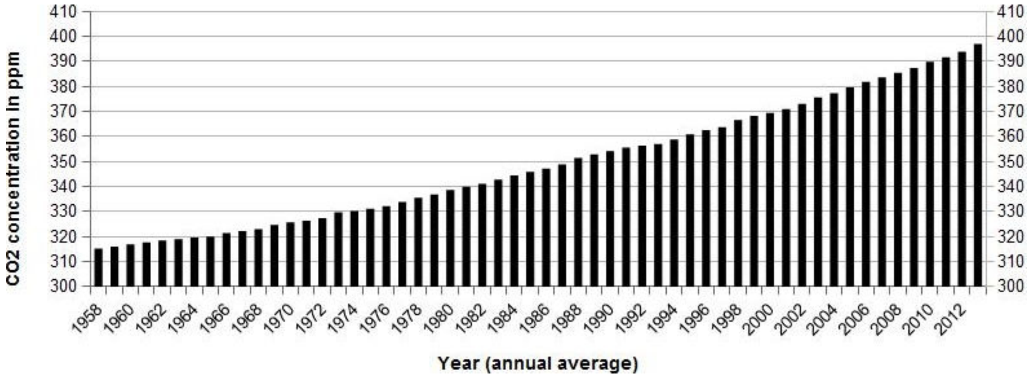


Figure 2.5: Atmospheric CO_2 concentration from March 1958 to September 2013 [2013h]

Climate Change The *CO_2Now Organization* provides a collection of global climate data from scientists around the world, showing the status of the global change [2013h]. These data show that the climate changes frequently occur and have different impacts on many aspects, especially regarding biodiversity, in each different zone of the world. For example, in the context of biodiversity in the Arctic zone the polar bears live on sea ice and other species living under sea ice are affected due to the elevation of temperatures in high latitudes [2013k, Staudinger 2012]. On the other hand, the climate change is impacting life cycles of humans and other species on the earth, because of the more uncomfortable and unstable environment [Silva 2013]. For example, the migration and adaption of human and other species to new locations due to frequent changes in environmental conditions [Staudinger 2012]. However, the effect of climate change might give rise to abundance and distribution of individual species around ecosystems such as crops grow, breeding stock, the tides of the sea, etc. [Blanco 2011, 2013k, Perrings 2010, 2013g].

Invasive Alien Species This is the actual most important risk for biodiversity loss globally. Invasive Alien Species, whether present deliberately or coincidentally, can create intense issues in the biological ecosystem that can lead to the disappearance of numerous species and to difficulties to survive for other local species. The report of *Global Biodiversity Outlook 3* report of the *Convention on Biological Diversity* (CBD) [2013q, 2013k] depicts for each different category of species the different portions that are at distinct extinction risk levels (Figure 2.3). It is shown that the among the 10,000 species listed, 2,000 species are in the risk or extinction zones. In [2013k], is also reported the sample data of alien species of 57 countries, where have been found more than 542 alien species. In addition, this issue will cause enormous investment expenses for farming, ranger services, fishery and other related human activities [2013k, 2013m].

Human Populations The human population has a growing factor at an exponential rate and, according to the *United Nation report on World Population to 2300* [2004y], the size of the human population will increase from a growth rate of 2.3 to a growth rate of 36.4 billion as shown in Figure 2.6. This will increase the consumption and may thus cause natural resources, as well as ecosystem services, to be insufficient. In order to preserve human life on the planet, the *Food and Agriculture Organization* (FAO) estimates and predicts data on the consumption of human population, and promotes the consumption of edible insects [Van Huis 2013].

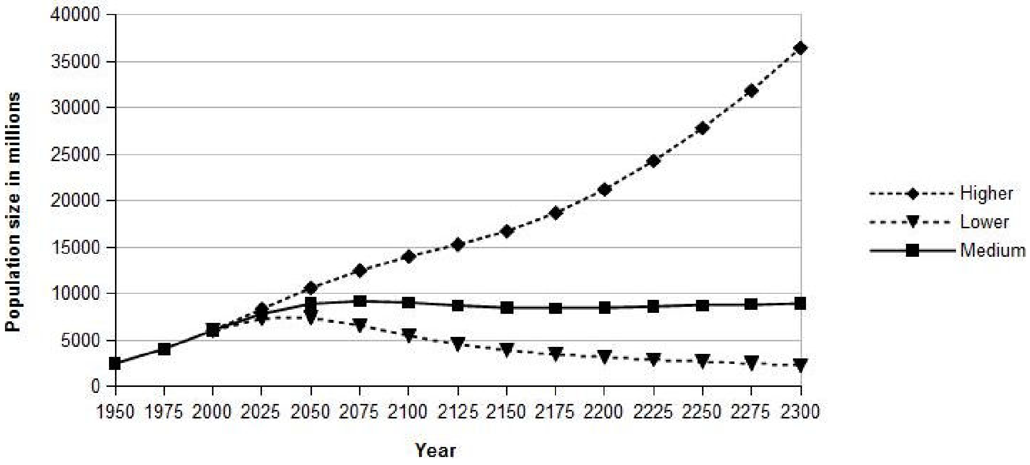


Figure 2.6: World human population projections according to three scenarios (lower, medium and higher growth rates) for period 1950-2300 [2004y]

Overexploitation Overexploitation, in term of humans use of natural resources, means an over-consumption of ecosystem services by humans (e.g., in fisheries, hunting, and industries). This can lead to the destruction of the volume of natural resources, for example, the volume of fish stock decreased by fishery's overexploitation. According to the Food and Agriculture Organization (FAO) report, among the overall 600 marine fish stocks worldwide, 17 percent are overexploited, 52 percent are fully exploited, 20 percent are moderately exploited, 3 percent are underexploited, 7 percent are depleted, and 1 percent are recovering from depletion [2013i]. Despite this small overexploitation percent (17%), this is an issue as natural resources and ecosystem services have limits to serve humans; the actual rate of resources consumption and the absence of natural resource protection causes risks for different species [Grillo 2011, Van Huis 2013].

2.2 Resources and Technologies for Biodiversity

This section is devoted to the presentation of the major information providers and different resources available, and the technologies used to represent data and knowledge related to biodiversity and environment.

2.2.1 Resources

The most prominent information providers that propose data and knowledge repositories used for biodiversity and environmental studies are shown in Figure 2.7. Each one provides contents related to different domains and categories depicted by the edges in the schema. See Appendix A.1 for a description of each of these seventeen information providers. These providers propose contents of different types (documents, databases, meta-data, spatio-temporal, etc.), in different categories (data, knowledge and documentations) and for different domains of application. Two main categories of resources are considered in this classification diagram. The *Data* category corresponds to resources depicting facts about species (animal and plants) and environmental conditions in specific areas. The *Frameworks* category corresponds to resources depicting both tacit and formal knowledge related to biodiversity and environment analytical application domains. Several providers, such as ESABII (East and Southeast Asia Biodiversity Information Initiative), IUCN (International Union for Conservation of Nature), and OECD (Organisation for Economic Co-operation and Development), give access to both data and frameworks resources.

In the *Biodiversity Policy* knowledge domain, information concern issues of principles, regulations and agreements on biodiversity. Amid these resources, we can cite BioNET, CBD, ESABII, IUCN and UNEP that provide information to serve and follow up among botanists, biologists and researchers in "globalization". For example, the *Agreement of the United Nations Decade on Biodiversity 2011-2020* aims to support and implement the Strategic Plan for Biodiversity[1].

The *Environment* domain refers to results of researches and repositories on this area to scientists or people who want to know status of environment on the earth,

[1] http://www.cbd.int/2011-2020

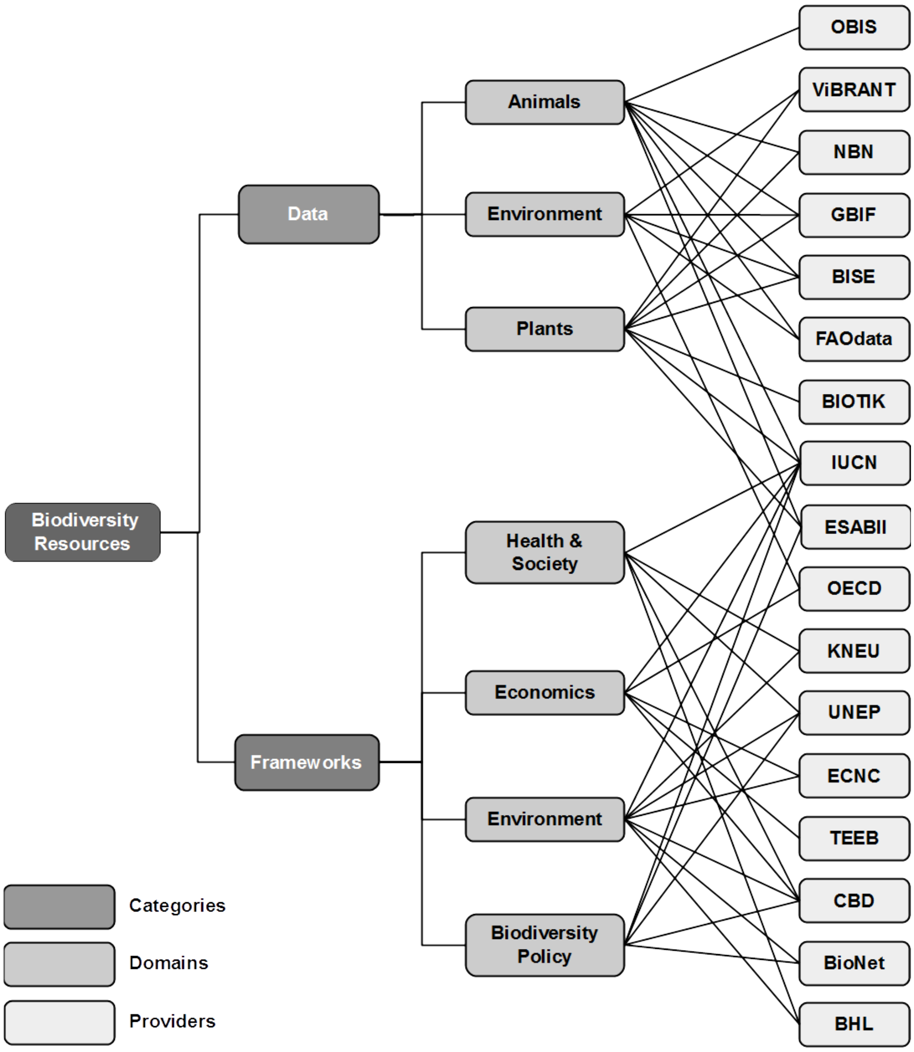


Figure 2.7: A classification diagram of biodiversity and environmental information

especially the biologist who works on this issue. The major actors in this category are BHL, BioNET, CBD, ECNC, IUCN, KNEU and UNEP that are organizations which regularly publish reports and results of studies on domains related to biodiversity and environment, such as the *Protected Areas Presented as Natural Solutions to Global Environmental Challenges at RIO +20* published by the IUCN or the *Global Environment Facility* (GEF) published by the CDB organization.

The *Economics* domain corresponds to information from scientists about the status of economics development, based on effects and values of ecosystem services. The foremost information providers in this category are CBD, IUCN, OECD and TEEB. Among reports and studies published by these organizations, we can cite *Restoring World's Forests Proven to Boost Local Economies and Reduce Poverty* by IUCN and *Green Growth and Sustainable Development* by OECD for example.

In the *Health/Society* domain, repositories supply knowledge information referring to natural resources of ecosystem services and effects. These information have been produced, and their validity was demonstrated, by researchers from world-wide organizations such as BHL, CBD, IUCN, KNEU and UNEP. These organizations provide summaries and proposals such as for example, *Human Health and Biodiversity* is projected by CBD[2], *Towards the Blue Society* is projected by IUCN and *Action for Biodiversity: Towards a Society in Harmony with Nature* is published by CBD.

Information providers of data repositories and resources are categorized among the *Animals*, *Environment* and *Plants* categories depending on the research topics and domains of data they provide. These organizations are: FAOdata which web-based portal provides global data on biodiversity (i.e., Data on Plants and Environment represented as datasets, statistics, spatial data, documents, images, etc.), BIOTIK that is a data repository on plants only in Southeast Asia and some other Asian countries, BISE that is a portal to serve data and datasets on biodiversity (Plants, Animals, and Environment) in European countries, GBIF that is a global data center (Data and Datasets) functioning as a hub of data collections (Plants, Animals, and Environment) from researchers around the world, NBN that is a portal to share biodiversity (Plants and Animals) data and datasets in United Kingdom, ViBRANT that is a web-based portal that aims to facilitate for research communities to merge and share biodiversity (Plants and Environment) data and datasets across European countries and some others, OBIS that is a web-based portal providing data on global marine species and visual spatial information on marine species from all the world's oceans (bio-geography).

2.2.2 Technologies

A very interesting technology that has been developed within the field of artificial intelligence as an outgrowth of early efforts aimed at representing knowledge consists of formal ontologies, which are a key for the semantic interoperability and integration of data and knowledge from different sources.

A definition of an ontology which makes justice of its complexity is the following, proposed in [Guarino 1998]:

> An ontology is a logical theory accounting for the intended meaning of a formal vocabulary, i.e., its ontological commitment to a particular conceptualization of the world. The intended models of a logical language using such a vocabulary are constrained by its ontological commitment. An ontology indirectly reflects this commitment (and the underlying conceptualization) by approximating these intended models.

In other words, an ontology may be regarded as "a kind of controlled vocabulary of well-defined terms with specified relationships between those terms, capable of interpretation by both humans and computers" [Whetzel 2013]. From a practical point of view, an ontology defines a set of concepts and relations relevant to a domain

[2] http://www.cbd.int/en/health

of interest, along with axioms stating their properties. An ontology thus includes a taxonomy of concepts, a formally defined vocabulary (called a *terminology*), and other artifacts that help structure a knowledge base. In a sense, such artifacts may be considered as a generalization of the concept of metadata in database technology [Plant 2010]. A knowledge base that uses the terms defined in an ontology becomes usable by and interoperable with any other system that has access to that ontology and is equipped by a logic *reasoner* for it [Obrst 2003].

Recently, an extensive standardization effort has been carried out by the World-Wide Web Consortium (W3C) in the framework of the Semantic Web movement. The Semantic Web is an extension of the World-Wide Web that enables people to share content beyond the boundaries of applications and websites [Daconta 2003]. The W3C has defined widely-accepted standards that make such an interoperability possible: the OWL 2 Web Ontology language defines the syntax that can be used to write ontologies; many reasoners are available today that are capable of using ontologies written in OWL 2 to make inferences on *facts* stored as RDF graphs [Hitzler 2009]. A query language, SPARQL, is available for retrieving facts from RDF graphs in much the same way as data is retrieved from a database [Seaborne 2008]. Data formatted using the RDF language and linked to ontologies are called *linked open data*, because their adoption of a standard format makes them usable to everybody and connected to all other data which refer to the same shared ontologies. Linked open data is the data layer of the Semantic Web.

Ontologies vary widely in scope and granularity. It is useful to distinguish four kinds of ontologies according to their level of generality [Guarino 1997]:

- *Top-level* or *upper* ontologies describe very general and fundamental concepts like space, time, matter, event, action, quality, etc., which are independent of a particular problem or domain.

- *Domain* and *task* ontologies describe, respectively, the vocabulary related to a generic domain (like biology) or a generic task or activity (like classifying or mapping), by specializing the terms introduced in the top-level ontology.

- *Application* ontologies describe concepts depending both on a particular domain and task, which are often specializations of both the related ontologies.

For example, a Plant Ontology (PO) [2013t], containing a conceptualization of plant structures (including plant cell, plant tissue, and sporophyte) and a controlled vocabulary for describing things like plant anatomy, plant morphology and plant development stage, may be described as a domain ontology.

In general, ontologies do not contain *facts*, i.e., data about the instances of the concepts they define and about their relationships or, when they do, these are limited to a few important facts that are useful to situate or organize the rest of the knowledge. Facts are usually stored in what we will call *fact repositories* which, in some cases, are implemented as or backed by a traditional RDBMS and may contain very large or huge amounts of data.

It is important to clarify that what may be considered an instance varies depending on the domain or task. When it comes to biodiversity, a plant species is generally treated as an instance, even though, strictly speaking, it should be regarded as a concept which groups together all specimens that share a set of phenotypic and genotypic characters. It is the specific application that dictates what should be treated as an instance or a concept.

In addition, ontologies have been extensively used in data integration approaches as they provide an explicit and computer-understandable conceptualization of a domain [Cruz 2005]. One of their major contributions to data integration and analysis is *mapping support*, that is the use of an ontology of terms, formalizing a thesaurus, for the mapping process to facilitate its automation. The term mapping refers here to the semantical linking of data through the concepts represented in the ontology. Ontologies also provide capabilities of generalization and specialization of data according to the ontology concepts and their relationships. Each data can then be considered at the most appropriate level of abstraction, or aggregation, regarding the application objectives [Kwuida 2014].

An illustration of relationships between objects of plant structure is given in Figure 2.8 [Jaiswal 2005]. In this oriented graph, each node represents a term defin-

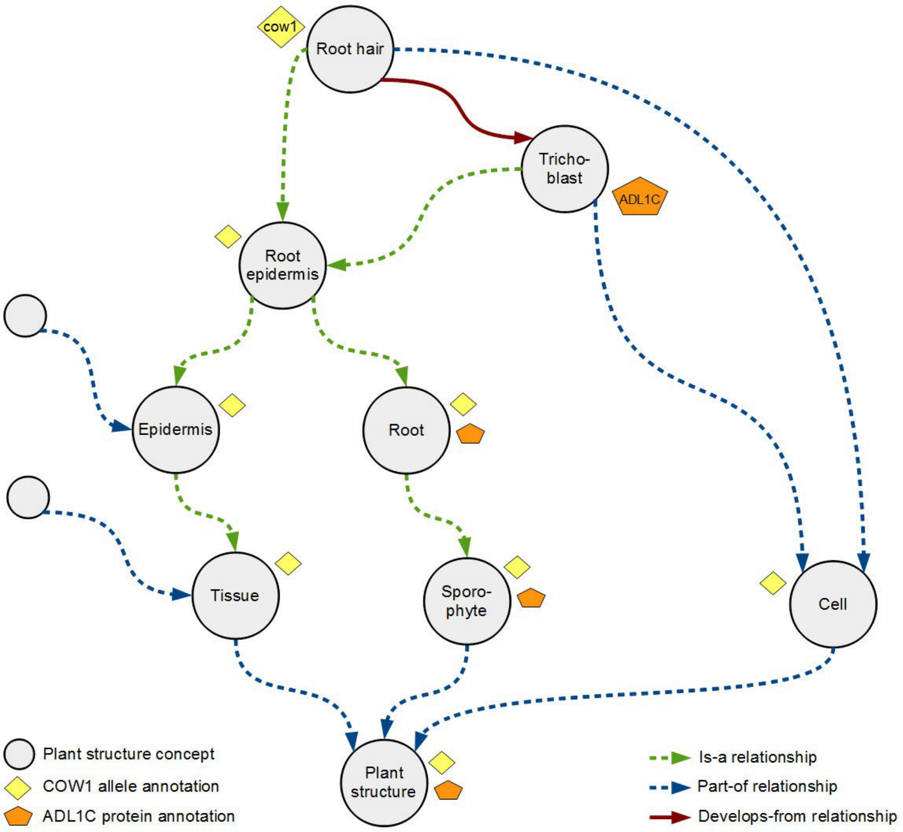


Figure 2.8: Ontological relationships between objects of plant structure

ing an object in plant structure and each arrow represents an oriented relationships between two of these objects. The different colors of arrows correspond to different types of oriented relationships (directional paths). Blue edges represent "is a" relationships, green edges represent "part of" relationships and red edges represent "develops from" relationships. Dot arrows show parent and child inheritance relationships between the two related nodes. The diamond and pentagon-shaped symbols show a specific expression of cell type based on the association of 'COW1' and 'ADL1C'. We can see in Figure 2.8 that the association of 'COW1' and 'ADL1C' is not inherited by children nodes for the "develops from" relationship for instance. Hence, the 'COW1' annotation does not come along with the 'Trichoblast' node. In this figure, the 'Cell', 'Sporophyte' and 'Tissue' are nodes representing objects at the first level of plant structure. The relationships show that the 'Root hair' and 'Trichoblast' are two different types of 'cell' and are also part of 'root epidermis'. They also show that the 'root hair' develops from 'Trichoblast'.

The Open Biological and Biomedical Ontologies (OBO Foundry) [Smith 2007] is the portal of an ontology consortium that provides approaches and tools to help ontology development in different scientific domains of interest. In addition, the OBO Foundry is a large repository of candidate and validated ontologies in a number of scientific domains. Currently, it contains more than 100 ontologies that were declared in different domains[3] and different ontology levels [4].

BioPortal [Whetzel 2011] is a Web portal repository of open ontology resources that allows to search and explore ontology terms through an interactive visualization graphical user interface (Figure 2.9). Ontobee [2013o, Xiang 2011] is a web tool to explore and browse ontological terms annotating linked data from different ontology repositories (Figure 2.10). Moreover, it supplies visualization and supports the SPARQL querying language, as well as management of data repositories of ontological terms, hierarchies and RDF format.

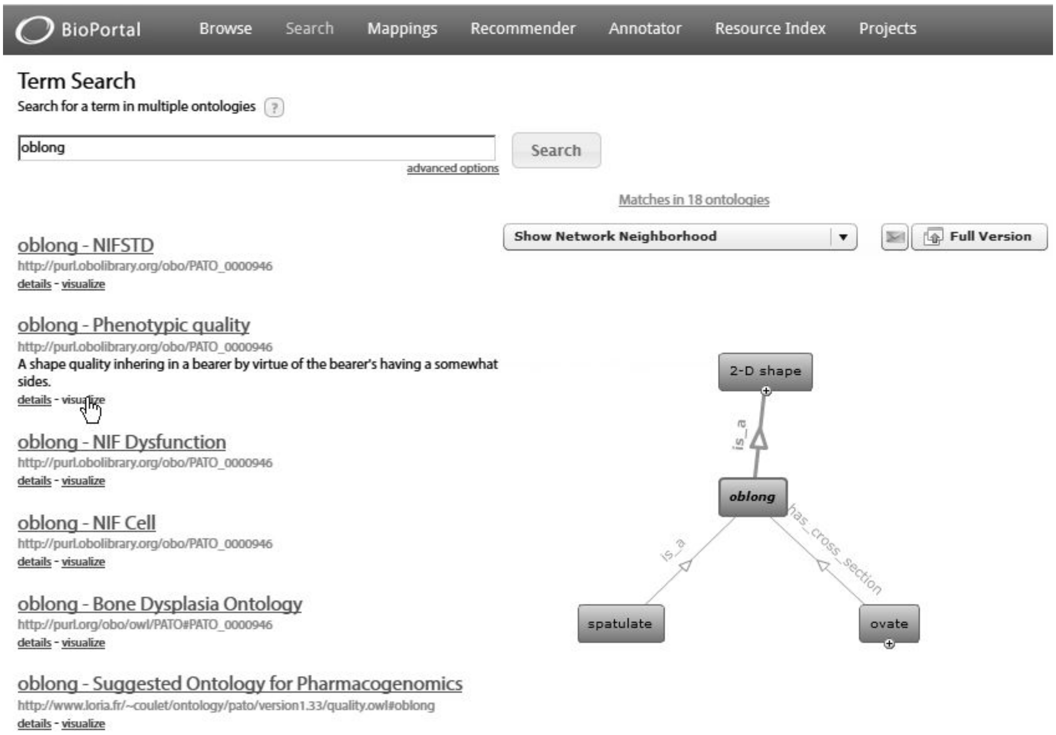
2.3 Data Mining for Biodiversity

2.3.1 Data Mining Concept

Data mining, also known as knowledge discovery from data (KDD), is a set of concepts, methods and tools for the rapid and efficient discovery of previously unknown information, represented as knowledge patterns and models, hidden inside massive information repositories [Han 2011]. The most prominent data mining approaches, gaining actually much importance in many application domains to support decision-making, are association rule and pattern mining, classification, clustering and regression. Since its emergence in the early 1990s, data mining made great strides and continues to flourish nowadays with the rapid evolution of automatic data acquisition systems, such as digital cameras, satellite remote sensing systems, bar code usage in retail, data streams in networks, text and image gathering tools,

[3] See appendix A.2 for a description of domains.

[4] Each table in Appendix B shown an ontology level.



Term Search

Search for a term in multiple ontologies

oblong

oblong - NIFSTD
http://purl.obolibrary.org/obo/PATO_0000946
details - visualize

oblong - Phenotypic quality
http://purl.obolibrary.org/obo/PATO_0000946
A shape quality inhering in a bearer by virtue of the bearer's having a somewhat sides.
details - visualize

oblong - NIF Dysfunction
http://purl.obolibrary.org/obo/PATO_0000946
details - visualize

oblong - NIF Cell
http://purl.obolibrary.org/obo/PATO_0000946
details - visualize

oblong - Bone Dysplasia Ontology
http://purl.org/obo/owl/PATO#PATO_0000946
details - visualize

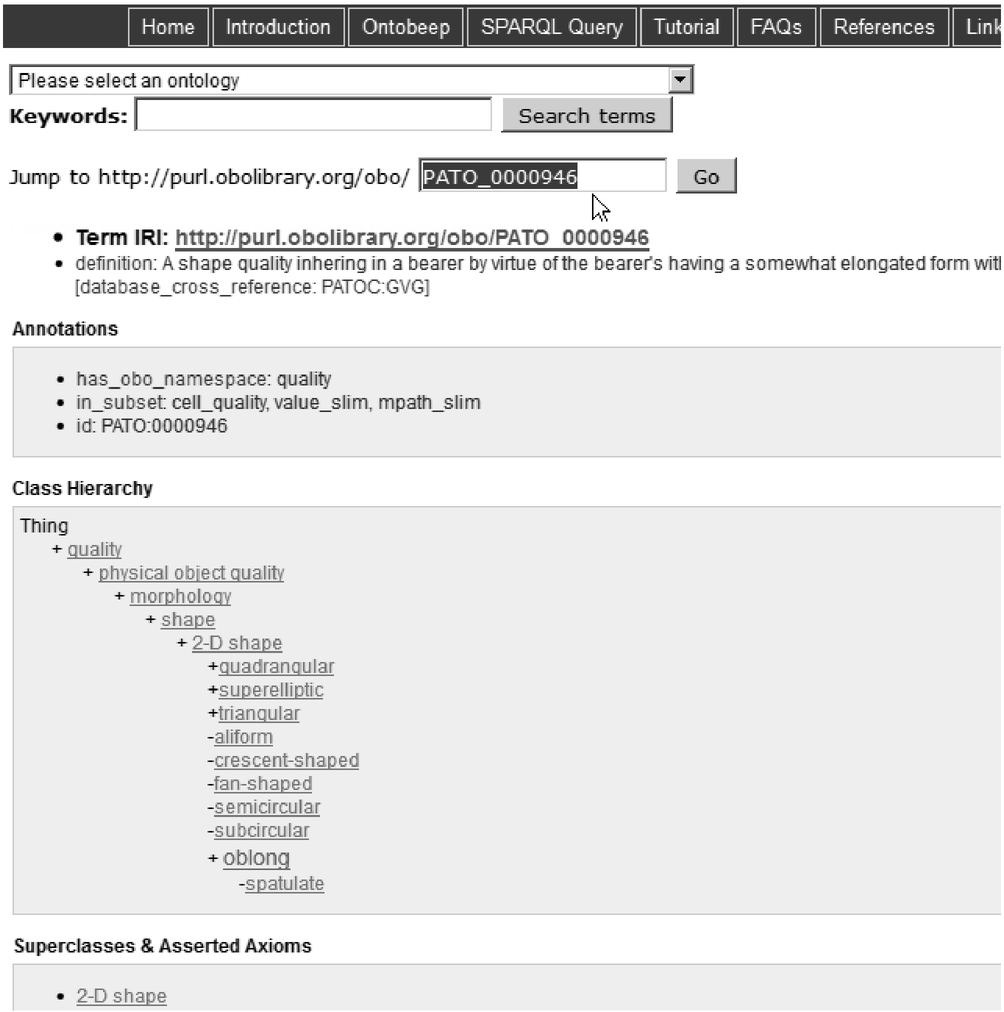
oblong - Suggested Ontology for Pharmacogenomics
http://www.loria.fr/~coulet/ontology/pato/version1.33/quality.owl#oblong
details - visualize

Figure 2.9: Visualization through the BioPortal: results for the search term 'oblong'

or Massively Parallel Signature Sequencing (MPSS) of genes for instance, and of storage systems, such as knowledge and data bases available on the World Wide Web, data warehouses and data marts, or publication repositories for instance. It is a multi-disciplinary field including solutions from database systems, statistics, knowledge-based systems, artificial intelligence, high-performance computing and data visualization.

A data mining process [Marbán 2009, Mariscal 2010] is an iterative and interactive process that typically involves the three following general phases (Figure 2.11). During the pre-processing phase, data preparation techniques, i.e., data cleaning, integration, transformation and reduction methods, are applied to generate datasets containing relevant, consistent and reliable data, from the viewpoint of the application objectives, from heterogeneous data sources. The modeling phase consists in applying algorithmic methods for extracting knowledge patterns and models from the prepared datasets. During the post-processing phase, extracted knowledge patterns and models are presented to the end-user for interpretation and evaluation in order to discover novel information. The interactive and iterative nature of the process relies on the fact that changes and decisions made in the different steps of the three phases can result in changes in later steps and in the extracted patterns and models. Feedback loops between the phases are thus necessary to converge toward a satisfactory solution.

Even if considerable progress has been made in data mining since its early be-



Keywords:

Jump to http://purl.obolibrary.org/obo/ PATO_0000946

- **Term IRI:** http://purl.obolibrary.org/obo/PATO_0000946
- definition: A shape quality inhering in a bearer by virtue of the bearer's having a somewhat elongated form wit [database_cross_reference: PATOC:GVG]

Annotations

- has_obo_namespace: quality
- in_subset: cell_quality, value_slim, mpath_slim
- id: PATO:0000946

Class Hierarchy

Thing
- + quality
  - + physical object quality
    - + morphology
      - + shape
        - + 2-D shape
          - +quadrangular
          - +superelliptic
          - +triangular
          - -aliform
          - -crescent-shaped
          - -fan-shaped
          - -semicircular
          - -subcircular
          - + oblong
            - -spatulate

Superclasses & Asserted Axioms

- 2-D shape

Figure 2.10: Browse the term 'oblong' by term ID 'PATO_000946' in Ontobee

ginning, many challenges still remain. Hence, generic data mining systems can have limitations regarding application specific problems and a trend toward application dedicated systems can be observed nowadays [Han 2011]. These systems aim to handle complex heterogeneous data types, including multimedia, such as images, spatial and text data for instance, to generate complex knowledge patterns, to integrate domain specific knowledge represented both in knowledge bases and as users' methodologies and processes, to develop a unified theory of data mining would address problems in many fields, not only the biological and environmental one [Han 2011, Yang 2006].

2.3.2 Techniques and Applications

Nowadays, many researchers and scientists attempt to solve biodiversity problems by using modern technologies, inventing approaches and techniques to measure and

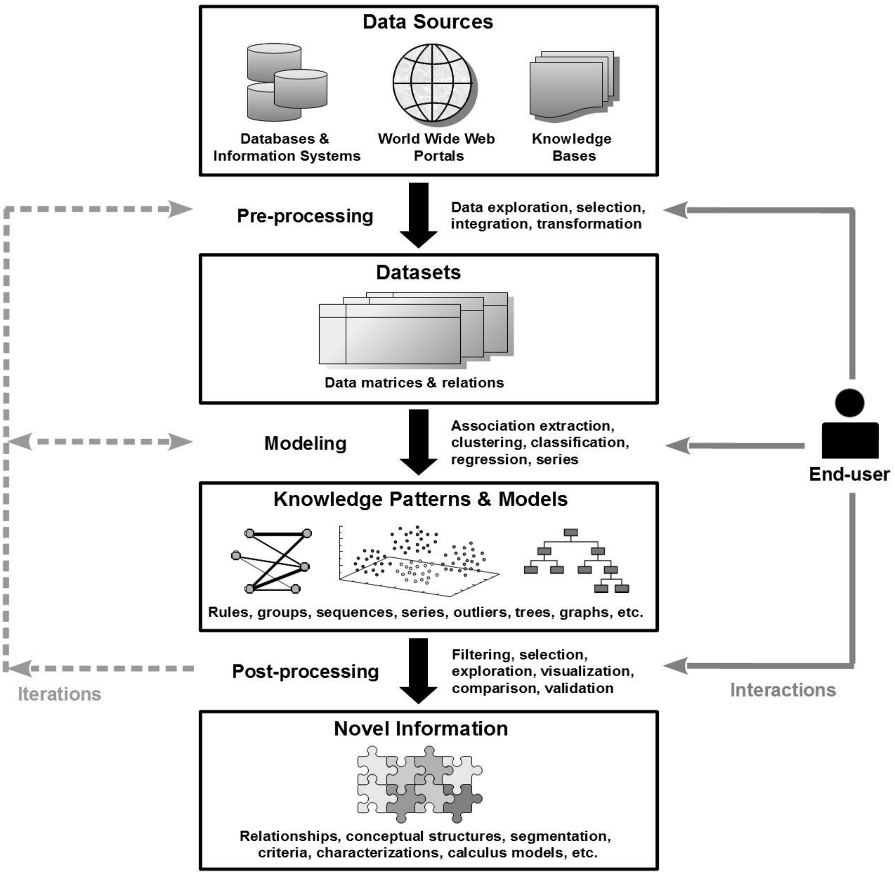

Data Sources
Databases & Information Systems
World Wide Web Portals
Knowledge Bases
Pre-processing
Data exploration, selection, integration, transformation
Datasets
Data matrices & relations
Modeling
Association extraction, clustering, classification, regression, series
End-user
Knowledge Patterns & Models
Rules, groups, sequences, series, outliers, trees, graphs, etc.
Post-processing
Filtering, selection, exploration, visualization, comparison, validation
Iterations
Interactions
Novel Information
Relationships, conceptual structures, segmentation, criteria, characterizations, calculus models, etc.

Figure 2.11: An illustration of the three main phases of a data mining process

solve occurrence biodiversity issues. Because of the deluge of information found in data repositories of environmental sciences, coming from both institutions and amateurs, the use of data mining techniques for discovering new knowledge is spreading rapidly. There is a vast literature on data mining applications for biodiversity and environmental studies as follows.

In [Conruyt 2012], the authors address the problem of the identification and classification of specimens through a knowledge-based discovery system. They extend the "classical" approach that consists of the three following phases: Grouping existing descriptions based on similarity measures by clustering, building and naming the classes identified by classification of groups, and reusing the formed concepts to identify the class of new observations. They combine inductive techniques and iterative neighbor search to take into account the structure (relationships between variables) and the content (missing, deviant and unknown values) of descriptions to improve the robustness of the classification. The resulting approach aims to help botanists and biologists to bring better evidences from knowledge based systems for the identification of new specimens through their observations. The proposed approach was validated with two knowledge bases built for coral classification and plants identification.

The problem of mountain biodiversity studies using data mining techniques is investigated in detail in [Spehn 2009]. It shows the importance of geophysical information systems for exploring and analyzing mountain biodiversity. The problems of the availability, quality and completeness of biodiversity data, that require consolidation before their use in analyses, are addressed thoroughly, demonstrating the importance of high-quality metadata, such as represented in ontologies and knowledge bases, for this complex task. Several case studies, covering all biodiversity levels, from genes to species and ecosystems, using data mining techniques to analyze biodiversity patterns and processes along elevation gradients are reported. These show the relevance of data mining approaches for biodiversity conservation and protected area management, and the study of climate change effects on mountain biodiversity.

Biodiversity in forest ecosystems is the subject of the study in [O'Sullivan 2010]. A data mining based approach is developed to predict biodiversity in forests by reasoning about their physical structure. The authors utilize an high-resolution scanning technology to capture different aspects of forests in three dimensional structure. These data are then related to the diversity of plants, invertebrates and birds in a range of forest types, to generate rich physical description datasets. These datasets are analyzed afterwards using five regression techniques from the popular Weka [Hall 2009] data mining application. Results show that this approach can accurately predict six biodiversity measures of the species richness and the abundance of beetles, birds and spiders. This is a step toward the automation of the creation of a world forest inventory rich with environmental concerns.

In [Hochachka 2007], the authors compare statistical and data mining methods for identifying relationships between a response and a set of predictors. They show that when little or no prior knowledge about the studied system is available, statistical models cannot accurately describe relationships between variations of the predictors and the response variable. Experiments were conducted using different datasets, including geographical, temporal, climate and species data, to compare results of six popular data mining tools with results of statistical techniques. The authors conclude that more use of data mining techniques should be made in environmental studies, whatever the degree of prior knowledge, and propose effective solutions to integrate data mining and statistical analyses into a thorough analysis.

The integration of geographical information from Geographical Information Systems (GIS) with species data, and its use in data mining studies is the object of the biodiversity informatics project of the W. P. Fraser Herbarium (SASK) [Peters 2009]. The participants to this project develop an integrated bio-geography GIS model, using Google Maps API, based on data mining concepts to map and explore flora data. This research project shows that these data can be explored on a map and analyzed in several ways to reveal patterns showing relationships and trends that are not discernible in other representations of information.

The general problems of information integration and descriptive data quality are addressed in [Paterson 2004]. These problems are considered in the context of taxonomic classification of plant specimens into taxa, i.e., groups, according to the similarities between their observed features, or characters. This classification process relies mainly on the identification and description of variations between comparable structures of the different species. Since several terminologies and methodologies are in use for composing character descriptions, most often these descriptions are inconsistently composed, difficult to interpret and re-use, and data from diverse sources are not comparable. The authors propose a new conceptual model for unambiguously representing quantitative and qualitative description elements. It makes use of ontology technology to represent concepts and relationships in the descriptive terms. This model was implemented in a Java tool to help taxonomists to classify specimens and describe characters of new specimens through defined and controlled vocabularies.

In [Raguenaud 2001], the authors propose two approaches for representing plant classifications that are multiple and overlapping since in taxonomic classification of plants, some groups of specimens are referred to by a name used in different contexts over time. In both approaches, graph structures are used to represent classifications and relationships between them, but the two corresponding data models are different due to their constraints and aims regarding their capabilities to perform users' tasks. In the first approach, named Database Approach, the data model is dedicated to the storage of plant information. In the second approach, named Visualization Approach, the data model aims at the efficient retrieval and graphical exploration of plant information and their relationships. The results show that the two data models should coexist in a unique system for the automatic processing capabilities of the first one and the efficient exploration and comparison of classifications allowed by the second one.

On the other hand, Figure 2.12 shows a hierarchical structure of relationships between data mining and biodiversity tasks. From top to down, data mining approaches and models can apply to solve the problems on biodiversity domain such as [Peters 2009] and [Spehn 2009] that used both patterns and regression for analyzing and exploring/visualizing biodiversity data.

- The *Link Analysis* category is a collection containing a variety of techniques to perform evaluation and validation on data [Donoho 2010].

- The *Similarity Analysis* category is a collection of statistical approaches to examine statistically a set of two similar things or more pairs of samplings.

- In the *Prediction* category, learning methods are applied to a set of instances for which the value of the objective variable is known in order to generate a predictive model that will be applied to new instances for predicting the value of this variable.

- *Genomics Domain* is a part of genetics, which studies all gemomes of living organisms. In biology/biodiversity, the genome refers to all genetic information in the DNA, which is vital to create and maintain for the survival of living species [Knapp 2004].

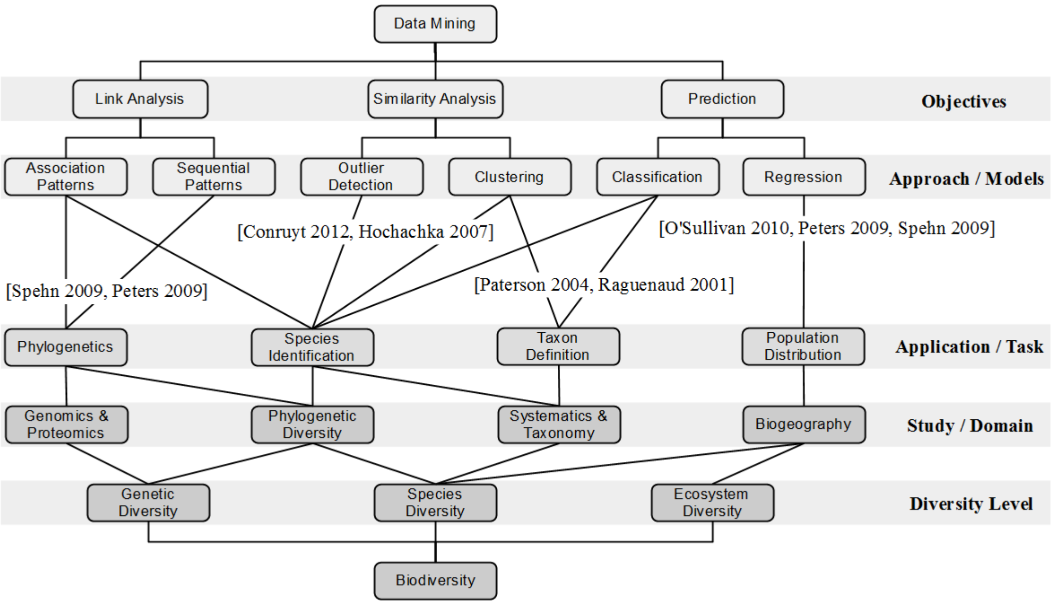


Figure 2.12: Data mining techniques and biodiversity applications

- *Proteomics Domain* studies proteins of all living organisms, including structures and functions. Protein is derived from a Greek word which means "primary". The protein is a biochemical compound which is vital for living species. It has a complex structure and molecular mass [Gotelli 2012].

- *Phylogenetic Diversity Domain* is a measure of biodiversity from the viewpoint of phylogenetic (evolutionary histories and relationships) differences among species and populations [Faith 2014].

- *Systematics Domain* (in biology/biodiversity) refers to the phylogenetic classification of living species (evolutionary histories and relationships). Systematics is as "the fact-finding field" of taxonomy [Porter 2008].

- *Taxonomy Domain* involves to bring the facts to the identification, description, nomenclature of species, and classification of species into a system based on their shared characteristics [Mason 1950]. This domain is essential to the study of biodiversity [Gaston 1992, Pennisi 2000, SCBD 1992].

- *Biogeography Domain* is the study of spatial and temporal distributions of species, habitation patterns, populations, and ecosystems. In addition, the climate and terrain are variants to the distribution of species, depending on the suitability of such environment and obstacles such as mountain, desert, and ocean [Cox 2010, Dansereau 1957].

Biodiversity applications showed that data mining can successfully discover new results and information to help environmental scientists to explain phenomena and get new insights in particular. However, these results can be improved by integrating data and knowledge from different related application domains into the biodiversity data mining process. For this integration, data, such as stored in application-level ontologies and databases, can be pre-processed using the structured representations of the domain knowledge stored in upper-level and domain-level ontologies. This approach can consolidate extracted information and help to solve complex problems of biodiversity and environmental studies that require to analyze data and knowledge from different domains together (e.g., environment, biology, geospatial topology).

Recently, many biodiversity and environment knowledge bases have been published as off-line and on-line resources, as well as some upper-level ontologies. At this point, the main challenge is to store and integrate the most relevant information from the different knowledge and data bases into a unified information system. This information system aims to provide *biodiversity background knowledge* (BBK) for generating different datasets for biodiversity studies. For efficient automatic processing, a database can be used to store application level data, and domain knowledge can be represented in a conceptual base as depicted in Figure 2.13.

Data can then be processed according to domain knowledge, integrated and consolidated using upper-level ontologies, and users' requirements. For example, hierarchical discretization of numerical data and hierarchical categorizations of discrete data stored in the conceptual base can be used to generate datasets containing data at different levels of abstraction. Using domain-level knowledge in the conceptual base, data can also be selected according to concepts at different levels of abstraction.

This background knowledge integration approach was applied to create a plant knowledge base from the BIOTIK plant Web portal, the BRAHMS repository, the NAPIS repository and the IUCN Red List Data Web portal. Conceptual bi-clusters and conceptual association rules, based on the Galois closure theoretical framework [Everett 1944, Ganter 1999], were extracted. In such conceptual knowledge patterns, each property extracted from the data, e.g., a rule or a cluster, is associated to the instances that support this property. This feature can improve the analysis process and help the analysts to understand the underlying phenomena as extracted patterns which are related to both application-level data and domain-level knowledge. Preliminary results show that extracted patterns and models can successfully link information from different domains and from different types [Inthasone 2014].

Background knowledge integration aims at consolidating relevant information available in different sources of knowledge. The resulting knowledge can be stored into databases and conceptual systems to build central services of a biodiversity information system. Application data can then be integrated and processed according to this knowledge, taking into account the specificities of the analysis context and application, to generate diverse datasets. For example, a background knowledge base can be used for the identification of new specimens in botanical researches, the study of the effects of global warming on biodiversity loss and the prediction of economic trends on food production and consumption.

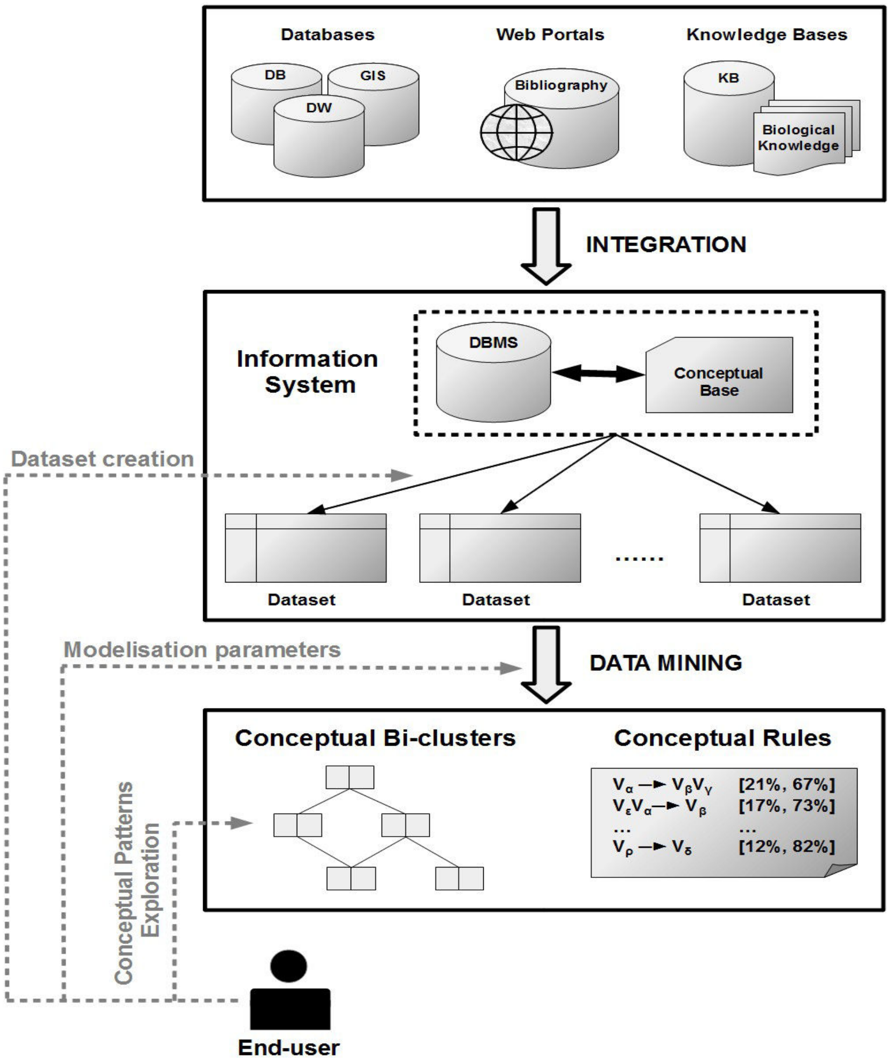


Figure 2.13: A data mining approach for integrating knowledge bases

2.4 Data Integration and Data Warehouse

2.4.1 Data Integration

Data Integration refers to a consolidation of different data from heterogeneous sources across the entire public and enterprise. In general the data integration architecture system consists of three parts, data sources which are stored in plain text or relational database systems, the extractor or loader which fetches data from the data source, and data warehouse which keeps all data tracks [Doan 2012]. Figure 2.14 shows a classical architecture of data integration process.

There is a vast literature on the integration of heterogeneous information sources and data visualization. In [Kuenne 2007], the main concept focuses on two aspects. The first one is to integrate Crop plant data from specimens of 300 ryegrass and

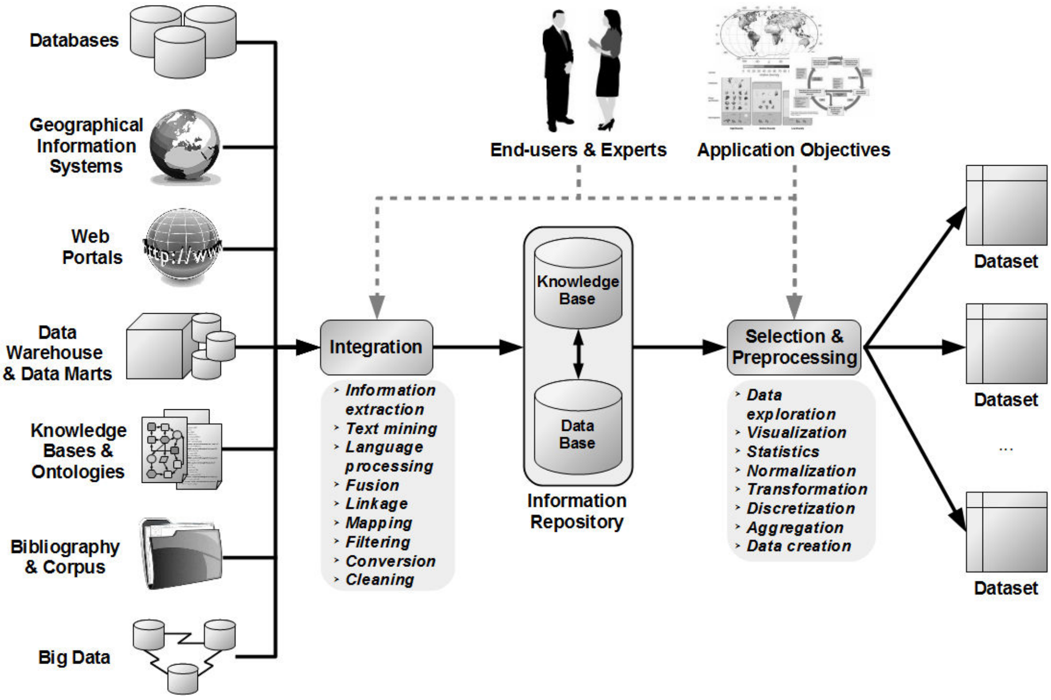


Figure 2.14: A classical data integration process

250 barley. The second one is to design and implement Crop plant data warehouse for supporting plant biological data analyses. A Kew project at Royal Botanic Gardens in UK [Bachman 2011] supplies a utility tool on the GeoCAT open source platform which enables users to visualize biodiversity data by using the extent of occurrence (EOO) and the area of occupancy (AOO) functions and analyse the risk status of biodiversity data. Model and tool development for the integration of information the TSIMMIS Project [Chawathe 1994] proposed a conceptual model called "Object Exchange Model (OEM)" and tools for combining data from multiple information sources. The authors [Wache 2001] present how to use ontologies in 25 extisting approaches including SIMS, TSIMMIS, OBSERVER, CARNOT, Infosleuth, KRAFT, PICSEL, DWQ, Ontobroker, SHOE and others for the integration of information from heterogeneous information sources. These approaches were analysed with the role of ontologies, Ontology Representation, Use of Mappings, and Ontology Engineering. The new approach for the integration of information the authors [Domenig 2000] present a query language extension approach to fetch heterogeneous information sources from different relational database systems. In [Cruz 2005], the authors present the role of ontologies which focuses on five case studies to apply for data integration task. Case study 1: Metadata Representation, is to prepare generating the XML source schemas to the local ontologies. Case study 2: Global Conceptualization, is to map between the local ontologies and global ontology. Case study 3: Support for High-level Queries, provides a conceptual view on local sources and an inference mechanism. Case study 4: Declarative Mediation,

is to declare a mediation for peer-to-peer query rewriting. Case study 5: Mapping Support, provides steps "Path Exploration", "Path Selection", "Semantic Derivation" for the mapping process.

2.4.2 Data Warehouse

Data warehouse (or DW for short) generally denotes, a huge storage of data and knowledge from heterogeneous sources, referring to a data vault which gathers a variety of different databases and transaction repositories in the whole enterprise to support scientists and business decision makers (See Figure 2.15).

> There are other definitions and viewpoints about DW:
> "A data warehouse is made up of all the data marts in an enterprise." - Ralph Kimball
>
> "A data warehouse is simply a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use it in a business context." - Barry Devlin and IBM Consultant
>
> "A data warehouse is a subject oriented, integrated, non volatile, time variant collection of data in support of management's decisions." - William H. Inmon

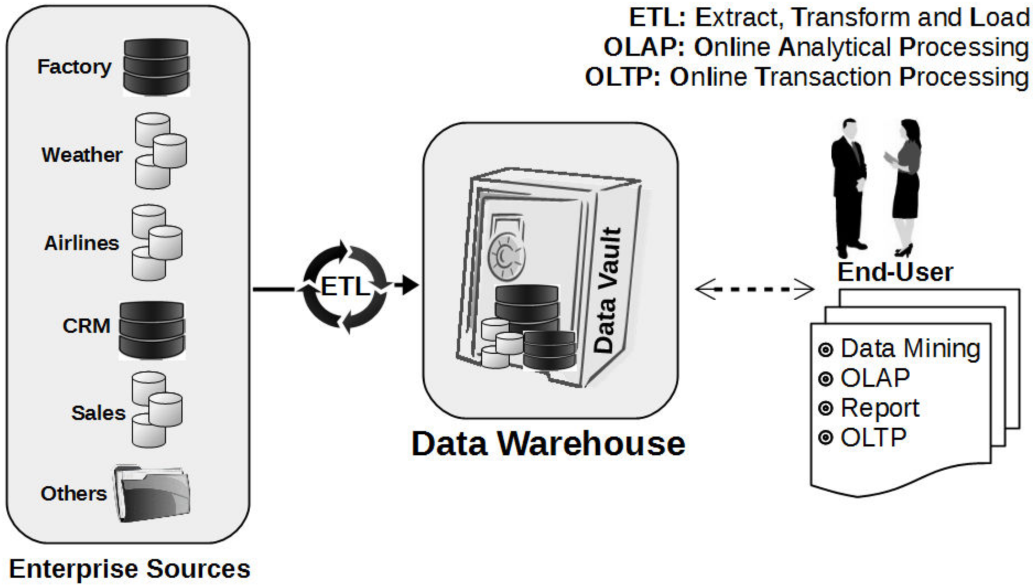


Figure 2.15: An overview and perspective of data warehouse

The DW discussion is divided into two subsections which focus on DW Architecture and DW Design as follows.

(a) DW Architecture: Figure 2.16 shows a typical DW architecture which consists of two main parts. The first one is formed by heterogeneous data sources are stored in operational systems (DBMS, Non-relational DBMS) such as Amazon Online Shopping and Wal-Mart Stores or stored in flat files/plain text files (text, CSV, Spreadsheet, etc). The second one is a set of DW layers (tiers) including Back-End tier, DW tier, OLAP tier and Front-End tier. During the ETL process, the Back-End tier extracts data and loads all transactions into the DW tier. DW tier is a data pool to keep all tracks of valid data and transactions including Data Mart and Matadata. OLAP tier acts as an online processing server to treat end-user's requests through Front-End tier. And the Front-End tier supports utility tools including OLAP tools, Reporting tools, Statistical tools, Data Mining tools, and Data visualization tools. These tools enable end-users to perform analyses for their purposes, such as online banking, daily stock exchange reports, weather forecasting, classification and identification of species for conservation, geospatial visualization like Google Maps [Inmon 2005, Kimball 2002].

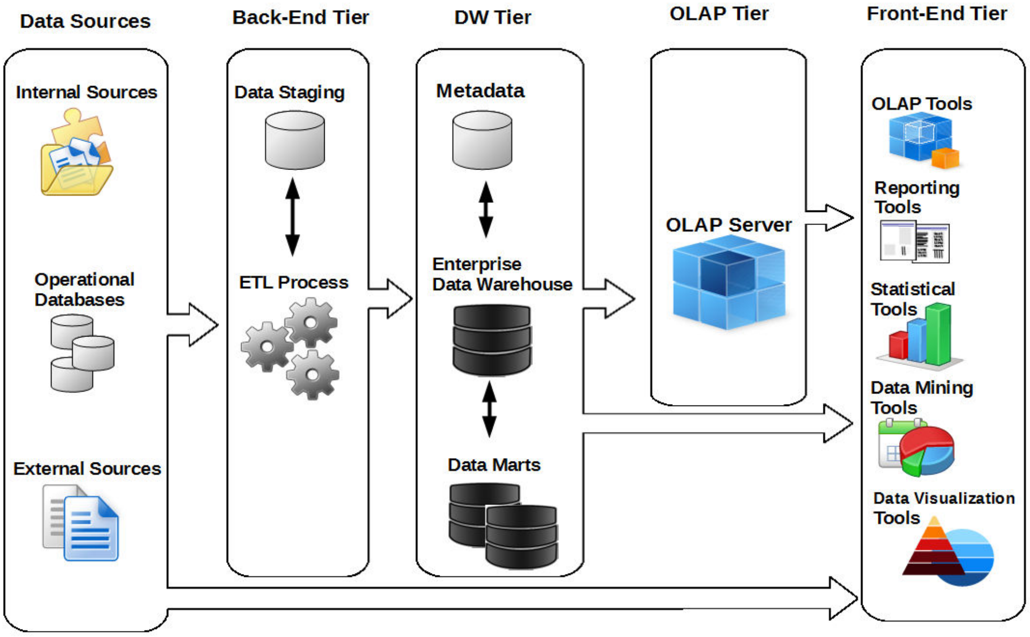


Figure 2.16: A typical perspective of data warehouse architecture [Malinowski 2009]

(b) DW Design: Data Warehouse design is a process, an important part of building a Data Warehouse, to plan DW data modelling. There are two typical model approaches which are most commonly used to design for a database and DW. The Relational Model [William H. Inmon] and Multidimensional Model for Kimball. Inmon's approach on the Relational Model consists in arranging data into tables with relationships between rows of data by primary keys and foreign keys. Figure 2.17 shows relationships between tables (entities) through their primary and foreign keys. On the other hand, Kimball's approach on the Multidimensional Model is presented for designing and building DW as follows:

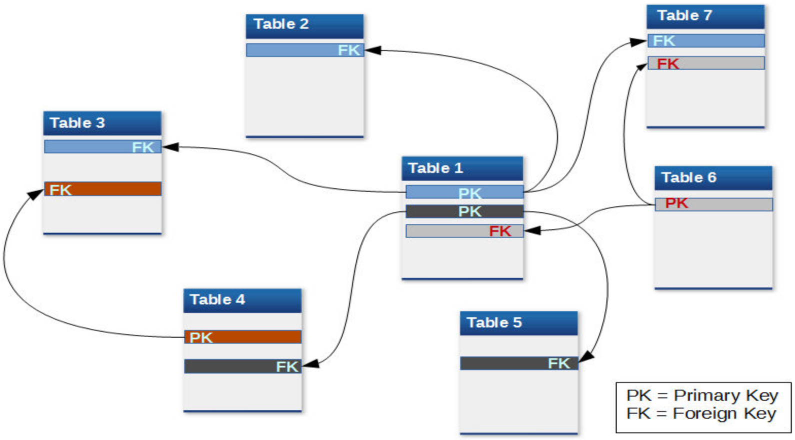


Figure 2.17: An overview of typical relational database design

- Star Join Approach: The Star Join structure (Figure 2.18 (Left)) consists of a fact table and dimension tables. The fact table is central, which contains data and keys from dimension tables. The dimension tables are denormalized while the fact table is normalized. Many advantages of Star Join approach are to design DW schemas quickly, perform often used data queries faster across entire tables including the fact table and dimension tables, but data redundancy may occur in some or even all tables.

- Snowflake Approach: The Snowflake structure (Figure 2.18 (Right)) consists of fact tables and dimension tables, which extend from the Star Join structure. All fact tables and dimension tables are normalized. And one of the advantages is to improve the Star Join Approach on data redundancy, but data query syntax is more complex.

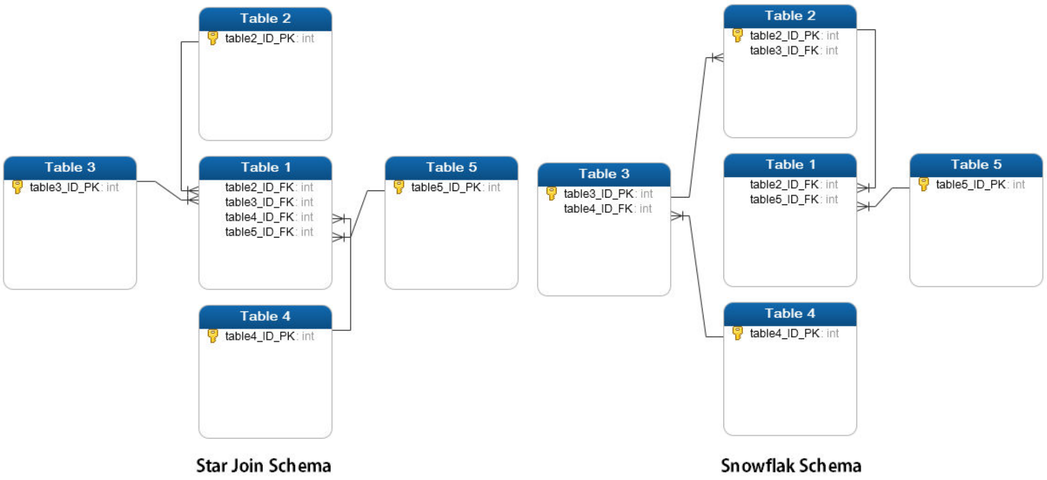


Figure 2.18: An overview of Star Join (left) and Snowflake (right) schemas

The DW design approach is applied widely to build DW systems for scientific analysis purposes, education, and enterprises. [Bassil 2012] presents the usage of the DW design approach to design and build a University Information System under MS Access 2010, used in order to generate datasets for data mining analysis purposes. Finally, [Kuenne 2007] presents an analysis including Data Integration and Data Warehouse Design for building a Crop Plant Bioinformatics Data Warehouse System.

In this chapter, finally the BioKET background and related work were addressed including biodiversity section, Resources and Technologies for Biodiversity, Data Mining for Biodiversity, and Data Integration and Data Warehouse. The next chapter is BioKET Data Warehouse.

Chapter 3

BioKET Data Warehouse

This chapter focuses on five technical issues. This will present on how to build, design, and implement the BioKET Data Warehouse, on how to integrate and visualize the BioKET data with geolocation data (geospatial data), and how to prepare patterns for generating the BioKET dataset and presenting the BioKET plant ontology mapping concept.

Contents

3.1 BioKET Data Warehouse Model

According to botanical description, a species has distinct features and properties. Therefore, the species data contains its variety of information including plant structure, medical information, bibliographic information, and geolocation information (Figure 3.1). With these information, firstly, we performed a data analysis of 17 categories of information of BIOTIK [2011b] data source, 20 categories of information of BRAHMS [2013e] data source, and 21 categories of information of NAPIS data source [2013n] as shown in Table 3.1, 3.2, and 3.3 respectively. Secondly, we designed the BioKET Data Warehouse model (Figure 3.2) using the Snowflake method, – 61 entities with relationships and a data dictionary detailed in Appendix D. Figure 3.2 shows the corresponding BioKET Data Warehouse ER schema which consists of 35 fact tables and 26 dimension tables. Finally, we translated the BioKET Data Warehouse ER schema (Listing E.1 in Appendix E) to a relational schema to be used with a MySQL database management system platform.

3.2 BioKET Plant Data Integration

The BioKET data warehouse is the consolidation of four main data sources:

- BIOTIK [2011b] (Western Ghats of India and National University of Laos), which contains 652 species records;

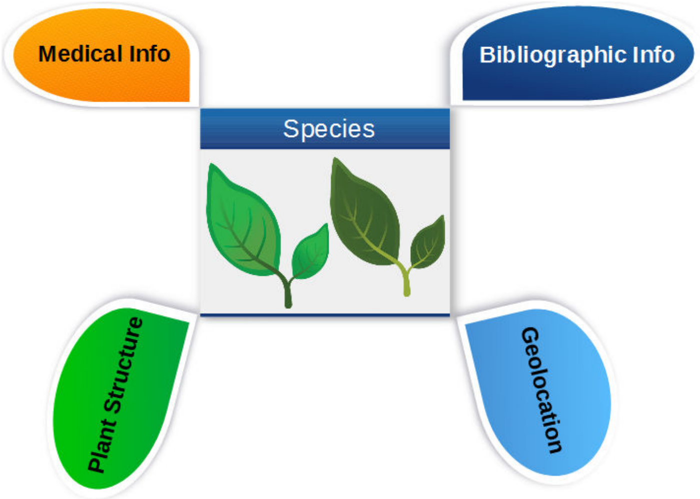


Figure 3.1: A species description contains a variety of information

- the herbarium from the BRAHMS repository [2013e] (National University of Laos, Faculty of Forestry), with 7548 species records;
- the herbarium from the NAPIS repository [2013n] (Lao Ministry of Public Health, Institute of Traditional Medicine), with 747 species records;
- the IUCN Red List Data [2014s], with 71570 species records.

These data sources are stored in different formats: BIOTIK and IUCN Red List are in HTML, while the two others use, respectively, the dBase and Paradox file formats. Integrating such diverse data sources (Figure 3.3) required performing the following tasks.

- The first step was to extract data from the BIOTIK and IUCN Red List repositories and use VBA Scripts (See Listing C.2 and Listing C.3 in Appendix C) to preprocess and store them in a tabular file format (Excel spreadsheet). Data from the BRAHMS and NAPIS repositories were also exported to a tabular file format (Excel spreadsheet) using the Microsoft Management and Navicat Premium database management tools.
- The second step, data cleaning, was performed by using advanced Excel functions (INDEX, TRIM, CONCATENATE, MATCH, IF, Remove Duplicates, Text to Columns, etc.).
- The last step was to use an administration tool (MySQL Workbench/Nivatcat Premium) importing the data thus obtained into the BioKET database, under MySQL.

A key factor for the integration and the enrichment of the data was the use of ontologies. Formal ontologies are a key for the semantic interoperability and integration of data and knowledge from different sources. Ontologies define controlled

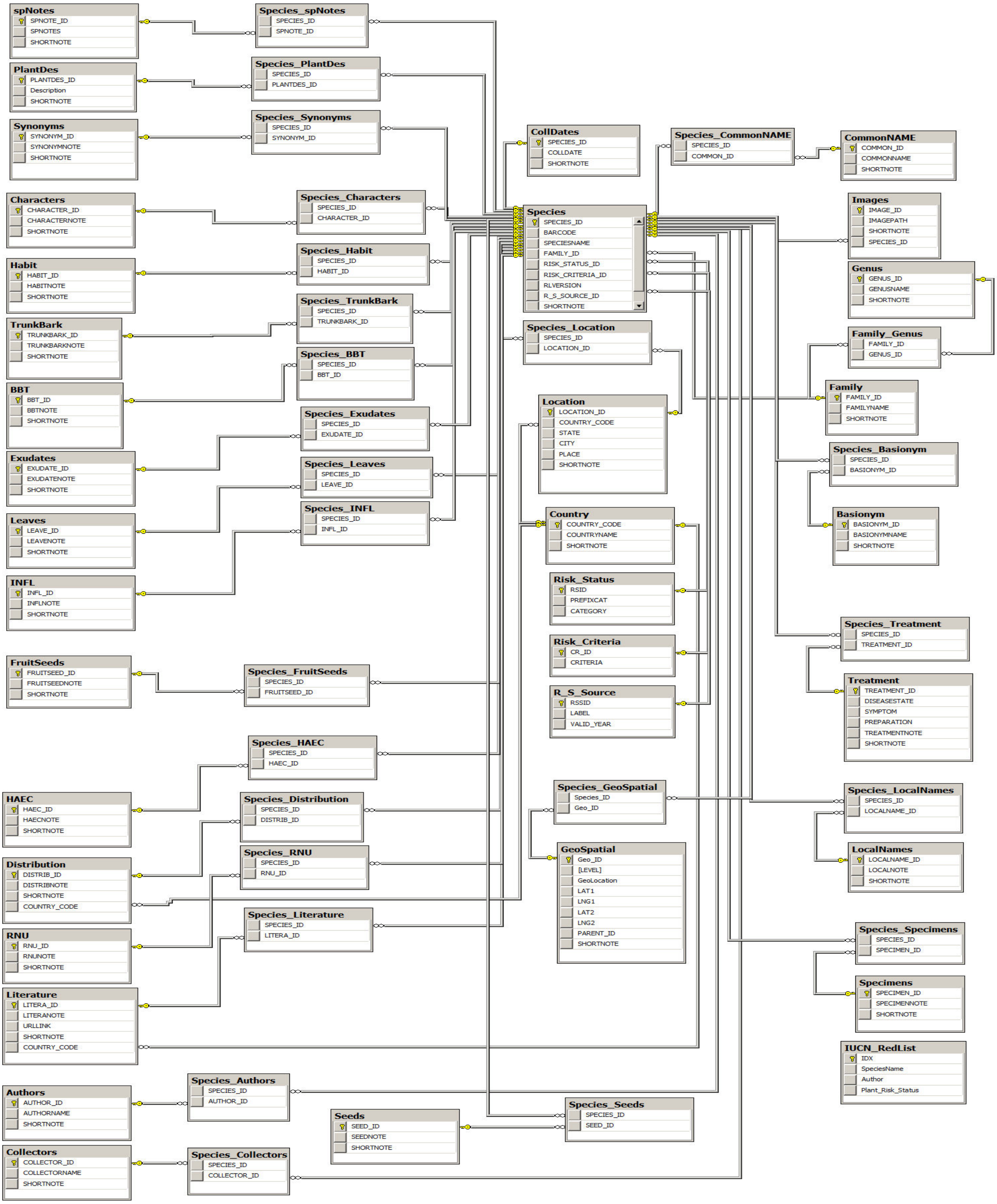


Figure 3.2: The BioKET data warehouse ER schema (61 entities)

Table 3.1: BIOTIK field elements

Index	Category of information	Instance
1	Species name	Acer laurinum Hassk. - ACERACEAE
2	Synonym	Acer decandrum Merr., Acer garettii Craib, Acer niveum Blume, Acer philippinum Merr.
3	Diagnostic characters	Evergreen trees bark scaly. Leaves simple, opposite, glabrous, below glaucous. Stipules absent. Flowers white. Fruit a winged samara
4	Habit	Evergreen tree up to 40 m tall
5	Trunk and bark	Bole straight, bark scaly, red-brown
6	Branches and branchlets or twigs	Twigs terete, glabrous
7	Exudates	Exudate absent
8	Leaves	Leaves simple, opposite, ovate or elliptic, glaucous below, glabrous, apex acuminate, base rounded. Midrib flat above, 3-5 basal secondary veins, secondary veins oblique, tertiary veins reticulate, Stipules absent
9	Inflorescences or flowers	Flowers white, large, arranged in an inflorescence, axillary, bisexual, pedicel longer than 0.5 cm long
10	Fruits	Fruit 3.5 - 7.5 cm, a double samara, locules smooth inside, not splitting open
11	Seeds	2 Seeds
12	Habitat and ecology	Usually undisturbed evergreen forest, common or scattered
13	Distribution	Burma (Myanmar), Cambodia, India (Assam), Indonesia, Laos (Khammouan), Malaysia, Nepal, Philippines, Thailand
14	Remark/Notes/Uses	No uses known, relatively rare
15	Specimens studied	BT 55 (Herbarium of Faculty of Sciences-NUoL, NHN-Leiden and CIRAD-Montpellier)
16	Literature	Gardner S., Sidisunthorn P. & Anusarnsunthorn V. 2000. A field guide to Forest Trees of Northern Thailand. Kobfai Publishing Project. Bangkok. Thailand
17	Images	2 files

vocabularies, consisting of consensual terms and relationships between these terms, that can be processed both by humans and by automatized process. From a practical point of view, an ontology defines a set of concepts and relations relevant to a domain of interest, along with axioms stating their properties. An ontology thus includes a taxonomy of concepts, a formally defined vocabulary (called a terminology), and other artifacts that help structure a knowledge base. A knowledge base

Table 3.2: BRAHMS field elements

Index	Category of information	Instance
1	Collector	Lamxay V.
2	Addcoll	Pheng, Somchanh
3	Specimen	229,1893
4	Collection Date	1/10/1998
5	Barcode	FOF0001893
6	Family	Orchidaceae
7	Genus	Habenaria/Pecteilis
8	SP1	Habenaria susannae/Pecteilis susannae
9	Author1	R. Br.
10	Determination	N/A
11	Country	Laos
12	Majorarea	Vientiane Capital
13	Minorarea	Xaythany District
14	Locality name	Nang Oua
15	Locality notes	Houay Ngang NBCA
16	Collection map(Lat,NS,Long,EW)	Lat=18.1105410, Long=102.5298028
17	Altitute(m,atl,maxatl)	N/A
18	Plant description	Herb, pale green stem, pale green adaxial, greengrey abaxial, inflorescences on top white flowers
19	Habitat/Site description	Dry dipterocarp forest
20	Cultnotes	N/A
21	Images	3 files

that uses the terms defined in an ontology becomes usable by and interoperable with any other system that has access to that ontology and is equipped by a logic reasoner for it [Obrst 2003].

It was thus important to construct a map among all the concepts in all the data sources and all the considered ontologies. It is worth noting that (i) some concepts are not equally represented in all the sources, (ii) some are represented in some sources and not in others and (iii) other concepts are not represented at all. The mapping process works (as shown in Figure 3.8) as follows: the textual descriptors of plants are segmented into small chunks, which are then matched with the labels of concepts in the target ontology. For instance, from the descriptor “evergreen tree up to 8 m”, we can infer that “evergreen” is related to “shedability”, “up to 8m” is related to “height” and “tree” is related to “plant type ”. In the process, new concepts may be generated (e.g., from the textual descriptor “branches ascending or horizontal”, where “branch”, “branch ascending”, and “branch horizontal” match concepts in the ontology, a new concept “branch ascending or horizontal”, subsumed by “branch” and subsuming the latter two is generated). The plant record can thus be automatically enriched with a large number of “implicit” fields, inferred from the ontology. We designed a relational data base of concepts that make it possible to

Table 3.3: NAPIS field elements

Index	Category of information	Instance
1	Interviewer	Khamphao
2	Province	Attapeu
3	Village	2 villages
4	Family	Fabaceae-PAPILIONOIDEAE
5	Genus	DALBERGIA
6	Species	BARIENSIS
7	Authority	Pierre
8	Synonym	N/A
9	Det by/date	02/02/2010
10	Common Name	Tonh pa dong deng
11	Rationale	ETHNOMEDICAL
12	Sample Identity	SL7153
13	Sample Part Description	SB
14	MUU disease state or system	PAIN
15	LN symptom	HEADACHE
16	Medical Use Part	SB
17	MUPR med-use preparation	Cut into small pieces and dry. Crush, wrap up with indigo cloth, steam for 20 minutes, stand to warm.
18	Informant MUIN informant,Note	2 healers
19	Age category	> 65 yrs
20	MUC original data, notes	Put on the head 3 times a day.

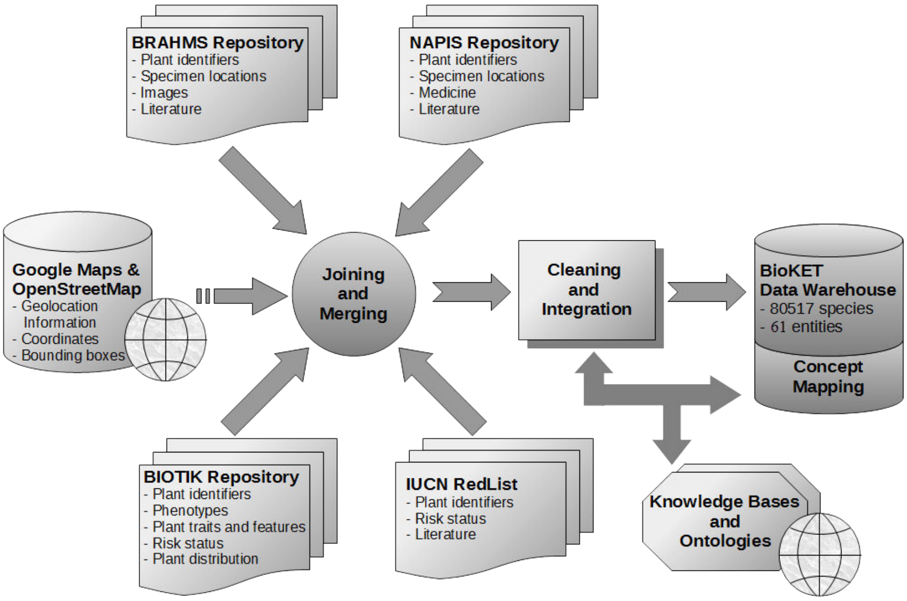


Figure 3.3: An overview of the BioKET heterogeneous data integration process

relate concepts represented in different ways in Section 3.1. Currently the BioKET database contains 61 relationship entities and a total of 80,517 records.

3.3 BioKET Geospatial Data Integration

As pointed out by many researchers (see, e.g., [Benniamin 2008]), to conserve organisms whether about plants or animals, one important step to take is to identify rare and endangered organisms in a given geographical area or country. The GIS technology is sufficiently mature. Indeed, several have been invented such as Google Maps, OpenStreetMap, DBpedia, Microsoft Bing Maps, NASA WorldView and others. Following [Peters 2009] we focused to integrate Google Maps with BioKET data. By using a VBA script (GoogleGeoLocation Function [1]) with input of real-world location data (i.e., continent, region, country, etc.) we extracted coordinates (latitude and longitude) from Google Maps Geocoding Service, and stored them in a GeoSpatial table.

An integration process of Google Maps Geocoding Service with terms of geographical information of each source (BIOTIK, BRAHMS, and NAPIS) is to match coordinates of locations on the earth with the terms in the BioKET database system (Species and Distribution table). For example, A species name "Aglaia elliptica" (in Species table) was found in China, Laos, Malaysia, Vietnam (in Distribution table) and these countries' coordinates were extracted and stored in GeoSpatial table: coordinates (35.8616600,104.1953970) of China, coordinates (19.8562700,102.4954960) of Laos, coordinates (4.2104840,101.9757660) of Malaysia, coordinates (14.0583240,108.2771990) of Vietnam.

The BioKET data warehouse integrated geographical information and 8,947 species of the 80,517 total species have descriptions of specimen location and risk status that may differ depending on the area considered. This information is described at different levels of precision, from continent to specific places such as cities or villages. For example, *Cratoxylum formosum* grows up in Myanmar, Southern China, Thailand, Indochina, and Laos (Khammouan) [2011b]. This species is also reported in the Lower Risk/Least Concern category by IUCN Red List data [2014s]. The integration of geolocation information allows to explore species properties in different areas using the GeoCAT (Geospatial Conservation Assessment Tool) platform [2014j]. GeoCAT is based on Google Maps to explore geographical information if coordinates, i.e., latitudes and longitudes, are provided.

The Google Maps does not support coordinates of directions (South, North, East, West, etc.) like “Southern China”, but Google Bounding Box (BBox) coordinates are provided. We propose to improve this issue by calculating the coordinates for each direction (Figure 3.4) from Google BBox coordinates.

In the geolocation domain, the BBox of an area on Earth is defined by two points corresponding to the minimal and maximal longitudes and latitudes of the area [2014f]. Figure 3.4 shows the 13 partitions of an area: the 9 elementary partitions and the North, South, East and West partitions that result of merging the 3

[1] See coding in Listing C.4, Appendix C

corresponding elementary partitions, e.g., NW, NC and NE for North. This multi-level partitioning allows to represent location related properties of species, such as risk status or abundance for instance, at different area covering levels.

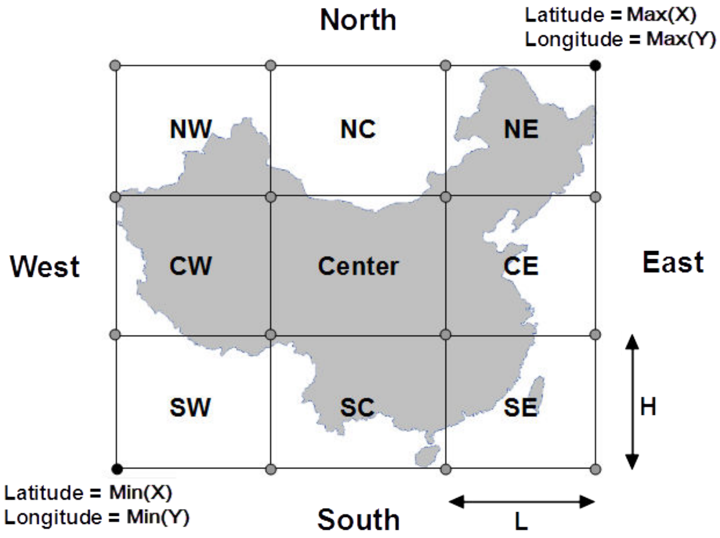


Figure 3.4: Thirteen partitions of a region/area on the earth

Formulas to calculate the BBox of each partition are given in Table 3.4. These computations use the L and H values computed from the minimal ($Min(X)$, $Min(Y)$) and maximal ($Max(X)$, $Max(Y)$) longitude and latitude coordinates of the BBox of the partitioned region/area as follows:

$$L = \frac{(Max(X) - Min(X))}{3}, \quad H = \frac{(Max(Y) - Min(Y))}{3}.$$

Table 3.4: Bounding Box computations for each partition

Area	Min Long	Min Lat	Max Long	Max Lat
South	$Min(Y)$	$Min(X)$	$Min(Y) + H$	$Max(X)$
North	$Min(Y) + 2H$	$Min(X)$	$Max(Y)$	$Max(X)$
West	$Min(Y)$	$Min(X)$	$Max(Y)$	$Max(X) + L$
East	$Min(Y)$	$Min(X) + 2L$	$Max(Y)$	$Max(X)$
SW	$Min(Y)$	$Min(X)$	$Min(Y) + H$	$Min(X) + L$
SC	$Min(Y)$	$Min(X) + L$	$Min(Y) + H$	$Min(X) + 2L$
SE	$Min(Y)$	$Min(X) + 2L$	$Min(Y) + H$	$Max(X)$
CW	$Min(Y) + H$	$Min(X)$	$Min(Y) + 2H$	$Min(X) + L$
Center	$Min(Y) + H$	$Min(X) + L$	$Min(Y) + 2H$	$Min(X) + 2L$
CE	$Min(Y) + H$	$Min(X) + 2L$	$Min(Y) + 2H$	$Max(X)$
NW	$Min(Y) + 2H$	$Min(X)$	$Max(Y)$	$Min(X) + L$
NC	$Min(Y) + 2H$	$Min(X) + L$	$Max(Y)$	$Min(X) + 2L$
NC	$Min(Y) + 2H$	$Min(X) + 2L$	$Max(Y)$	$Max(X)$

This computation of partitions can be applied to all objects defined by a geolocation bounding box, from continent level to place level as shown in Figure 3.5. For example, using the BBox of China (Country Level), that is {73.4994137, 18.1535216, 134.7728100, 53.5609740}, the BBox of Southern China (Part of Country Level) will be computed as {73.4994137, 18.1535216, 93.9238791, 53.5609740}.

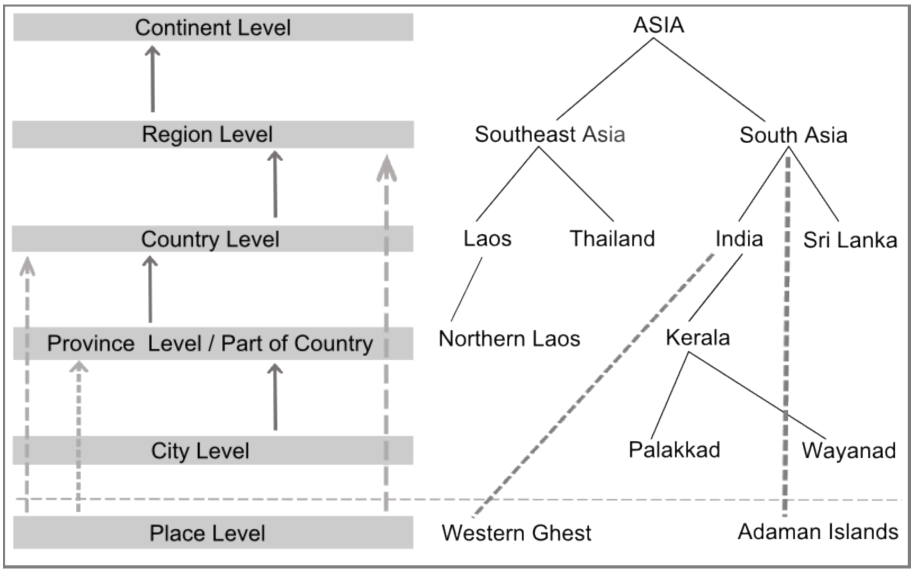


Figure 3.5: The world regional hierarchy

3.4 BioKET Data Visualization on GeoCAT

GeoCAT (Geospatial Conservation Assessment Tool)[2] is an interactive tool to plot data in biodiversity and simply use for analyzing species specimens based on "The extent of occurrence (EOO) and the area of occupancy (AOO)" [Bachman 2011] in particular geolocation on the earth. Currently the GeoCAT supports the synchronization and visualization species data from GBIF, Flickr, iNaturalist and Picasa. On the other hand, it allows to import also from datasets with standard formats such CSV, GeoCAT and DWC. Following the GeoCAT documents and the user guideline, we generated a dataset that contained mandatory data based on the GeoCAT template (Table 3.5) for CSV import, and we visualized the BioKET data simply as shown in Figure 3.6.

3.5 BioKET Plant Ontology Mapping Concept

The BioKET data warehouse contains a variety of plant descriptions. Some descriptions used taxonomic terms and some used improper terms (general terms) to describe a plant structure, i.e. different plant descriptions but the same meaning. To reduce the number of duplicated descriptions, we proposed to perform mapping the

[2] GeoCAT Platform at http://geocat.kew.org

Table 3.5: GeoCAT data template

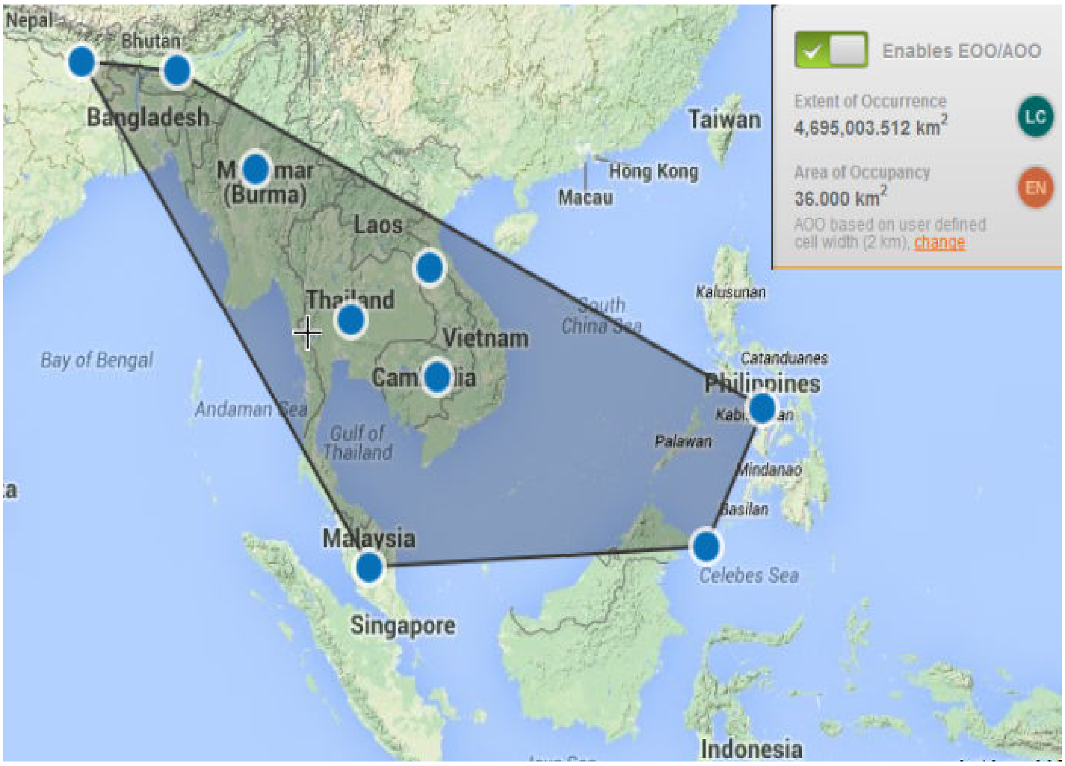
Index	Column Name	Instance
1	basisOfRecord	Observation
2	catalogNumber	BioKET000001
3	catalogue_id	1
4	changed	True
5	collectionCode	1
6	collector	BioKET.MinD
7	coordinateUncertaintyText	100
8	coordinateuncertaintyinmeters	1000
9	country	Laos
10	county	LA
11	eventDate	03/03/2014
12	identifiedBy	ID1
13	institutionCode	NUOL-I3S
14	latitude	17.6384066
15	locality	N/A
16	longitude	105.2194808
17	occurrenceDetails	i3s.unice.fr/~pasquier/BioKET/Species/1
18	occurrenceRemarks	In Khammoune province, Laos
19	recordSource	BT 55 (biotik.org)
20	scientificname	Acer laurinum
21	stateProvince	Khammoune province
22	verbatimElevation	100

terms with terms of Plant Ontology (PO), Phenotypic Quality Ontology (PATO), and Plant Trait Ontology (TO) by following steps as shown in Figure 3.7.

First, get started to check a term in files (PO,PATO,TO) by using script languages or manual techniques. If the term can be found, then map it with an ontology term including term IRI (Internationalized Resource Identifier) and parent term IRI. If the term cannot be found, then check it with BioPortal if it is found then map the term, if it is not found then set a undefined term.

For example, a plant description "Evergreen tree up to 12 m tall" contains two terms "Evergreen tree" and "up to 12 m tall". The terms "Evergreen tree" and "up to 12 m tall" are both found in the PATO file and then map both terms with their IRI and parent IRI which IRI and parent IRI of "Evergreen tree" are "http://purl.obolibrary.org/obo/PATO_0001733" and "http://purl.obolibrary.org/obo/PATO_0001729", and IRI and parent IRI of "up to 12 m tall" are "http://purl.obolibrary.org/obo/PATO_0001733" and "http://purl.obolibrary.org/obo/PATO_0000119" respectively. In case they are both not found at all in PATO, PO, TO, and BioPortal then set both terms to undefined term with none IRI and parent term IRI.

For each completed process, the term and a value pair of term IRI and parent term IRI will be transferred and stored sequentially into Value Concept and Abstract Concept. The BioKET plant ontology mapping concept and processes were depicted



☑ Enables EOO/AOO

Extent of Occurrence
4,695,003.512 km^2

Area of Occupancy
36.000 km^2
AOO based on user defined cell width (2 km), change

Figure 3.6: The BioKET data visualization on the GeoCAT platform

in Figure 3.8.

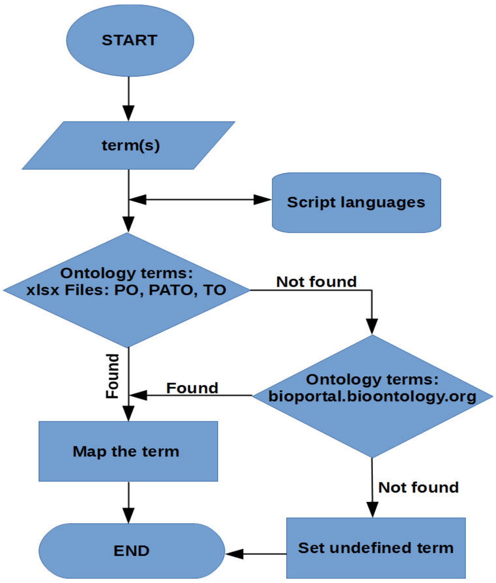
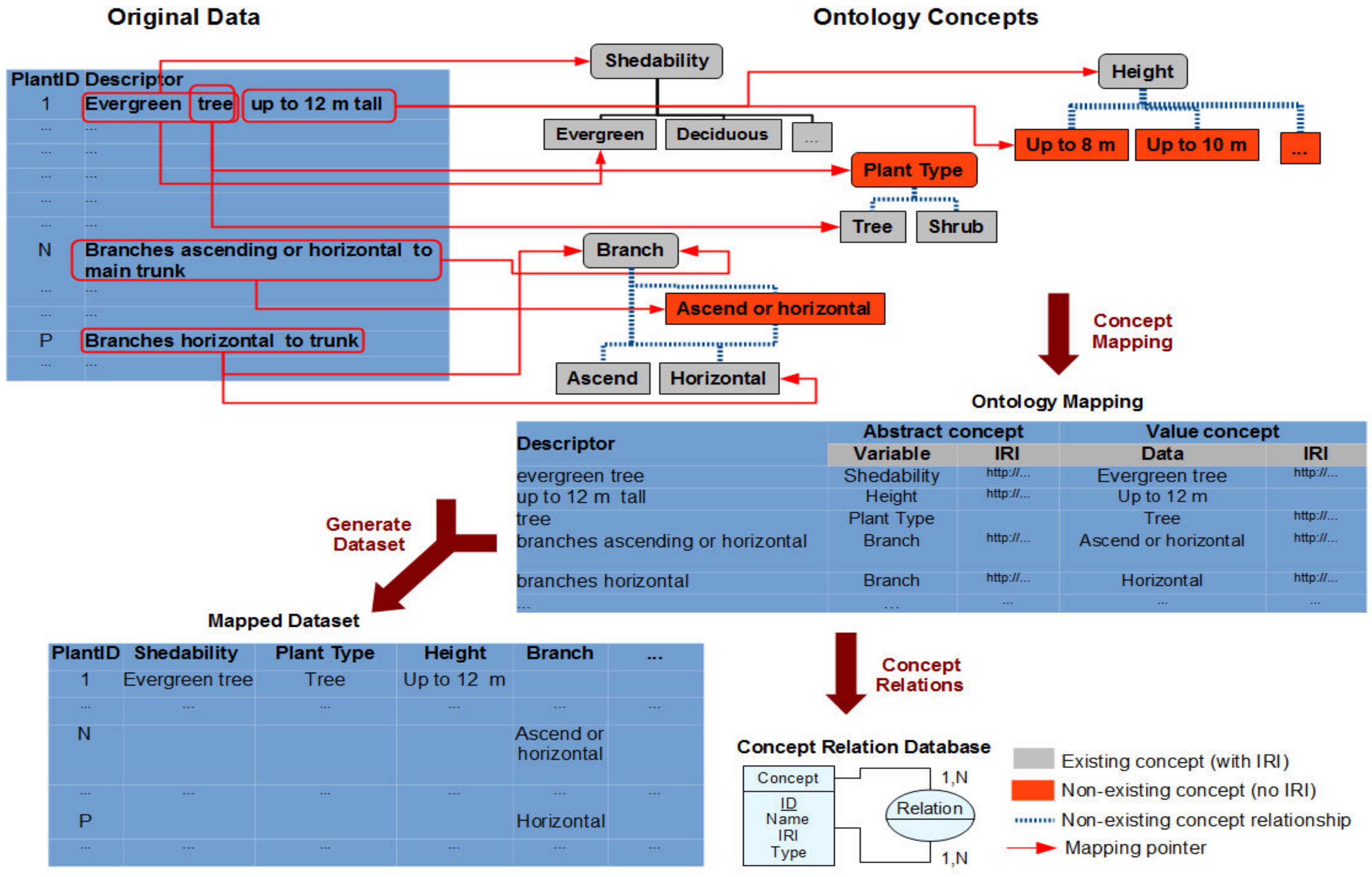


Figure 3.7: The BioKET plant ontology mapping process



Ontology Mapping

Descriptor	Abstract concept		Value concept	
	Variable	IRI	Data	IRI
evergreen tree	Shedability	http://...	Evergreen tree	http://...
up to 12 m tall	Height	http://...	Up to 12 m	
tree	Plant Type		Tree	http://...
branches ascending or horizontal	Branch	http://...	Ascend or horizontal	http://...
branches horizontal	Branch	http://...	Horizontal	http://...
...	...	...	...	...

Mapped Dataset

PlantID	Shedability	Plant Type	Height	Branch	...
1	Evergreen tree	Tree	Up to 12 m		
...	...	...	...	...	...
N				Ascend or horizontal	
...	...	...	...	...	...
P				Horizontal	
...	...	...	...	...	...

Figure 3.8: The BioKET plant ontology mapping concept

Finally the BioKET data warehouse was designed and built with 35 fact tables and 26 dimension tables. The data integration task was integrated the four sources (BIOTIK, BRAHMS, NAPIS, IUCN Red List) and Google Maps. The BioKET data was visualized on GeoCAT, and plant ontology mapping concept was demonstrated. The BioKET dataset and experimental results will be presented in the next chapter.

Chapter 4

Experiments and Results

This chapter focuses on experimental analyses of a dataset constructed from the BioKET Data Warehouse. These experiments were conducted using biclustering and association rule extraction methods that are classical descriptive data mining approaches. We present experimentation design, evaluation of performances of the application of Weka classical data mining tools and of the Galois closure based TFIST approach in terms of execution times and memory usage, and the evaluation of extracted patterns with regards to the literature of the application domain.

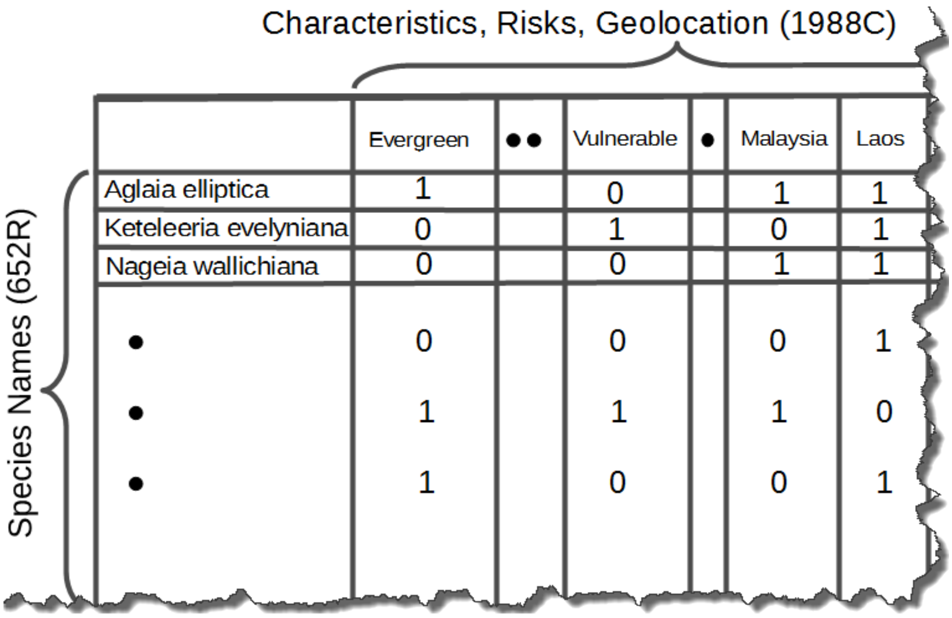
Contents

4.1 Experimentation Design

For these experiments, a dataset containing information on 652 species was constructed from the BioKET Data Warehouse using BioKET Query Scripts (see Appendix C, listing C.1). This dataset, containing 652 rows and 1988 attributes, includes three categories of information on plants as shown in Figure 4.1.

The first category corresponds to 1826 binary attributes describing morphological and environmental properties. The second category corresponds to 9 binary attributes describing risk status of species: Lower risk, Endangered, Least concern, Vulnerable, Critically Endangered, Rare, Data deficient, Rare & Threatened, Possibly extinct. The last category corresponds to 153 binary attributes describing geolocation of species at different levels: Continent, Region, Country, Part of, Province, City and Place.

The experiments were conducted on a Dell PowerEdge R710 server with 2 Intel Xeon X5675 processors at 3.06 GHz, each possessing 6 cores, 12 MB cache memory, 24 GB of DDR3 RAM at 1333 MHz and 2 Hot Plug SAS hard disks of 600 GB at 15000 rounds/min with RAID 0 running under the 64 bits CentOS Linux operating system.



	Evergreen	●●	Vulnerable	●	Malaysia	Laos
Aglaia elliptica	1		0		1	1
Keteleeria evelyniana	0		1		0	1
Nageia wallichiana	0		0		1	1
●	0		0		0	1
●	1		1		1	0
●	1		0		0	1

Figure 4.1: The BioKET dataset schema

4.2 Frequent Patterns Extraction

The dataset was analyzed using the Weka implementation of Apriori and the TFIST approach both written in Java language. The Apriori algorithm generates association rules for user-defined minimum support and confidence thresholds [Agrawal 1994]. The TFIST approach, that is based on the frequent closed itemsets framework [Mondal 2012], extracts minimal covers of conceptual association rules, (i.e. condensed representations of association rules) and biclusters jointly.

Apriori is based on the subset lattice framework. In the subset lattice, nodes represent all possible combinations of variable values in the dataset and edges are inclusion relationships between these nodes, i.e., depicting inclusion relationships between two sets of variable values. Association rules generated by Apriori are conditional rules with the form $\{V_1 \longrightarrow V_2, \mathit{support}, \mathit{confidence}, \mathit{lift}\}$ where V_1 and V_2 are sets of variable values (characteristics), with $V_1 \cap V_2 = \emptyset$. Statistical measures computed for each rule are:

- $\mathit{support} = P(V_1 \cup V_2)$ (or $\mathit{count}(V_1 \cup V_2) = |I_1|$ if given as an absolute number) evaluates the scope, or weight, of the rule in the dataset. It corresponds to the proportion of instances containing V_1 and V_2 among all instances.

- $\mathit{confidence} = \frac{P(V_1 \cup V_2)}{P(V_1)}$ evaluates the precision of the rule. It corresponds to the proportion of instances containing V_2 among those containing V_1. Rules with $\mathit{confidence} = 1$, that have no counter-example in the dataset, are called *exact* rules. Rules with $\mathit{confidence} < 1$ are called *approximate* rules.

- $\mathit{lift} = \frac{P(V_1 \cup V_2)}{P(V_1)P(V_2)}$ corresponds to the correlation between occurrences of V_1 and V_2:

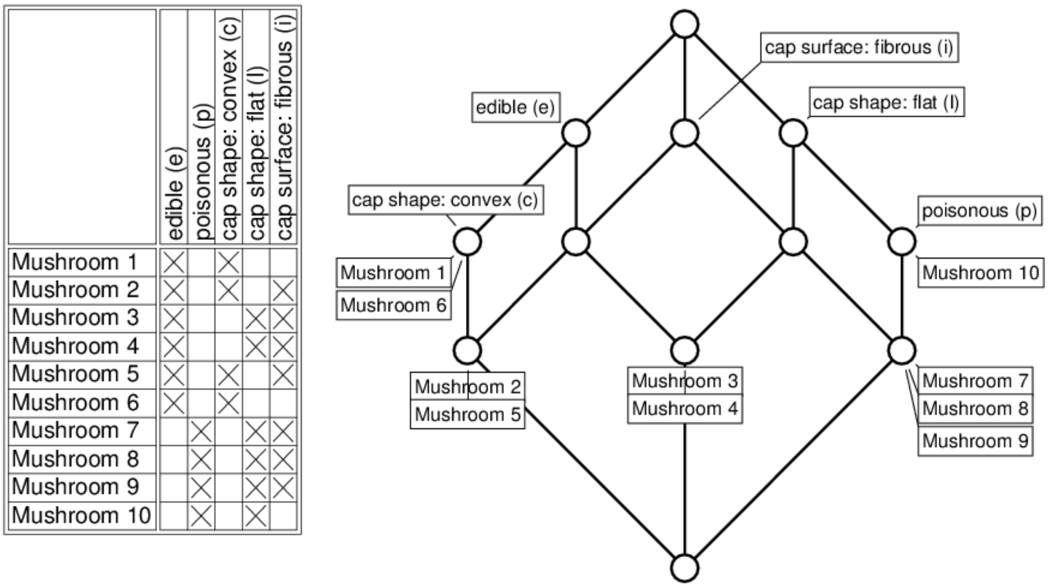
- *lift* > 1 means there is positive correlation between V_1 and V_2,
- *lift* $= 1$ means V_1 and V_2 are independent,
- *lift* < 1 means there is a negative correlation between V_1 and V_2.

Association rules are extracted from the dataset given two user-defined threshold parameters: *minsupport*, that corresponds to the minimal number (proportion) of supporting instances required for a rule to be considered valid, and *minconfidence*, that corresponds to the minimal value of *confidence* required for a rule to be considered valid. Only rules with *support* $\geq$ *minsupport* and *confidence* $\geq$ *minconfidence* are generated. From the viewpoint of the subset lattice, the association rule $V_1 \longrightarrow V_2$ is constructed from the two nodes corresponding to V_1 and $V_1 \cup V_2$ given their *support* computed from the dataset.

The frequent closed itemsets framework is related to concept lattices that are theoretical structures defined according to the Galois connection of a finite binary relation. Given a set of instances (objects) described by a list of properties (variables values), the concept lattice is a hierarchy of concepts in which each concept associates a set of instances, called *extent*, sharing the same value for a certain set of properties, called *intent*. Concepts are partially ordered in the lattice according to the inclusion relation: Each sub-concept in the lattice contains a subset of the instances and a superset of the properties in the related concepts above it. In Figure 4.2, an example dataset (left) and the corresponding concept lattice (right) are depicted. This dataset contains 10 instances (Mushroom 1 to 10) and, for each of them, 5 binary properties (Edible, Poisonous, Cap shape:convex, Cap shape:flat and Cap surface:fibrous). In the binary matrix representing the dataset, an 'X' means that the mushroom corresponding to the row possesses the binary property corresponding to the column, and an empty cell means that the mushroom doesn't possess this binary property. The concept lattice generated form this dataset contains 12 nodes, each of them corresponding to a concept, and edges depict inclusion relations between the intent and the extent of the linked concepts, i.e., the relationships between concepts and their sub-concepts. In this minimal representation of the lattice, properties are inherited from sup-concepts and instances are inherited from sub-concepts. For instance, the left-most upper node depicts the concept {{Mushroom1, Mushroom2, Mushroom5, Mushroom6}, {Cap shape:convex, Edible} and the left-most lower node depicts the concept {{Mushroom2, Mushroom5}, {Cap shape:convex, Edible, Cap surface:fibrous}.

In data mining, concept lattices serve as a theoretical framework for the efficient extraction of non-redundant loss-less condensed representations of association rules and hierarchical conceptual biclustering.

Conceptual biclusters are clusters with the form $\{V_N, I_M\}$ where V_N is a set of variable values (properties) and I_M is the maximal set of instances (species) possessing all properties in V_N. In other words, a bicluster is a sub-matrix associating a subset of rows and a subset of columns such that all these rows have a similar value for each of these columns. Conceptual biclusters are partially ordered according to the inclusion relation and form a lattice: the concept lattice. This hierarchical



	edible (e)	poisonous (p)	cap shape: convex (c)	cap shape: flat (l)	cap surface: fibrous (i)
Mushroom 1	╳		╳		
Mushroom 2	╳		╳		╳
Mushroom 3	╳			╳	╳
Mushroom 4	╳			╳	╳
Mushroom 5	╳		╳		╳
Mushroom 6	╳		╳		
Mushroom 7		╳		╳	╳
Mushroom 8		╳		╳	╳
Mushroom 9		╳		╳	╳
Mushroom 10		╳		╳	

Figure 4.2: An example concept lattice

organization allows to explore groups of instances (species) and properties (characteristics) at different levels of abstraction: the highest biclusters in the lattice regroup a large number of properties shared by small groups of instances; the lowest biclusters regroup small set of properties that are common to large group of instances.

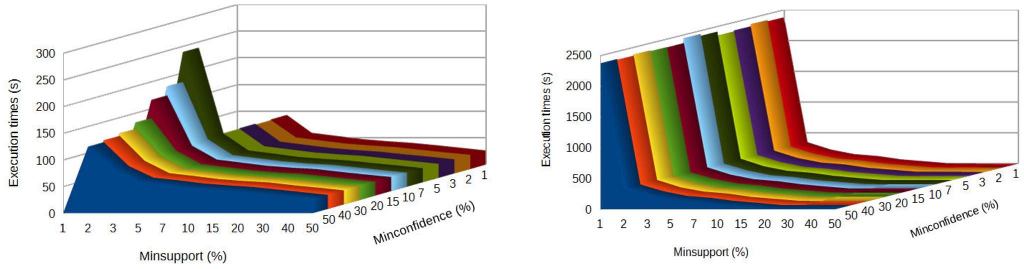
Conceptual association rules are conditional rules with the form $\{V_1 \longrightarrow V_2,\ I_1,$ *support*, *confidence*, *lift*$\}$ where V_1 and V_2 are sets of variable values (characteristics), with $V_1 \cap V_2 = \emptyset$ and I_1 is the set of instances (species) supporting the rule, i.e., the list of instances possessing all variable values in $V_1 \cup V_2$.

TFIST extracts simultaneously conceptual biclusters and association rules according to two parameters: The *minsupport* threshold, that corresponds to the minimal number (proportion) of supporting instances required for a rule to be considered valid and a bicluster to be considered relevant, and the *minconfidence* threshold, that corresponds to the minimal value of *confidence* required for a rule to be considered valid.

4.3 Experimental Results

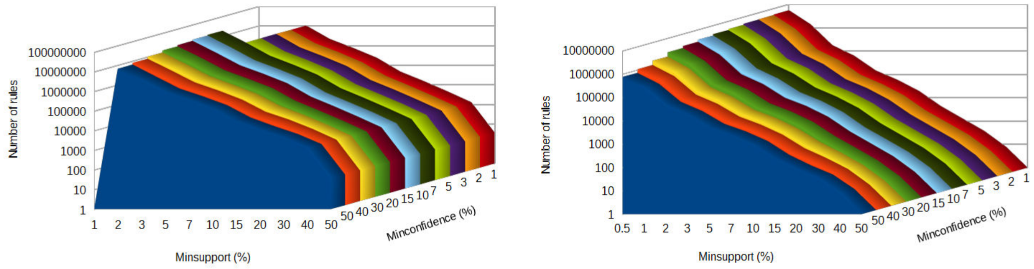
For each experiment, i.e., for a specific set of parameter values, ten runs were performed and, execution times and memory usage measures are the averages of these runs.

The *minsupport* and *minconfidence* thresholds were both varied between 50% and 1%. Figure 4.3 shows that the peak of Apriori-Weka's execution times is roughly 240 times at the *minsupport* threshold 1% and the *minconfidence* threshold 5%, whereas the peaks of TFIST's execution times are mostly closed to 2500 times at the *minsupport* threshold 1%. Figure 4.4 and 4.5 respectively show that TFIST was able to generate rules at any variants of the *minsupport* threshold. Memory usage of TFIST goes up to about 2 GB for *minsupport* threshold of 3% and less. Apriori-Weka was not able to extract association rules for a *minsupport* threshold of 1%, and for a *minsupport* threshold of 2% with *minconfidence* threshold less than 7%. Memory usage of Apriori-Weka goes up to around 600 MB for *minsupport* is equal to 2% and *minconfidence* is equal to 7%.



Apriori-Weka: Execution times TFIST: Execution times

Figure 4.3: BioKET experiments: execution times



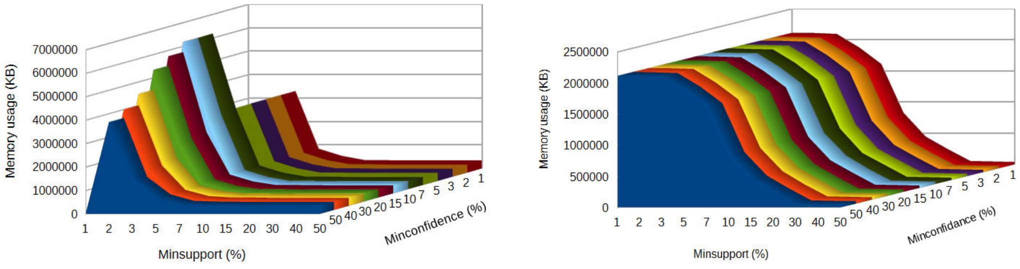
Apriori-Weka: Number of rules TFIST: Number of rules

Figure 4.4: BioKET experiments: number of rules

In Figure 4.6, the *minsupport* threshold was varied between 50% and 0.5%. Unfortunately, Apriori-Weka could not handle to process for pattens extraction, whereas TFIST was able to generate the numbers of patterns. For the *minsupport* threshold of 0.5%, the peak in the number of patterns extracted is roughly 100000 including exact rules, biclusters, and generators.

4.4 Extracted Patterns Evaluation

In this section, we present some interesting conceptual association rules obtained from TFIST. We would like to stress that these results mainly depend on the data collected within BioKET, which, as far as we know, is the only data warehouse consolidating different biodiversity information sources. These rules make it possible



Apriori-Weka: Memory usage

TFIST: Memory usage

Figure 4.5: BioKET experiments: memory usage

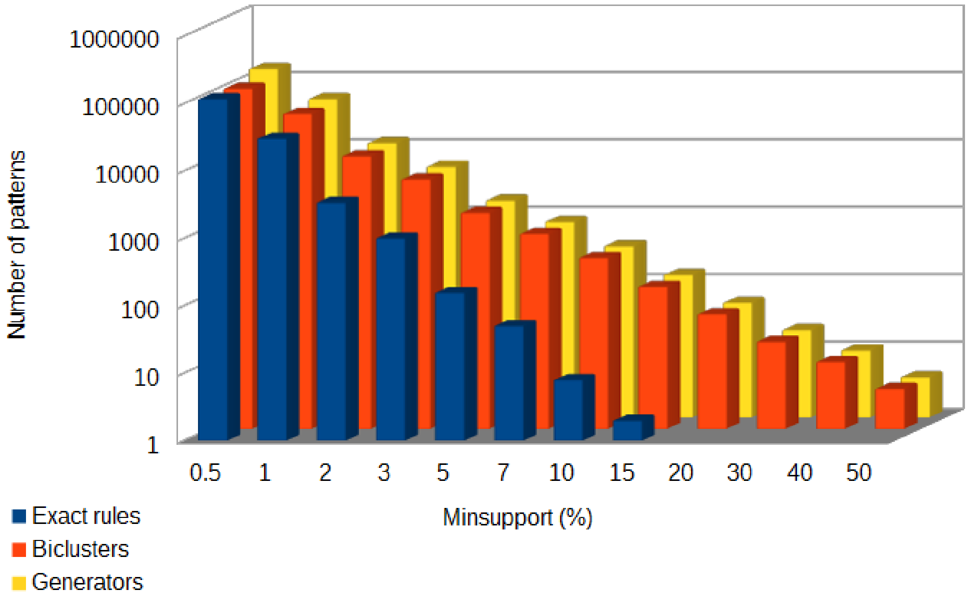


Figure 4.6: Number of patterns generated by TFIST

to estimate the risk status of a plant species according to IUCN RedList categories (*Lower Risk*, *Endangered*, *Least Concern*, *Vulnerable*, *Critically Endangered*, *Rare*, *Data Deficient*, *Rare & Threatened*, *Possibly Extinct*) with respect to their characteristics and *vice-versa*. For this experiment, the *minsupport* threshold was set to 1%, which corresponds to 6 species in the dataset, and the *minconfidence* threshold was set to 50%.

One of the obtained rules with the highest lift (11.75) is:

$$\text{INFL:pedicels up to 3 mm long}, \text{BBT:Twigs terete}, \text{INFL:axillary} \Rightarrow \text{RS:Lower Risk}. \tag{4.1}$$

According to this rule, concerning the six species with pedicels up to 3 mm long, twigs terete, and axillary inflorescence, 66,67% belong to the *lower risk* category. The six identified species are *Cratoxylum cochinchinense*, *Cratoxylum formosum*, *En-*

gelhardtia serrata, *Engelhardtia spicata*, *Irvingia malayana*, and *Knema globularia*. This result is corroborated, for example, by information from Singapore flora[1]

The following rule states, with 83.33% confidence, that a plant species classified as *Rare* has simple leaves:

$$\mathsf{RS{:}Rare} \Rightarrow \mathsf{LEAVES{:}Leaves\ simple.} \tag{4.2}$$

This rule is corroborated, for example, by [Fritsch 2011], which describes *Gaultheria paucinervia*, a new species restricted to the eastern slopes of Mt. Kinabalu in Sabah State, Borneo, Malaysia, which has been confused with *Gaultheria borneensis Stapf*, but differs in its more erect habit and larger stature, longer nonappressed leaf trichomes, purple (vs. white) fruiting calyx, and lower elevation range, among other features. *Gaultheria paucinervia* has not yet been assessed for the IUCN Red List, (but is in the Catalogue of Life: *Gaultheria paucinervia P.W. Fritsch & C.M. Bush* apparently). Besides, by taking into account the features in the geographical data source, the TFIST algorithm finds the rule:

$$\mathsf{RS{:}Rare, GEO{:}Western\ Ghats} \Rightarrow \mathsf{LEAVES{:}Leaves\ simple,} \tag{4.3}$$

which identifies species *Bentinckia condapanna*, *Drypetes malabarica*, *Glycosmis macrocarpa*, *Holigarna grahamii*, *Lasianthus jackianus*, *Pittosporum dasycaulon*, and *Vepris bilocularis*, all found in the Western Ghats.

The following rule states, with 79.59% confidence, that a plant species classified as *Vulnerable* has simple leaves:

$$\mathsf{RS{:}Vulnerable} \Rightarrow \mathsf{LEAVES{:}Leaves\ simple.} \tag{4.4}$$

This result is corroborated, for example, by [Van So 2000, Jøker 2000]. In [Van So 2000], the author discusses the applicability of the Accelerated Pioneer-Climax Series (APCS) method for restoring forests to degraded areas in Southern Vietnam using many local species such as *Hopea odorata* directly concerned by the above rule and which has been identified as *vulnerable* in the IUCN red list. Wickneswari [Ratnam 2014], instead, proposes a document which can help the readers to understand the entire life cycle of *Hopea odorata Roxb* in Malaysia, Vietnam, Cambodia, and Thailand.

The following rule, whose lift is 1.189 and whose support is 4.14%, states, with 55.1% confidence, that a plant species classified as *vulnerable* has both *glabrous* and *simple* leaves:

$$\mathsf{RS{:}Vulnerable} \Rightarrow \mathsf{LEAVES{:}glabrous, LEAVES{:}Leaves\ simple.} \tag{4.5}$$

Indeed, [Rahangdale 2014], proposing a deep and comprehensive botanical study of two rock outcrops in India, corroborates this rule.

[1] URL: `http://florasingapura.com/Home.php` .The aim of this site is to to bridge the gap between the terse technical descriptions of plants found in various botanical text books and what is observed in the Singapore forests.

Another interesting rule with a support of 3.37% and a lift of 1.07, states, with 59.46% confidence, that a plant species classified as having a *lower risk* has alternate leaves:

$$\mathsf{RS{:}Lower\ Risk} \Rightarrow \mathsf{LEAVES{:}alternate.} \tag{4.6}$$

This result is corroborated, for example, by results obtained by Craenel [De Craene 2009]. Species concerned include *Aglaia elliptica, Aphanamixis polystachya*, and *Prunus arborea.* As seen for Rule 4.3, the integration of geolocation information with multiple heterogeneous biological data can show common properties related to species with a specific risk status and/or in a specific area. For instance, the following rule with a lift of 4.26 states that 88.9% of species having a *lower risk* in the Indochina geographic region (i.e., 8 species) have leaves with entire margin:

$$\mathsf{RS{:}Lower\ Risk, GEO{:}Indochina} \Rightarrow \mathsf{LEAVES{:}Margin\ entire.} \tag{4.7}$$

Another example of such rule is the following, showing that 88.2% of *endangered* species in Western Ghats have alternate leaves:

$$\mathsf{RS{:}Endangered, GEO{:}Western\ Ghats} \Rightarrow \mathsf{LEAVES{:}Alternate.} \tag{4.8}$$

This rule, whose lift is 1.96, concerns 15 species. Such patterns can help comparisons between different geographical areas, at different levels of abstraction. For instance, considering the Malaysia geographic region, a part of Indochina, only 61.5% of species having a *lower risk* have leaves with entire margin as stated by the following rule, whose lift is 2.95 and which concerns 8 species:

$$\mathsf{RS{:}Lower\ Risk, GEO{:}Malaysia} \Rightarrow \mathsf{LEAVES{:}Margin\ entire.} \tag{4.9}$$

If we consider the Agasthyamalai area, lying at the extreme southern end of the Western Ghats mountain range along the western side of Southern India, we can see from the following rule that only 50% of *endangered* species in this area have alternate leaves, whereas the percentage is of 88.2% in the whole Western Ghats:

$$\mathsf{RS{:}Endangered, GEO{:}Agasthyamalai} \Rightarrow \mathsf{LEAVES{:}Alternate.} \tag{4.10}$$

This rule, which has a lift of 4.27, concerns 10 species.

All the above rules have been constructed from the BioKET data warehouse presented in the previous chapter. Although some of the species are not yet included in the IUCN red list, combining information from different data sources allowed us to infer their risk status using the rules constructed by TFIST. This is the case, e.g., for the species related to Rule 4.3, with the sole exception of *Bentinckia condapanna*, whose risk category is explicitly in IUCN. Indeed, *Glycosmis macrocarpa*'s taxon has not yet been assessed for the IUCN Red List, but is listed in the Catalogue of Life as *Glycosmis macrocarpa Wight.* The same holds for *Drypetes malabarica* (in the Catalogue of Life as *Drypetes malabarica (Bedd.) Airy Shaw*), *Lasianthus jackianus* (in the Catalogue of Life as *Lasianthus jackianus Wight*), *Pittosporum dasycaulon* (in the Catalogue of Life as *Pittosporum dasycaulon Miq*), and *Vepris bilocularis* (in the Catalogue of Life as *Vepris bilocularis (Wight & Arn.) Engl.*).

On the other hand, all the above rules were summarized into the following tables. Table 4.1 shows a significant correlation between threatened plant status and plant features. For example, if a plant species is in the rare category, then with 83.33% confidence implies that the plant species has a feature of simple leaf. Table 4.2 shows a correlation among plant features, threatened plant status, and plant location. For example, if 88.90% of plant species are in the lower risk category and found in Indochina (Cambodia, Laos, Vietnam, Myanmar, Thailand, Malaysia, Singapore), then they have entire margin leaves.

Table 4.1: The correlation of threatened plant status and plant features

Classified in:	**Confidence value:**	**Plant species has:**
Rare category	83.33%	simple leaf
Vulnerable category	79.59%	simple leaf
Lower Risk category	59.46%	alternate leaf
Vulnerable category	55.10%	glabrous and simple leaves

Table 4.2: The correlation of threatened plant status, plant location and features

Percentage of Plant species:	**Classified and Found in:**	**Plant species have:**
88.90%	Lower Risk category and in Indochina (Cambodia, Laos, Vietnam, Myanmar, Thailand, Malaysia, Singapore)	leaves with entire margin
88.20%	Endangered category and in India, Western Ghats	alternate leaves
61.50%	Lower Risk category and in Malaysia	leaves with entire margin
50.00% [a]	Endangered category and in India, Agasthyamalai	leaves with entire margin

[a] This value is a part of 88.20% plant species in the whole Western Ghats, India because Agasthyamalai is in Western Ghats

Chapter 5

Conclusions and Further Work

This final chapter summarizes what has been done during this research work and presents perspectives from the viewpoint of the extension of this work and related applications.

Contents

5.1 Conclusions

Biodiversity refers to the variety and abundance of living organisms (plants, animals and other living beings) in a particular area or region. In an ecosystem, each species is part of the web of life and has a fundamental role in the circle of life. Hence, all species interact and depend upon one another for what each supplies, e.g., food, oxygen, shelter, and soil enrichment. Maintaining biodiversity of species in ecosystems is thus a necessity to preserve the web of life, and according to the biologist Edward O. Wilson, known as the “father of biodiversity”: “It is reckless to suppose that biodiversity can be diminished indefinitely without threatening humanity itself” [Wilson 1992].

Biodiversity loss is a major issue for all living species and preserving biological diversity in ecosystems requires to analyze and understand the parameters of this loss. This is a complex task for scientists as information from many domains (biology, geography, environment, pollution, etc.) must be considered and linked. This information can be categorized into two types: Knowledge, i.e., abstract concepts and relationships between them, that can be general or specific to a peculiar domain, and data, i.e., known and inventoried facts on concrete objects, which are described using knowledge concepts represented in ontologies. An important part of this information is available through Web portals and repositories, but this information is most cases scattered, weakly documented and in formats that hinder their integration and analysis, and thus the discovery of new information. The definition of a methodology to integrate and structure data and knowledge into an unified information system, that can serve as an integrated community resource, is therefore a major concern for biodiversity and environment studies [Parr 2012].

Data mining regroups theories, concepts and techniques for the analysis of large sets of weakly-structured heterogeneous data. The pre-processing, modeling and post-processing approaches proposed in this domain are thence adequate to both integrate and analyze biodiversity and environment data and knowledge. This is the core of the integration process to construct a biodiversity information system from which different datasets can be generated according to the specificities of the application or the analysis domain. This process allows to integrate and link information from different domains and of different types (e.g., text, images, spatial data) and to extract them together in data mining patterns and models without the requirement of peculiar treatments.

The BioKET data warehouse was obtained by consolidation of a number of heterogeneous data sources on biodiversity. As far as we know, this is the first data warehouse containing that amount of heterogeneous data which can be used for conducting data-intensive studies about biodiversity. For this research, the scope of BioKET is to focus on plant data. Plant data were integrated with Google Maps Geocoding Service, and interfaces for visualizing BioKET plant data on GeoCAT and generating datasets from BioKET data were developed. We demonstrated the usefulness of BioKET by applying association rule extraction and biclustering methods, based on the Apriori and Galois closure approaches, on datasets generated from BioKET to analyze the risk status of plants endemic to Laos and Southeast Asia. The evaluation of the extracted patterns against the botanical literature shows that knowledge on plant conservation can be inferred from BioKET.

5.2 Further Work

As a continuation of this work, we plan to extend the scope of the BioKET data warehouse to other types of biodiversity data, such as zoological data that are both larger and more complex. To deal with a such big amount of data, from the viewpoint of scalability, optimization and performance, we also plan to transfer the BioKET data warehouse from relational database management system platform to a non-relational NoSQL big data management system platform.

Moreover, an application integrating the BioKET data warehouse for plant recognition and specimen census and observations, using both geolocation and data mining techniques, is under development. We illustrate below the functioning of this plant recognition and census system in the use case model. This illustration is based on interactions between four following elements, namely End-user, Plant Recognition Application (PR App), GIS System, and Plant Recognition System (PRS), as shown in Figure 5.1.

This application begins with the communication to the PR App plant recognition application of plant snapshots and geolocation data (latitude and longitude coordinates) by the end-user. The PRS plant recognition system then receives these information through the PR App, and processes subsequent requests from the PR App. Different situations must then be considered, leading to different processes,

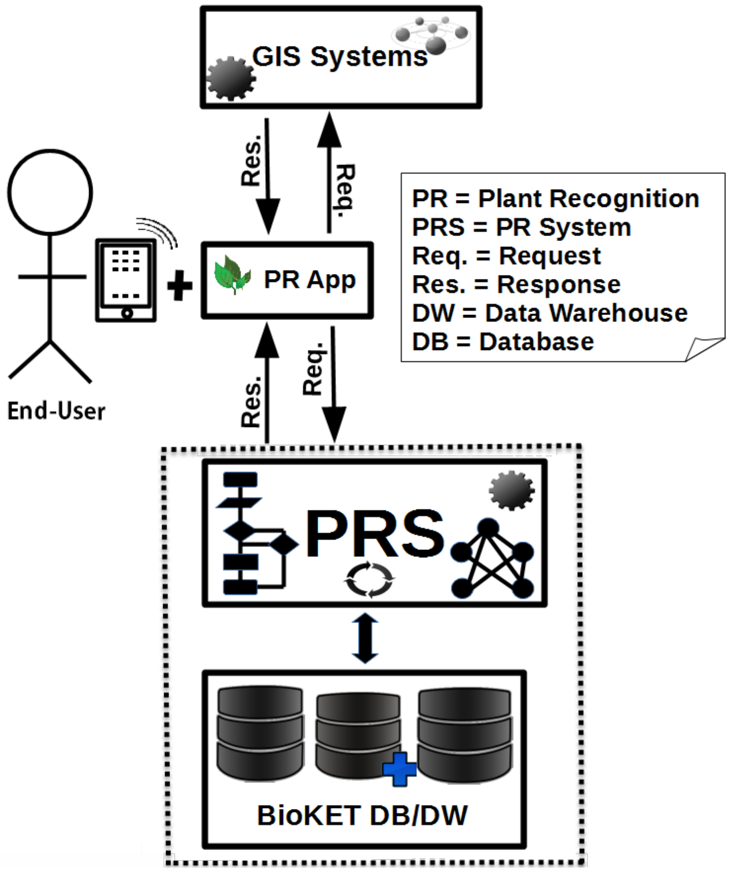


Figure 5.1: The BioKET Plant Recognition System concept

as illustrated in the three use cases described in the following. Each use case is illustrated as a workflow of ordered events, each event corresponding to an operation denoted as a function call. The use cases are depicted as collaboration diagrams, to illustrate the processing of the tasks, using the notations from the Agile Modeling collaboration diagram guideline [agi 2014].

The first situation is illustrated by the collaboration diagram depicted in Figure 5.2. The first event, initiating the workflow, is the sending by the end-user of an inquiry to the PR App in order to identify the plant subject of the picture (event 1). Then, the PR App dispatches the request to the PR Operator (event 2). The PR Operator obtains geolocation coordinates from the GIS System (event 3) and receives a result (event 4) from the PR System. The PR Operator then dispatches the result to the user through the PR App's interface (event 5). This result is constituted by a list of potential matches to identify the plant on the picture. These potential matches are determined according to similarity comparison between the picture sent and images of plants in the PR System taking into account geolocation information. Each result is assessed by a probability determined according to comparison results and the correspondence with reported geolocations of specimens of the species.

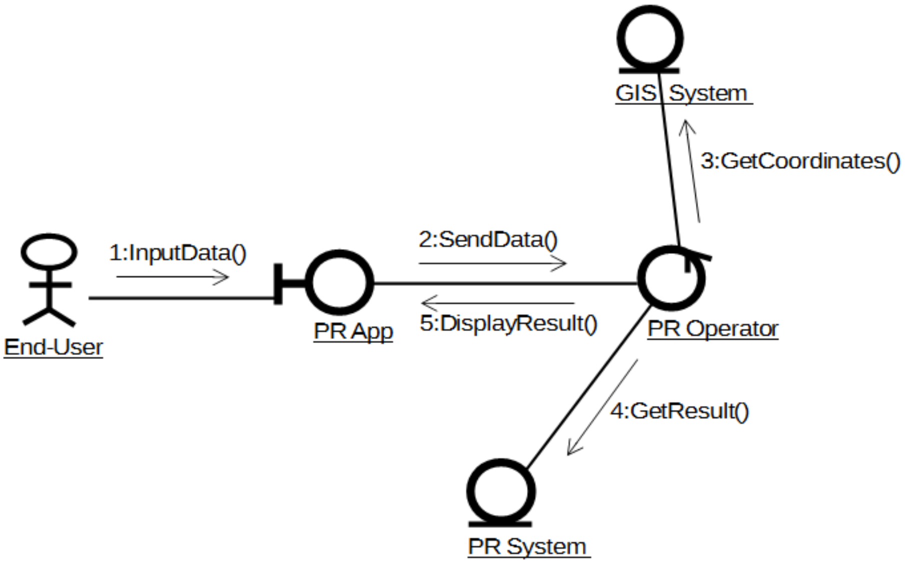


Figure 5.2: Use case 1 of the plant recognition workflow (positive match)

If the end-user validates one match among the list received, that is, he confirms the plant corresponds to one plant in the result, then the recognition process ends. This information can then be used to update the list of specimens of this species found at this geolocation position. Otherwise, that is if no result in the list is validated by the end-user or the resulting list is empty, a new process is initiated to refine the recognition process by providing more information on the plant to identify to the PR System. This new process is illustrated by the collaboration diagram depicted in Figure 5.3.

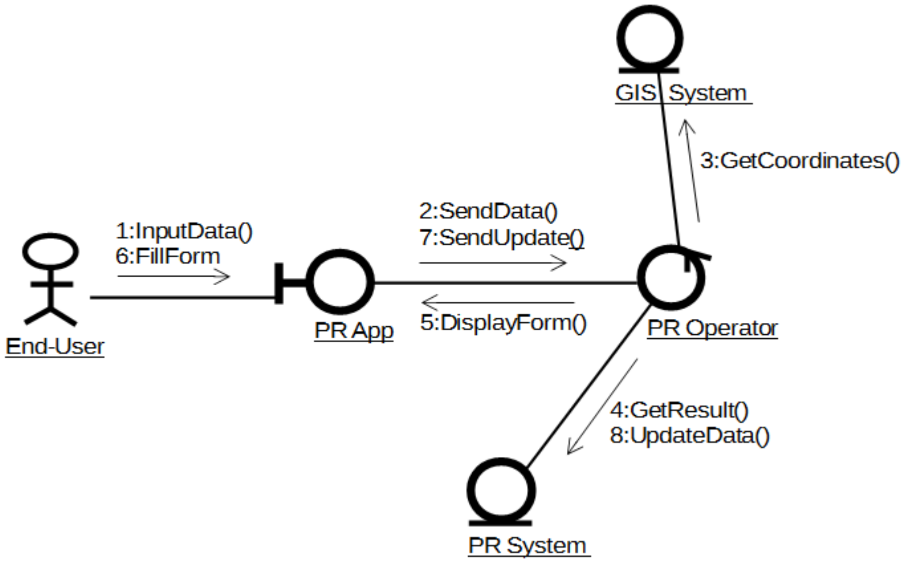


Figure 5.3: Use case 2 of the plant recognition workflow (no positive match)

First event, the end-user ends an recognition inquiry to the PR App (event 1) and the PR App dispatches the request to the PR Operator (event 2). The PR Operator obtains geolocation coordinates (event 3) and receives a result (event 4) from the PR System. We consider here the situation where no relevant match was found by the PR System. The PR Operator the dispatches a response as a feedback form to the user through the PR App's interface (event 5). This form allows the end-user to provide detailed information on the plant to recognize, such as for instance plant feature descriptors, and this form is sent to the PR App (event 6). The PR App then dispatches these feedback data to the PR Operator (event 7) which dispatches in turn a command to update data on the PR System (event 8).

The third use case corresponds to the situation where several positive matches are returned to the end-user in the result list and complementary steps are required to identify the plant among this list. This process is illustrated by the collaboration diagram depicted in Figure 5.4.

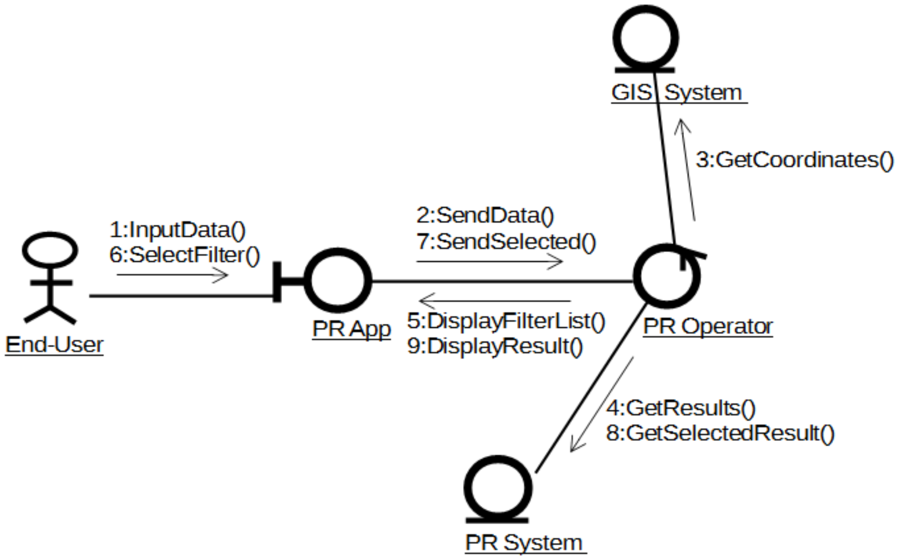


Figure 5.4: Use case 3 of the plant recognition workflow (several positive matches)

First, the end-user sends an inquiry to the PR App (event 1). The PR App dispatches the request to the PR Operator (event 2), the PR Operator obtains coordinates from GIS System (event 3) and receives results from the PR System (event 4). The list of results is then presented to the end-user through the PR App's interface with a filtering form that allows him/her to precise descriptive characteristics of the plant to recognize (event 5). The end-user completes this filtering criteria and results are sent to the PR App (event 6). The PR App dispatches these results to the PR Operator (event 7) which in turn receives itself a corresponding result from the PR System (event 8). The PR Operator then dispatches the specific result to the end-user through the PR App's interface (event 9).

The processing of geolocation information is an important part of the recognition process. Different solutions to compute the area of presence of specimens of a species, taking into account the diversity of completeness of geolocation information for the different species, can be applied. Examples of such solutions are presented in Figure 5.5.

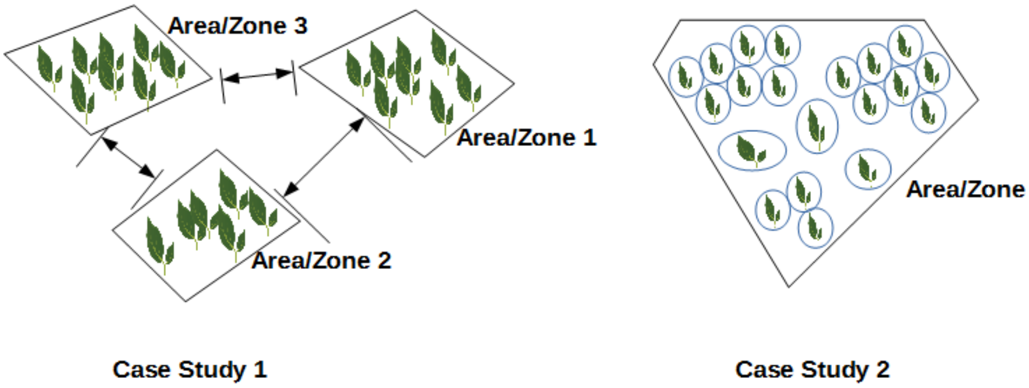


Figure 5.5: Two possible solutions to compute specific coordinates

The following solutions can be a factor for the species recognition process. The first solution by calculating the distance of different areas/zones based on coordinates (latitude and longitude) of GIS precision scales/density, the end-user can obtain the geolocation information of the different/same species: how far from the area/zone that she/he has discovered the species. The second solution if all species are located in the same area/zone, the end-user can then discover the species by using/calculating values of Bounding Box(BBox) coordinates of the area/zone to point the geolocation of the species. For example, botanists can predict and verify new speciments by simply sending coordinates of the speciments when they are in the fieldwork.

Appendix A

Resources and Ontology Domains

A.1 Biodiversity and Environment Resources

BHL: Biodiversity Heritage Library The BHL is a consortium of natural history and botanical libraries that cooperate to digitize and make accessible the legacy literature of biodiversity held in their collections and to make that literature available for open access and responsible use as a part of a global "biodiversity commons". The BHL consortium works with the international taxonomic community, rights holders, and other interested parties to ensure that this biodiversity heritage is made available to a global audience through open access principles. The BHL was found in 2005.

BioNET BioNET - the global network for taxonomy - is an international initiative dedicated to promoting the science and use of taxonomy, especially in the economically poorer countries of the world. To date the network comprises ten government-endorsed regional networks, the 'Locally Owned and Operated Partnerships' (LOOPs), encompassing institutions and 3,000 individuals in over 100 countries, and a Secretariat in the UK hosted by CABI, an international not-for-profit organization.

BIOTIK BIOTIK, stands for Biodiversity Informatics and co-Operation in Taxonomy for Interactive shared Knowledge base. Its aims is to provide data on plants in Laos, Cambodia, India and some countries in Asia.

BISE: Biodiversity Information System for Europe The BISE is a single entry point for data and information on biodiversity in the EU. Bringing together facts and figures on biodiversity and ecosystem services, it links to related policies, environmental data centers, assessments and research findings from various sources. It is being developed to strengthen the knowledge base in support of the implementation of the EU biodiversity strategy and the assessment of its progress.

CBD: Convention on Biological Diversity The CBD was opened for signature on 5 June 1992 at the United Nations Conference on Environment and Development (the Rio "Earth Summit"). The Convention on Biological Diversity was inspired by the world community's growing commitment to sustainable development. The CBD represents a dramatic step forward in the conservation of biological diversity, the sustainable use of its components, and the fair and equitable sharing of benefits arising from the use of genetic resources.

ECNC: European Center for Nature Conservation The ECNC is an independent organization working for the conservation and sustainable use of Europe's nature, biodiversity and landscapes. Since its establishment in 1993 ECNC has developed a working partnership with an extensive network of organizations and institutes from all over Europe. The ECNC provides its expertise to national and regional governments, intergovernmental organizations such as the United Nations, the European Commission, the European Environment Agency and the Council of Europe, and to institutions working in financing, land use and research.

ESABII: East and Southeast Asia Biodiversity Information Initiative The ESABII was launched to pursue capacity building in taxonomy and the development of an information system on biodiversity in East and Southeast Asia in order to contribute to the promotion of biodiversity conservation and the implementation of the CBD Strategic Plan in the area.

FAOdata: FAO's Data Warehouse FAOdata brings together statistics, maps, pictures and documents on nutrition, food and agriculture from throughout the Food and Agriculture Organization of the United Nations(FAO), providing easy access, a powerful search engine and data visualizations all in one convenient location. FAOdata already unites data from 24 of our databases related to 198 countries and includes 64 Statistical datasets, 235,025 Maps, 61,714 Pictures, and 8685 Tags.

IUCN: International Union for Conservation of Nature The IUCN is the world's oldest and largest global environmental organization network. It helps the world find pragmatic solutions to our most pressing environment and development challenges. The IUCN was founded in 1948.

GBIF: Global Biodiversity Information Facility The GBIF was established by governments in 2001 to encourage free and open access to biodiversity data, via the Internet. Through a global network of countries and organizations, GBIF promotes and facilitates the mobilization, access, discovery and use of information about the occurrence of organisms over time and across the planet.

KNEU: Biodiversity Knowledge Biodiversity Knowledge is an initiative by researchers and practitioners to help all societal actors in the field of biodiversity and ecosystem services to make better informed decisions.

NBN: National Biodiversity Network The NBN is a collaborative venture in the United Kingdom committed to making biodiversity information available through various media, including on the Internet via the NBN Gateway, the data search web site of the NBN.

OBIS: The Ocean Biogeographic information System OBIS provides a portal or gateway to many datasets containing information on where and when marine species have been recorded. The datasets are integrated so you can search them all seamlessly by species name, higher taxonomic level, geographic area, depth, and time; and then map and find environmental data related to the locations.

OECD: Organization for Economic Co-operation and Development The OECD is an international economic organization of 34 countries founded in 1961 to stimulate economic progress and world trade. It is a forum of countries committed to democracy and the market economy, providing a platform to compare policy experiences, seek answers to common problems, identify good practices and coordinate domestic and international policies of its members.

TEEB: The Economics of Ecosystems and Biodiversity TEEB is a global initiative focused on drawing attention to the economic benefits of biodiversity. Its objective is to highlight the growing cost of biodiversity loss and ecosystem degradation. TEEB presents an approach that can help decision-makers recognize, demonstrate and capture the values of ecosystems & biodiversity, including how to incorporate these values into decision-making.

UNEP: United Nations Environment Program The UNEP is an international institution (a programme, rather than an agency of the UN) that coordinates United Nations environmental activities, assisting developing countries in implementing environmentally sound policies and practices. It was founded as a result of the United Nations Conference on the Human Environment in June 1972 and has its headquarters in the Gigiri neighborhood of Nairobi, Kenya. The UNEP also has six regional offices and various country offices.

ViBRANT: Virtual Biodiversity Research and Access Network for Taxonomy ViBRANT is a European Union FP7 funded project starting in December 2010 that will support the development of virtual research communities involved in biodiversity science. ViBRANT provides a more integrated and effective framework for those managing biodiversity data on the Web.

A.2 Definitions of Ontology Domains

Adverse event: Any unfavorable or unintended symptom, sign, or disease including an abnormal laboratory finding temporally associated with the use of a medical treatment or procedure that may or may not be considered related to the medical treatment or procedure. Such events can be intervention related, dose related, route related, patient related, or caused by an interaction with another drug(s) or procedure(s) [Whetzel 2011].

Algorithm: A defined procedure for solving a problem. Applied to a problem-solving procedure implemented in software to be executed by a computer [Whetzel 2011].

All: Any types.

Anatomy: A branch of biology and Medicine that studies primarily the internal structure and design of the structure of living things. It is a general term that includes human anatomy, animal anatomy (zootomy) and plant anatomy (phytotomy). Anatomy is divided into various sub specialties in some of its facets anatomy is closely related to Embryology, Histology, comparative anatomy and comparative embryology, through common roots in evolution. Anatomy is subdivided into gross anatomy (or macroscopic anatomy) and microscopic anatomy. Gross anatomy (also called topographical anatomy, regional anatomy, or anthropotomy) is the study of anatomical structures that can be seen by unaided vision with the naked eye. Microscopic anatomy is the study of minute anatomical structures assisted with microscopes, which includes histology (the study of the organization of tissues), and cytology (the study of cells). The history of anatomy has been characterized, over time, by a continually developing understanding of the functions of organs and structures in the body including the clinical understanding of how damage to these structures effects other functions in the body. Methods have also advanced dramatically, advancing from examination of animals through dissection of cadavers (dead human bodies) to technologically complex techniques developed in the 20th century including X-ray technology, Sonogram and MRI technology. Anatomy should not be confused with anatomical pathology (also called morbid anatomy or histopathology), which is the study of the gross and microscopic appearances of diseased organs [Whetzel 2011].

Behavior: The actions or reactions of an object or organism, usually in relation to the environment or surrounding world of stimuli [Whetzel 2011].

Biochemistry Study of the chemical substances and vital processes occurring in living organisms [Whetzel 2011].

Bioinformatics: To derive knowledge from computer analysis of biological data. These can consist of the information stored in the genetic code, but also experimental results from various sources, patient statistics, and scientific literature [Whetzel 2011].

Biology: Science that studies living organisms [Whetzel 2011].

Biological Process: A process that takes place within a living organism.

Biological function: An activity occurring within an organism, between organisms or among organisms and the mechanisms underlying such events [Whetzel 2011].

Biological Sequence: A single, continuous molecule of nucleic acid or protein. It can be thought of as a multiple inheritance class hierarchy. One hierarchy is that of the underlying molecule type: DNA, RNA, or protein [2013c].

Experiment: A coordinated set of actions and observations designed to generate data, with the ultimate goal of discovery or hypothesis testing [Whetzel 2011].

Environment: The external elements and conditions which surround, influence, and affect the life and development of an organism or population [Whetzel 2011].

Genomic: The complete genomic content of an organism, and possibly the full DNA sequence of that organism. It is contained in a set of chromosomes in eukaryotes, a single chromosome in bacteria, or a DNA or RNA molecule in viruses [Whetzel 2011].

Geography: The science that deals with the world and its inhabitants; a description of the earth, or a portion of the earth, including its structure, features, products, political divisions, and the people by whom it is inhabited [Whetzel 2011].

Health: Refers to a person's state of physical, mental and social well-being; usually it refers specifically to the state of being in good health, a state of complete physical, mental and social well-being, and does not consist only of the absence of disease or infirmity [Whetzel 2011].

Information: Knowledge derived from study, experience, or instruction that has been gathered or received by communication [Whetzel 2011].

Immunology: Study of the immune system and its reaction to pathogens, as well as its malfunctions (autoimmune diseases, allergies, rejection of organ transplants) [Whetzel 2011].

Lipid: An oily organic compound insoluble in water but soluble in organic solvents; essential structural component of living cells (along with proteins and carbohydrates) [Whetzel 2011].

Medicine: Branches of medical science that deal with nonsurgical techniques [Whetzel 2011].

Molecular structure: The location of the atoms, groups or ions relative to one another in a molecule, as well as the number and location of chemical bonds [2013d].

Neuroscience: A branch of science that deals with the study of the nervous system [Whetzel 2011].

Phenotype: Product of interactions between genes, and between genes and the environment [Whetzel 2011].

Provenance: Where something originated or was nurtured in its early existence [Whetzel 2011].

Proteins: A group of complex organic macromolecules composed of one or more chains (linear polymers) of alpha-L-amino acids linked by peptide bonds and ranging in size from a few thousand to over 1 million Daltons. Proteins are fundamental genetically encoded components of living cells with specific structures and functions dictated by amino acid sequence [Whetzel 2011].

Resource: Available source of wealth; a new or reserve supply that can be drawn upon when needed [Whetzel 2011].

Software Tool: A program that is employed in the development, repair, or enhancement of other programs or of hardware. Traditionally a set of software tools addressed only the essential needs during program development: a typical set might consist of a text editor, compiler, link loader, and some form of debug tool [John Daintith, February 2010, A Dictionary of Computing].

Taxonomy: Theories and techniques of naming, describing, and classifying organisms, and the study of the relationships of taxa [Whetzel 2011].

Upper: A top-level or foundation ontology that describes general concepts that are the same across all domains of knowledge.

Appendix B

List of Ontologies

Table B.1: Top-level ontologies.

Index	Ontology Name	Domain	Prefix
1	Basic Formal Ontology	upper	BFO
2	OBO relationship types (legacy)	all	OBO_REL
3	Relation ontology	all	RO

Table B.2: Domain-level ontologies.

Index	Ontology Name	Domain	Prefix
1	Biological process	biological process	GO
2	Cellular component	anatomy	GO
3	Chemical entities of biological interest	biochemistry	CHEBI
4	Molecular function	biological function	GO
5	Ontology for biomedical investigations	experiments	OBI
6	Phenotypic quality	phenotype	PATO
7	Plant Ontology	anatomy and development	PO
8	PRotein Ontology (PRO)	proteins	PR
9	Xenopus anatomy and development	anatomy	XAO
10	Zebrafish anatomy and development	anatomy	ZFA
11	Ascomycete phenotype ontology	phenotype	APO
12	Bilateria anatomy	anatomy	BILA
13	Biological Spatial Ontology	anatomy	BSPO
14	C. elegans development	anatomy	WBls
15	C. elegans gross anatomy	anatomy	WBbt
16	C. elegans phenotype	phenotype	WBPhenotype
17	Cell type	anatomy	CL
18	Chemical Information Ontology	biochemistry	CHEMINF
19	Common Anatomy Reference Ontology	anatomy	CARO
20	Dendritic cell	anatomy,immunology	DC_CL
21	Dictyostelium discoideum anatomy	anatomy	DDANAT
22	Drosophila development	anatomy	FBdv
23	Drosophila gross anatomy	anatomy	FBbt
24	Environment Ontology	environment	ENVO
Continued on next page ...			

Table 2: Domain-level ontologies (continued).

Index	Ontology Name	Domain	Prefix
25	Evidence codes	experiments	ECO
26	Fission Yeast Phenotype Ontology	phenotype	FYPO
27	Fly taxonomy	taxonomy	FBsp
28	Foundational Model of Anatomy (subset)	anatomy	FMA
29	Fungal gross anatomy	anatomy	FAO
30	Human developmental anatomy,abstract version, v2	anatomy	EHDAA2
31	Human disease ontology	health	DOID
32	human phenotype ontology	phenotype	HP
33	Hymenoptera Anatomy Ontology	anatomy	HAO
34	Infectious disease	health	IDO
35	Mammalian phenotype	phenotype	MP
36	Mass spectrometry	experiments	MS
37	Medaka fish anatomy and development	anatomy	MFO
38	Mosquito gross anatomy	anatomy	TGMA
39	Mosquito insecticide resistance	environment	MIRO
40	Mouse adult gross anatomy	anatomy	MA
41	Mouse gross anatomy and development, abstract	anatomy	EMAPA
42	Mouse gross anatomy and development, timed	anatomy	EMAP
43	Mouse pathology	health	MPATH
44	Ontology for General Medical Science	medicine	OGMS
45	Ontology of Adverse Events	adverse events, health	OAE
46	Ontology of Medically Related Social Entities	medicine	OMRSE
47	Pathogen transmission	health	TRANS
48	Plant Trait Ontology	phenotype	TO
49	Platynereis stage ontology	anatomy	PD_ST
50	Porifera Ontology	anatomy	PORO
51	Protein modification	proteins	MOD
52	Protein-protein interaction	experiments	MI
53	RNA ontology	molecular structure	RNAO
54	Sequence types and features	biological sequence	SO
55	Spider Ontology	anatomy	SPD
56	Subcellular anatomy ontology	anatomy	SAO
57	Suggested Ontology for Pharmacogenomics	health	SOPHARM
58	Symptom Ontology	health	SYMP
59	Systems Biology	biochemistry	SBO
60	Teleost Anatomy Ontology	anatomy	TAO
61	Teleost taxonomy	taxonomy	TTO
62	Terminology of Anatomy of Human Embryology	anatomy	TAHE
63	Terminology of Anatomy of Human Histology	anatomy	TAHH
64	Tick gross anatomy	anatomy	TADS
65	Uber anatomy ontology	anatomy	UBERON
66	Uber anatomy ontology, basic version	anatomy	uberon-basic
67	Units of measurement	phenotype	UO

Continued on next page ...

Table 2: Domain-level ontologies (continued).

Index	Ontology Name	Domain	Prefix
68	Vaccine ontology	health	VO
69	verteberate Homologous Organ Groups	anatomy	vHOG
70	Vertebrate Skeletal Anatomy Ontology	anatomy	VSAO
71	Zebrafish developmental stages	anatomy	ZFS

Table B.3: Application-level ontologies.

Index	Ontology Name	Domain	Prefix
1	Adverse Event Reporting Ontology	health	AERO
2	Anatomical Entity Ontology	anatomy	AEO
3	Biological imaging methods	experiments	FBbi
4	BRENDA tissue / enzyme source	anatomy	BTO
5	Cardiovascular Disease Ontology	health	CVDO
6	Chemical Methods Ontology	health	CHMO
7	eagle-i resource ontology	resources	ERO
8	Emotion Ontology	health	MFOEM
9	Event (INOH pathway ontology)	biological process	IEV
10	eVOC (Expressed Sequence Annotation for Humans)	experiments	EV
11	Exposure ontology	health	ExO
12	Gene Regulation Ontology	genomic	BOOTStrep
13	Influenza Ontology	health	FLU
14	Information Artifact Ontology	information	IAO
15	Kinetic Simulation Algorithm Ontology	algorithms	KISAO
16	Lipid Ontology	lipids	LiPrO
17	Malaria Ontology	health	IDOMAL
18	Microarray experimental conditions	experiments	MO
19	Minimal anatomical terminology	anatomy	MAT
20	Molecule role (INOH Protein name/-family name ontology)	proteins	IMR
21	NCBI organismal classification	taxonomy	NCBITaxon
22	NCI Thesaurus	health	ncithesaurus
23	Neuro Behavior Ontology	behavior	NBO
24	NIF Cell	neuroscience	NIF_Cell
25	NIF Dysfunction	neuroscience	NIF_Dysfunction
26	NIF Gross Anatomy	neuroscience	NIF_GrossAnatomy
27	NMR-instrument specific component of metabolomics investigations	experiments	NMR
28	Pathway ontology	biological process	PW
29	Plant Environmental Conditions	environment	EO
30	Protein covalent bond	proteins	RESID
31	Proteomics data and process provenance	proteins	ProPreO
Continued on next page ...			

Table 3: Application-level ontologies (continued).

Index	Ontology Name	Domain	Prefix
32	Sample processing and separation techniques	provenance	SEP
33	Taxonomic rank vocabulary	taxonomy	TAXRANK
34	The Drug Ontology	health	DRON
35	Uber anatomy ontology	anatomy	UBERON
36	Uber anatomy ontology, basic version	anatomy	uberon-basic

Appendix C

BioKET DW: SQL and VBA Scripts

Listing C.1: BioKET DW: Query Scripts

```
-- Fetch Geolocation Information
 Select vsp.SPECIES_ID,
         vsp.SPECIESNAME,vgeo.GeoLocation,vgeo.`LEVEL`,
         vgeo.LAT1,vgeo.LNG1, vgeo.LAT2, vgeo.LNG2, vgeo.Parent
From species as vsp
JOIN  species_geospatial as vsg on vsp.SPECIES_ID = vsg.SPECIES_ID
JOIN  vgeospatial as vgeo on  vgeo.Geo_ID = vsg.Geo_ID
WHERE vsp.SPECIES_ID in (SELECT spid from tmp652)
ORDER BY vsp.SPECIES_ID

-- Specie Risk Status Matrix Table
SELECT sp.SPECIES_ID, sp.SPECIESNAME,
 If(rs.CATEGORY='Lower Risk',1,'?')
 as '[RS]Lower Risk',
 If(rs.CATEGORY='Endangered',1,'?')
 as '[RS]Endangered',
 If(rs.CATEGORY='Least concern',1,'?')
 as '[RS]Least concern',
 If(rs.CATEGORY='Vulnerable',1,'?')
 as '[RS]Vulnerable',
 If(rs.CATEGORY='Critically Endangered',1,'?')
 as '[RS]Critically Endangered',
 If(rs.CATEGORY='Rare',1,'?')
 as '[RS]Rare',
 If(rs.CATEGORY='Data Deficient ',1,'?')
 as '[RS]Data Deficient ',
 If(rs.CATEGORY='Rare & Threatened',1,'?')
 as '[RS]Rare & Threatened',
 If(rs.CATEGORY='Possibly Extinct ',1,'?')
 as '[RS]Possibly Extinct '
FROM species sp
LEFT JOIN RISK_STATUS rs on sp.RISK_STATUS_ID=rs.RSID
```

```
WHERE  sp.SPECIES_ID in (SELECT tmp652.SPID from tmp652)

-- The joining of tables: Species, Family, Authors, and Synonyms
 Select vs.SPECIES_ID, vs.BARCODE, TRIM(vs.SPECIESNAME),
TRIM(vf.FAMILYNAME), TRIM(va.AUTHORNAME),
GROUP_CONCAT(vsn.SYNONYMNOTE SEPARATOR '; ') as Synonym_Name
From species as vs
INNER JOIN  species_synonyms as vssn on vs.SPECIES_ID=vssn.SPECIES_ID
INNER JOIN  synonyms as vsn on vssn.SYNONYM_ID = vsn.SYNONYM_ID
INNER JOIN  species_authors as vsa on vs.SPECIES_ID = vsa.SPECIES_ID
INNER JOIN  ‘authors‘ as va on vsa.AUTHOR_ID = va.AUTHOR_ID
INNER JOIN  family as vf on vs.FAMILY_ID = vf.FAMILY_ID
GROUP BY vs.SPECIES_ID

-- The joining of tables: Species, Family, Authors, and Habit
 Select vs.SPECIES_ID, vs.BARCODE, vs.SPECIESNAME, vf.FAMILYNAME,
va.AUTHORNAME,
GROUP_CONCAT(CONCAT(vh.HABITNOTE) SEPARATOR ', ') as HABIT_NOTE
From species as vs
INNER JOIN  species_habit as vsh on vs.SPECIES_ID=vsh.SPECIES_ID
INNER JOIN  habit as vh on vsh.HABIT_ID = vh.HABIT_ID
INNER JOIN  species_authors as vsa on vs.SPECIES_ID = vsa.SPECIES_ID
INNER JOIN  ‘authors‘ as va on vsa.AUTHOR_ID = va.AUTHOR_ID
INNER JOIN  family as vf on vs.FAMILY_ID = vf.FAMILY_ID
GROUP BY vs.SPECIES_ID

-- The joining of tables: Species, Family, Authors, and TrunkBark
 Select vs.SPECIES_ID, vs.BARCODE, vs.SPECIESNAME, vf.FAMILYNAME,
va.AUTHORNAME,
GROUP_CONCAT(vt.TRUNKBARKNOTE SEPARATOR ', ') as TRUNKBARK_NOTE
From species as vs
INNER JOIN  species_trunkbark as vst on vs.SPECIES_ID = vst.SPECIES_ID
INNER JOIN  trunkbark as vt on vst.TRUNKBARK_ID = vt.TRUNKBARK_ID
INNER JOIN  species_authors as vsa on vs.SPECIES_ID= vsa.SPECIES_ID
INNER JOIN  ‘authors‘ as va on vsa.AUTHOR_ID = va.AUTHOR_ID
INNER JOIN  family as vf on vs.FAMILY_ID = vf.FAMILY_ID
GROUP BY vs.SPECIES_ID

-- The joining of tables: Species, Family, Authors, and Leaves
 Select vs.SPECIES_ID, vs.BARCODE, vs.SPECIESNAME, vf.FAMILYNAME,
va.AUTHORNAME,
GROUP_CONCAT(vl.LEAVENOTE SEPARATOR ', ') as LEAVE_NOTE
From species as vs
INNER JOIN  species_leaves as vsl on vs.SPECIES_ID = vsl.SPECIES_ID
INNER JOIN  ‘leaves‘ as vl on vsl.LEAVE_ID = vl.LEAVE_ID
INNER JOIN  species_authors as vsa on vs.SPECIES_ID= vsa.SPECIES_ID
INNER JOIN  ‘authors‘ as va on vsa.AUTHOR_ID = va.AUTHOR_ID
INNER JOIN  family as vf on vs.FAMILY_ID = vf.FAMILY_ID
GROUP BY vs.SPECIES_ID

-- The joining of tables: Species, Family, Authors, and BBT
Select vs.SPECIES_ID, vs.BARCODE, vs.SPECIESNAME, vf.FAMILYNAME,
va.AUTHORNAME,
GROUP_CONCAT(vb.BBTNOTE SEPARATOR ', ')
as Branchs_Branchlets_Twigs_NOTE
From species as vs
INNER JOIN  species_BBT as vsb on vs.SPECIES_ID = vsb.SPECIES_ID
INNER JOIN  BBT as vb on vsb.BBT_ID = vb.BBT_ID
INNER JOIN  species_authors as vsa on vs.SPECIES_ID= vsa.SPECIES_ID
INNER JOIN  ‘authors‘ as va on vsa.AUTHOR_ID = va.AUTHOR_ID
INNER JOIN  family as vf on vs.FAMILY_ID = vf.FAMILY_ID
GROUP BY vs.SPECIES_ID

-- The joining of tables: Species, Family, Authors, and INFL
Select vs.SPECIES_ID, vs.BARCODE, vs.SPECIESNAME, vf.FAMILYNAME,
va.AUTHORNAME,
GROUP_CONCAT(vin.INFLNOTE SEPARATOR ', ')
as Inflorescences_or_Flowers_NOTE
From species as vs
INNER JOIN  species_infl as vsin on vs.SPECIES_ID = vsin.SPECIES_ID
INNER JOIN  infl as vin on vsin.INFL_ID = vin.INFL_ID
INNER JOIN  species_authors as vsa on vs.SPECIES_ID= vsa.SPECIES_ID
INNER JOIN  ‘authors‘ as va on vsa.AUTHOR_ID = va.AUTHOR_ID
INNER JOIN  family as vf on vs.FAMILY_ID = vf.FAMILY_ID
GROUP BY vs.SPECIES_ID

-- The joining of tables: Species, Family, Authors, and FruitSeeds
Select vs.SPECIES_ID, vs.BARCODE, vs.SPECIESNAME, vf.FAMILYNAME,
va.AUTHORNAME,
GROUP_CONCAT(vfs.FRUITSEEDNOTE SEPARATOR ', ') as Fruit_Seed_NOTE
From species as vs
INNER JOIN  species_fruitseeds as vsfs on vs.SPECIES_ID = vsfs.SPECIES_ID
INNER JOIN  fruitseeds as vfs on vsfs.FRUITSEED_ID = vfs.FRUITSEED_ID
INNER JOIN  species_authors as vsa on vs.SPECIES_ID= vsa.SPECIES_ID
INNER JOIN  ‘authors‘ as va on vsa.AUTHOR_ID = va.AUTHOR_ID
INNER JOIN  family as vf on vs.FAMILY_ID = vf.FAMILY_ID
GROUP BY vs.SPECIES_ID
```

```
-- The joining of tables: Species, Family, Authors, and HAEC
 Select vs.SPECIES_ID, vs.BARCODE, vs.SPECIESNAME, vf.FAMILYNAME,
 va.AUTHORNAME,
GROUP_CONCAT(vhc.HAECNOTE SEPARATOR ', ') as HABITAT_ECOLOGY_NOTE
From species as vs
INNER JOIN  species_haec as vshc on vs.SPECIES_ID = vshc.SPECIES_ID
INNER JOIN  haec as vhc on vshc.HAEC_ID = vhc.HAEC_ID
INNER JOIN  species_authors as vsa on vs.SPECIES_ID= vsa.SPECIES_ID
INNER JOIN  ‘authors‘ as va on vsa.AUTHOR_ID = va.AUTHOR_ID
INNER JOIN  family as vf on vs.FAMILY_ID = vf.FAMILY_ID
GROUP BY vs.SPECIES_ID

-- The joining of tables: Species, Family, Authors, and Exudates
 Select vs.SPECIES_ID, vs.BARCODE, vs.SPECIESNAME, vf.FAMILYNAME,
 va.AUTHORNAME,
GROUP_CONCAT(ve.EXUDATENOTE SEPARATOR ', ') as EXUDATE_NOTE
From species as vs
INNER JOIN  species_Exudates as vse on vs.SPECIES_ID = vse.SPECIES_ID
INNER JOIN  exudates as ve on vse.EXUDATE_ID = ve.EXUDATE_ID
INNER JOIN  species_authors as vsa on vs.SPECIES_ID= vsa.SPECIES_ID
INNER JOIN  ‘authors‘ as va on vsa.AUTHOR_ID = va.AUTHOR_ID
INNER JOIN  family as vf on vs.FAMILY_ID = vf.FAMILY_ID
GROUP BY vs.SPECIES_ID
```

Listing C.2: BioKET DW: Convert an HTML to a spreadsheet format

```
'Declare API for MS Office 64bit
'Private Declare PtrSafe Function GetAsyncKeyState Lib "kernel32" _
 (ByVal vkey As Long) As Integer

Declare Function GetKeyState Lib "User32" _
(ByVal vKey As Integer) As Integer
Const SHIFT_KEY = 16

Function Refresh4Next() As Boolean
    Refresh4Next = GetKeyState(SHIFT_KEY) < 0
End Function

Sub ConvertHtmlToExcel()
    Dim wb As Workbook
    Dim strFile As String
    Dim strPath As String

        'To disable popups of Excel features
    With Application
        .EnableEvents = False
        .DisplayAlerts = False
        .ScreenUpdating = False
        .Visible = False
    End With

    'Set a path of location where stored HTML files
    strPath = "D:\InputOutput\"
    strFile = Dir(strPath & "*.html")

    Do While strFile <> ""
        Do While Refresh4Next()
            DoEvents
        Loop
        Set wb = Workbooks.Open(strPath & strFile)
        strFile = Mid(strFile, 1, Len(strFile) - 5) & ".xlsx"
        'wb.SaveAs strPath & strFile, XlFileFormat.xlWorkbookNormal
        With wb
             .Worksheets(1).Name = "Sheet1"
             .SaveAs strPath & strFile, FileFormat:=xlOpenXMLWorkbook
             .Close
       End With
```

```
        Set wb = Nothing
        strFile = Dir
    Loop

        'To enable Excel features
    With Application
        .EnableEvents = True
        .DisplayAlerts = True
        .ScreenUpdating = True
        .Visible = True
    End With
End Sub
```

Listing C.3: BioKET DW: Read and Save a partial data from spreadsheet files

```
Option Explicit
Sub ReadSavePartialxlsx()

  Dim FileName As String
  Dim FileNumber As Long
  Dim PathCrnt As String
  Dim RowDestCrnt As Long
  Dim SheetDest As String
  Dim TgtValue As String
  Dim WBookSrc As Workbook
  Dim RetValue As String
  Dim Rng As Range
  Dim RowCrnt As Long

  PathCrnt = ActiveWorkbook.Path & "\Excel-files"

  'Set a sheet name for a new sheet
  SheetDest = "CommonName"
  RowDestCrnt = 2

  'Set a number to start from XLSX files e.g. 1.xlsx, 2.xlsx
  FileNumber = 1

  Do While True

    FileName = Dir$(PathCrnt & "\" & FileNumber & ".xlsx")
    If FileName = "" Then
      ' File does not exist
      Exit Sub
    End If
    Set WBookSrc = Workbooks.Open(PathCrnt & "\" & FileName)

    With WBookSrc.Worksheets("Sheet1")

    'Set a value for searching
    RetValue = "Common name"

      ' Set a range to find out the value
   Set Rng = .Columns("A:A").Find(What:=RetValue, _
   After:=.Range("A1"), LookIn:=xlFormulas, _
   LookAt:=xlPart, SearchOrder:=xlByRows, _
   SearchDirection:=xlNext, MatchCase:=False, _
   SearchFormat:=False)

        ' The entered value was found
        If Rng Is Nothing Then
        ' The entered value could not be found
          TgtValue = ""
      Else
          RowCrnt = Rng.Row
          TgtValue = .Cells(RowCrnt, "A").Value
      End If

    End With
    WBookSrc.Close SaveChanges:=False
    With Worksheets(SheetDest)
      .Cells(RowDestCrnt, "A").Value = Mid(FileName, _
      1, Len(FileName) - 5)
      .Cells(RowDestCrnt, "B").Value = TgtValue
    End With
    RowDestCrnt = RowDestCrnt + 1
    FileNumber = FileNumber + 1

  Loop
End Sub
```

Listing C.4: BioKET DW: Extract coordinates from Google Maps

```
Option Explicit
'___________________________________________________________________
'BioKET's Google GeoLocation Function
'___________________________________________________________________
Function GoogleGeoLocation(WorldData As String) As String
```

```
Dim xhrRequest As XMLHTTP60
Dim StrQuery As String
Dim domResponse As DOMDocument60
Dim ixnStatus As IXMLDOMNode
Dim ixnLat As IXMLDOMNode
Dim ixnLng As IXMLDOMNode

GoogleGeoLocation = ""

Set xhrRequest = New XMLHTTP60
StrQuery = "http://maps.googleapis.com/" & "" _
& "maps/api/geocode/xml?sensor=false&address="

StrQuery = StrQuery & Replace(WorldData, " ", "+")
xhrRequest.Open "GET", StrQuery, False
xhrRequest.send

Set domResponse = New DOMDocument60
domResponse.LoadXML xhrRequest.responseText
Set ixnStatus = domResponse.SelectSingleNode("//status")
If (ixnStatus.Text <> "OK") Then
    Exit Function
End If

Set ixnLat = domResponse.SelectSingleNode _
("/GeocodeResponse/result/geometry/location/lat")

Set ixnLng = domResponse.SelectSingleNode _
("/GeocodeResponse/result/geometry/location/lng")

GoogleGeoLocation = ixnLat.Text & ", " & ixnLng.Text

End Function
```

Listing C.5: BioKET DW: Extract Google Bounding Box coordinates

```
Option Explicit
'_______________________________________________________________
'BioKET's Google Bounding Box Function
'_______________________________________________________________
Function GoogleGeoBBox(WorldData As String) As String

Dim xhrRequest As XMLHTTP60
Dim sQuery As String
Dim domResponse As DOMDocument60
Dim ixnStatus As IXMLDOMNode
Dim southwestLat As IXMLDOMNode
Dim southwestLng As IXMLDOMNode
Dim northeastLat As IXMLDOMNode
Dim northeastLng As IXMLDOMNode
Dim strAddressEncode As String

GoogleGeoBBox = ""
strAddressEncode = URLEncode(WorldData)

Set xhrRequest = New XMLHTTP60
sQuery = "http://maps.googleapis.com/maps/" & "" _
& "api/geocode/xml?sensor=false&address="
sQuery = sQuery & strAddressEncode

xhrRequest.Open "GET", sQuery, False
xhrRequest.send

Set domResponse = New DOMDocument60
domResponse.LoadXML xhrRequest.responseText
Set ixnStatus = domResponse.SelectSingleNode("//status")
If (ixnStatus.Text <> "OK") Then
    Exit Function
End If

Set southwestLat = domResponse.SelectSingleNode _
("/GeocodeResponse/result/geometry/bounds/southwest/lat")
Set southwestLng = domResponse.SelectSingleNode _
("/GeocodeResponse/result/geometry/bounds/southwest/lng")
Set northeastLat = domResponse.SelectSingleNode _
("/GeocodeResponse/result/geometry/bounds/northeast/lat")
Set northeastLng = domResponse.SelectSingleNode _
("/GeocodeResponse/result/geometry/bounds/northeast/lng")

GoogleGeoBBox = southwestLng.Text & "," & southwestLat.Text _
& ", " & northeastLng.Text & "," & northeastLat.Text

End Function

Public Function URLEncode(StringVal As String, _
Optional SpaceAsPlus As Boolean = False) As String
  Dim StringLen As Long: StringLen = Len(StringVal)
  If StringLen > 0 Then
    ReDim result(StringLen) As String
    Dim i As Long, CharCode As Integer
    Dim Char As String, Space As String

    If SpaceAsPlus Then Space = "+" Else Space = "%20"

    For i = 1 To StringLen
      Char = Mid$(StringVal, i, 1)
      CharCode = Asc(Char)

      Select Case CharCode
      Case 97 To 122, 65 To 90, 48 To 57, 45, 46, 95, 126
        result(i) = Char
      Case 32
        result(i) = Space
      Case 0 To 15
        result(i) = "%0" & Hex(CharCode)
      Case Else
        result(i) = "%" & Hex(CharCode)
      End Select
    Next i
    URLEncode = Join(result, "")
  End If
End Function
```

Listing C.6: BioKET DW: Auto Checker Script for Ontology mapping

```
Option Explicit
Sub OntologyChecker()

  Dim FileName As String
  Dim FileNumber As Long
  Dim PathCrnt As String
  Dim RowDestCrnt As Long
  Dim SheetDest As String
  Dim TermIRI As String
  Dim ParentTermIRI As String
  Dim WBookSrc As Workbook
  Dim Term As String
  Dim Rng As Range
  Dim RowCrnt As Long

  PathCrnt = ActiveWorkbook.Path & "\OntologyXLSX"
```

```
 SheetDest = "SetSheetFound"
 RowDestCrnt = 2

 'Set a number to start from XLSX files e.g. 1.xlsx, 2.xlsx,...
 FileNumber = 1

Do While True

   FileName = Dir$(PathCrnt & "\" & FileNumber & ".xlsx")
   If FileName = "" Then
     ' File does not exist
     Exit Sub
   End If

   Set WBookSrc = Workbooks.Open(PathCrnt & "\" & FileName)

   With WBookSrc.Worksheets("WorkSheet1")

    'Set a value for searching
   Term = "decreased rate"

  'Set a range to find out the value
  Set Rng = .Columns("B:B").Find(What:=Term, _
  After:=.Range("B1"), LookIn:=xlFormulas, _
  LookAt:=xlWhole, SearchOrder:=xlByRows, _
  SearchDirection:=xlNext, MatchCase:=False, _
  SearchFormat:=False)

       ' The entered value was found
       If Rng Is Nothing Then
       ' The entered value could not be found
         TermIRI = ""
         ParentTermIRI = ""
     Else
         RowCrnt = Rng.Row
         TermIRI = .Cells(RowCrnt, "A").Value
         ParentTermIRI = .Cells(RowCrnt, "C").Value
     End If

   End With
   WBookSrc.Close SaveChanges:=False
   With Worksheets(SheetDest)
     .Cells(RowDestCrnt, "A").Value = Mid(FileName, _
     1, Len(FileName) - 5)
```

```
            .Cells(RowDestCrnt, "B").Value = Term
            .Cells(RowDestCrnt, "C").Value = TermIRI
            .Cells(RowDestCrnt, "D").Value = ParentTermIRI

        End With
        RowDestCrnt = RowDestCrnt + 1
        FileNumber = FileNumber + 1

    Loop
End Sub
```

Appendix D

BioKET DW: Data Dictionary

Table D.1: Species

Attribute Name	**Data Type**	**Length**	**Status**	**Description**
SPECIES_ID	Integer	Auto	PK	Identity Number of Species
BARCODE	NVARCHAR	32		Identity Code of Species/ Barcode
SPECIESNAME	NVARCHAR	300		Species Name
FAMILY_ID	Integer	Auto	FK	Family ID
RISK_STATUS_ID	Integer	Auto	FK	Risk Status ID
RISK_CRITERIA_ID	Integer	Auto	FK	Risk Criteria ID
RLVERSION	NVARCHAR	15		Red List Version
R_S_SOURCE_ID	Integer	Auto	FK	Risk Status Source ID
SHORTNOTE	NVARCHAR	200		Short Note

Table D.2: GeoSpatial

Attribute Name	**Data Type**	**Length**	**Status**	**Description**
Geo_ID	NVARCHAR	30	PK	Geolocation Identity Number
LEVEL	NVARCHAR	255		Geolocation Level
GeoLocation	NVARCHAR	255		Location Name
LAT1	Decimal	(15,7)		First Latitude
LNG1	Decimal	(15,7)		First Longitude
LAT2	Decimal	(15,7)		Second Latitude
LNG2	Decimal	(15,7)		Second Longitude
PARENT_ID	NVARCHAR	30		Parent Node
SHORTNOTE	NVARCHAR	255		Short Note

Table D.3: Risk Status

Attribute Name	**Data Type**	**Length**	**Status**	**Description**
RSID	Integer	Auto	PK	Risk Status Identity Number
PREFIXCAT	NVARCHAR	15		Prefix of Risk Category
CATEGORY	NVARCHAR	150		Category of Risk

Table D.4: Risk Criteria

Attribute Name	Data Type	Length	Status	Description
CR_ID	Integer	Auto	PK	Risk Criteria Identity Number
Criteria	NVARCHAR	150		Criteria by IUCN Red List

Table D.5: R_S_Source (Risk Status Sources)

Attribute Name	Data Type	Length	Status	Description
RSSID	Integer	Auto	PK	Risk Criteria Identity Number
Label	NVARCHAR	200		Organization/Project Name
Valid_year	NVARCHAR	10		Valid/published year

Table D.6: Family

Attribute Name	Data Type	Length	Status	Description
Family_ID	Integer	Auto	PK	Family Identity Number
FamilyName	NVARCHAR	200		Family Name
SHORTNOTE	NVARCHAR	60		Short Note

Table D.7: Genus

Attribute Name	Data Type	Length	Status	Description
Genus_ID	Integer	Auto	PK	Genus Identity Number
GenusName	NVARCHAR	200		Genus Name
SHORTNOTE	NVARCHAR	60		Short Note

Table D.8: Habit

Attribute Name	Data Type	Length	Status	Description
HABIT_ID	Integer	Auto	PK	Habit Identity Number
HABITNOTE	NVARCHAR	200		Habit Description
SHORTNOTE	NVARCHAR	60		Short Note

Table D.9: TrunkBark (Trunk and Bark)

Attribute Name	Data Type	Length	Status	Description
TRUNKBARK_ID	Integer	Auto	PK	TrunkBark Identity Number
TRUNKBARKNOTE	NVARCHAR	200		TrunkBark Description
SHORTNOTE	NVARCHAR	60		Short Note

Table D.10: BBT (Branches and Branchlets or Twigs)

Attribute Name	Data Type	Length	Status	Description
BBT_ID	Integer	Auto	PK	BBT Identity Number
BBTNOTE	NVARCHAR	200		BBT Description
SHORTNOTE	NVARCHAR	60		Short Note

Table D.11: Leaves

Attribute Name	Data Type	Length	Status	Description
LEAVE_ID	Integer	Auto	PK	Leave Identity Number
LEAVENOTE	NVARCHAR	200		Leave Description
SHORTNOTE	NVARCHAR	60		Short Note

Table D.12: INFL (Inflorescences or flowers)

Attribute Name	Data Type	Length	Status	Description
INFL_ID	Integer	Auto	PK	INFL Identity Number
INFLNOTE	NVARCHAR	200		INFL Description
SHORTNOTE	NVARCHAR	60		Short Note

Table D.13: Fruit Seeds

Attribute Name	Data Type	Length	Status	Description
FRUITSEED_ID	Integer	Auto	PK	Fruit Seed Identity Number
FRUITSEEDNOTE	NVARCHAR	200		Fruit Seed Description
SHORTNOTE	NVARCHAR	60		Short Note

Table D.14: HAEC (Habitat Ecology)

Attribute Name	Data Type	Length	Status	Description
HAEC_ID	Integer	Auto	PK	HAEC Identity Number
HAECNOTE	NVARCHAR	200		HAEC Description
SHORTNOTE	NVARCHAR	60		Short Note

Table D.15: Exudates

Attribute Name	Data Type	Length	Status	Description
EXUDATE_ID	Integer	Auto	PK	Exudation Identity Number
EXUDATENOTE	NVARCHAR	200		Exudation Description
SHORTNOTE	NVARCHAR	60		Short Note

Table D.16: Characters

Attribute Name	Data Type	Length	Status	Description
CHARACTER_ID	Integer	Auto	PK	Characteristic Identity Number
CHARACTERNOTE	NVARCHAR	200		Characteristic Description
SHORTNOTE	NVARCHAR	60		Short Note

Table D.17: Basionym

Attribute Name	Data Type	Length	Status	Description
BASIONYM_ID	Integer	Auto	PK	Basionym Identity Number
BASIONYMNAME	NVARCHAR	200		Basionym Name
SHORTNOTE	NVARCHAR	60		Short Note

Table D.18: Collectors

Attribute Name	Data Type	Length	Status	Description
COLLECTOR_ID	Integer	Auto	PK	Collector Identity Number
COLLECTORNAME	NVARCHAR	200		Collector Name
SHORTNOTE	NVARCHAR	60		Short Note

Table D.19: Common Name

Attribute Name	Data Type	Length	Status	Description
COMMON_ID	Integer	Auto	PK	Common Name Identity
COMMONNAME	NVARCHAR	200		Common Name
SHORTNOTE	NVARCHAR	60		Short Note

Table D.20: Distribution

Attribute Name	Data Type	Length	Status	Description
DISTRIB_ID	Integer	Auto	PK	Distribution Identity Number
DISTRIBNOTE	NVARCHAR	200		Distribution Description
SHORTNOTE	NVARCHAR	100		Short Note
COUNTRY_CODE	NVARCHAR	16	FK	Country Code

Table D.21: Location

Attribute Name	Data Type	Length	Status	Description
LOCATION_ID	Integer	Auto	PK	Location Identity Number
COUNTRY_CODE	NVARCHAR	16	FK	Country Code
STATE	NVARCHAR	200		State Name
CITY	NVARCHAR	200		City Name
PLACE	NVARCHAR	300		Place Name
SHORTNOTE	NVARCHAR	60		Short Note

Table D.22: Country

Attribute Name	Data Type	Length	Status	Description
COUNTRY_CODE	NVARCHAR	16	PK	Country Code Identity
COUNTRYNAME	NVARCHAR	200		Country Name
SHORTNOTE	NVARCHAR	60		Short Note

Table D.23: IUCN RedList

Attribute Name	Data Type	Length	Status	Description
IDX	NVARCHAR	8	PK	Identity Number
SpeciesName	NVARCHAR	200		Species Name
Author	NVARCHAR	100		Author Name
Plant_Risk_Status	NVARCHAR	200		Plant Risk Status

Table D.24: Specimens

Attribute Name	Data Type	Length	Status	Description
SPECIMEN_ID	Integer	auto	PK	Specimen Identity Number
SPECIMENNOTE	NVARCHAR	200		Specimen Description
SHORTNOTE	NVARCHAR	60		Short Note

Table D.25: Local Names

Attribute Name	Data Type	Length	Status	Description
Localname_ID	Integer	auto	PK	Local Name Identity
LocalNOTE	NVARCHAR	400		Local Name Description
SHORTNOTE	NVARCHAR	60		Short Note

Table D.26: Images

Attribute Name	Data Type	Length	Status	Description
IMAGE_ID	Integer	auto	PK	Image Identity Number
IMAGEPATH	NVARCHAR	255		Image Path
SPECIES_ID	Integer	auto	FK	Species Identity
SHORTNOTE	NVARCHAR	60		Short Note

Table D.27: Species Notes (spNotes)

Attribute Name	Data Type	Length	Status	Description
SPNOTE_ID	Integer	auto	PK	spNote Identity Number
SPNOTES	NVARCHAR	255		Species Note
SHORTNOTE	NVARCHAR	60		Short Note

Table D.28: Plant Description (PlantDes)

Attribute Name	Data Type	Length	Status	Description
PlantDes_ID	Integer	auto	PK	PlantDes Identity Number
Description	LONGTEXT	auto		Plant Description
SHORTNOTE	NVARCHAR	60		Short Note

Table D.29: Literature

Attribute Name	Data Type	Length	Status	Description
LITERA_ID	Integer	auto	PK	Literature Identity Number
LITERANOTE	TEXT	auto		Literature Description
URLLink	NVARCHAR	150		URL Link
SHORTNOTE	NVARCHAR	60		Short Note
COUNTRY_CODE	NVARCHAR	16	FK	Country code

Table D.30: RNU (Remarks/Notes/Uses)

Attribute Name	Data Type	Length	Status	Description
RNU_ID	Integer	auto	PK	RNU Identity Number
RNUNOTE	NVARCHAR	200		RNU Description
SHORTNOTE	NVARCHAR	60		Short Note

Table D.31: Seeds

Attribute Name	Data Type	Length	Status	Description
SEED_ID	Integer	auto	PK	Seed Identity Number
SEEDNOTE	NVARCHAR	200		Seed Description
SHORTNOTE	NVARCHAR	60		Short Note

Table D.32: Treatment

Attribute Name	Data Type	Length	Status	Description
TREATMENT_ID	Integer	auto	PK	TREATMENT Identity Number
DISEASESTATE	NVARCHAR	200		DISEASE STATE
SYMPTOM	NVARCHAR	200		SYMPTOM
PREPARATION	NVARCHAR	400		PREPARATION
TREATMENTNOTE	NVARCHAR	510		TREATMENT NOTE
SHORTNOTE	NVARCHAR	60		Short Note

Table D.33: Collection Dates (CollDates)

Attribute Name	Data Type	Length	Status	Description
SPECIES_ID	Integer	auto	PK	Species Identity Number
COLLDATE	DATE	auto		Collection Date
SHORTNOTE	NVARCHAR	60		Short Note

Table D.34: Authors

Attribute Name	Data Type	Length	Status	Description
AUTHOR_ID	Integer	auto	PK	Author Identity Number
AUTHORNAME	NVARCHAR	100		Author Name
SHORTNOTE	NVARCHAR	60		Short Note

Table D.35: Synonyms

Attribute Name	Data Type	Length	Status	Description
SYNONYM_ID	Integer	Auto	PK	Synonym Identity Number
SYNONYMNOTE	NVARCHAR	200		Synonym Name
SHORTNOTE	NVARCHAR	60		Short Note

Table D.36: Species_Authors

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
AUTHOR_ID	Integer	auto	FK

Table D.37: Species_Collectors

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
COLLECTOR_ID	Integer	auto	FK

Table D.38: Species_Seeds

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
SEED_ID	Integer	auto	FK

Table D.39: Species_Literature

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
LITERA_ID	Integer	auto	FK

Table D.40: Species_RNU

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
RNU_ID	Integer	auto	FK

Table D.41: Species_Distribution

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
DISTRIB_ID	Integer	auto	FK

Table D.42: Species_HAEC

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
HAEC_ID	Integer	auto	FK

Table D.43: Species_FruitSeeds

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
FRUITSEED_ID	Integer	auto	FK

Table D.44: Species_INFL

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
INFL_ID	Integer	auto	FK

Table D.45: Species_Leaves

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
LEAVE_ID	Integer	auto	FK

Table D.46: Species_Exudates

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
EXUDATE_ID	Integer	auto	FK

Table D.47: Species_BBT

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
BBT_ID	Integer	auto	FK

Table D.48: Species_TrunkBark

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
TRUNKBARK_ID	Integer	auto	FK

Table D.49: Species_Habit

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
HABIT_ID	Integer	auto	FK

Table D.50: Species_Characters

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
CHARACTER_ID	Integer	auto	FK

Table D.51: Species_Synonyms

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
SYNONYM_ID	Integer	auto	FK

Table D.52: Species_PlantDes

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
PLANTDES_ID	Integer	auto	FK

Table D.53: Species_spNotes

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
SPNOTE_ID	Integer	auto	FK

Table D.54: Species_CommonName

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
COMMON_ID	Integer	auto	FK

Table D.55: Species_Location

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
LOCATION_ID	Integer	auto	FK

Table D.56: Species_Basionym

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
BASIONYM_ID	Integer	auto	FK

Table D.57: Species_LocalNames

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
LOCALNAME_ID	Integer	auto	FK

Table D.58: Species_Specimens

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
SPECIMEN_ID	Integer	auto	FK

Table D.59: Family_Genus

Attribute Name	Data Type	Length	Status
FAMILY_ID	Integer	auto	FK
GENUS_ID	Integer	auto	FK

Table D.60: Species_Treatment

Attribute Name	Data Type	Length	Status
FAMILY_ID	Integer	auto	FK
TREATMENT_ID	Integer	auto	FK

Table D.61: Species_GeoSpatial

Attribute Name	Data Type	Length	Status
SPECIES_ID	Integer	auto	FK
GEO_ID	Integer	auto	FK

Appendix E

BioKET DW Structure: Entities

Listing E.1: BioKET DW Structure: Tables and Views

```
SET FOREIGN_KEY_CHECKS=0;

-- ----------------------------
-- Table structure for authors
-- ----------------------------
DROP TABLE IF EXISTS ‘authors‘;
CREATE TABLE ‘authors‘ (
  ‘AUTHOR_ID‘ int(11) NOT NULL,
  ‘AUTHORNAME‘ varchar(100) NOT NULL,
  ‘SHORTNOTE‘ varchar(60) DEFAULT NULL,
 PRIMARY KEY (‘AUTHOR_ID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for basionym
-- ----------------------------
DROP TABLE IF EXISTS ‘basionym‘;
CREATE TABLE ‘basionym‘ (
  ‘BASIONYM_ID‘ int(11) NOT NULL,
  ‘BASIONYMNAME‘ varchar(200) NOT NULL,
  ‘SHORTNOTE‘ varchar(60) DEFAULT NULL,
 PRIMARY KEY (‘BASIONYM_ID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for bbt
-- ----------------------------
DROP TABLE IF EXISTS ‘bbt‘;
CREATE TABLE ‘bbt‘ (
  ‘BBT_ID‘ int(11) NOT NULL,
  ‘BBTNOTE‘ varchar(200) NOT NULL,
  ‘SHORTNOTE‘ varchar(60) DEFAULT NULL,
 PRIMARY KEY (‘BBT_ID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

```
-- ----------------------------
-- Table structure for characters
-- ----------------------------
DROP TABLE IF EXISTS ‘characters‘;
CREATE TABLE ‘characters‘ (
  ‘CHARACTER_ID‘ int(11) NOT NULL,
  ‘CHARACTERNOTE‘ varchar(200) NOT NULL,
  ‘SHORTNOTE‘ varchar(60) DEFAULT NULL,
 PRIMARY KEY (‘CHARACTER_ID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for colldates
-- ----------------------------
DROP TABLE IF EXISTS ‘colldates‘;
CREATE TABLE ‘colldates‘ (
  ‘SPECIES_ID‘ int(11) NOT NULL,
  ‘COLLDATE‘ date NOT NULL,
  ‘SHORTNOTE‘ varchar(60) DEFAULT NULL,
 PRIMARY KEY (‘SPECIES_ID‘),
 CONSTRAINT ‘colldates_ibfk_1‘
 FOREIGN KEY (‘SPECIES_ID‘) REFERENCES ‘species‘ (‘SPECIES_ID‘)
 ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for collectors
-- ----------------------------
DROP TABLE IF EXISTS ‘collectors‘;
CREATE TABLE ‘collectors‘ (
  ‘COLLECTOR_ID‘ int(11) NOT NULL,
  ‘COLLECTORNAME‘ varchar(100) NOT NULL,
  ‘SHORTNOTE‘ varchar(60) DEFAULT NULL,
 PRIMARY KEY (‘COLLECTOR_ID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for commonname
-- ----------------------------
DROP TABLE IF EXISTS ‘commonname‘;
CREATE TABLE ‘commonname‘ (
  ‘COMMON_ID‘ int(11) NOT NULL,
  ‘COMMONNAME‘ varchar(200) NOT NULL,
  `SHORTNOTE` varchar(30) DEFAULT NULL,
  PRIMARY KEY (`COMMON_ID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for country
-- ----------------------------
DROP TABLE IF EXISTS `country`;
CREATE TABLE `country` (
  `COUNTRY_CODE` varchar(16) NOT NULL,
  `COUNTRYNAME` varchar(200) NOT NULL,
  `SHORTNOTE` char(60) DEFAULT NULL,
  PRIMARY KEY (`COUNTRY_CODE`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for distribution
-- ----------------------------
DROP TABLE IF EXISTS `distribution`;
CREATE TABLE `distribution` (
  `DISTRIB_ID` int(11) NOT NULL,
  `DISTRIBNOTE` varchar(200) NOT NULL,
  `SHORTNOTE` varchar(100) DEFAULT NULL,
  `GEO_IDS_Temp` varchar(200) DEFAULT NULL,
  `COUNTRY_CODE` varchar(11) DEFAULT NULL,
  PRIMARY KEY (`DISTRIB_ID`),
  KEY `FK_Distribution_Country` (`COUNTRY_CODE`) USING BTREE
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for exudates
-- ----------------------------
DROP TABLE IF EXISTS `exudates`;
CREATE TABLE `exudates` (
  `EXUDATE_ID` int(11) NOT NULL,
  `EXUDATENOTE` varchar(200) NOT NULL,
  `SHORTNOTE` varchar(60) DEFAULT NULL,
  PRIMARY KEY (`EXUDATE_ID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for family
-- ----------------------------
DROP TABLE IF EXISTS `family`;
```

```
CREATE TABLE ‘family‘ (
   ‘FAMILY_ID‘ int(11) NOT NULL,
   ‘FAMILYNAME‘ varchar(200) NOT NULL,
   ‘SHORTNOTE‘ varchar(60) DEFAULT NULL,
  PRIMARY KEY (‘FAMILY_ID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

—— ————————————————————————
—— Table structure for family_genus
—— ————————————————————————
DROP TABLE IF EXISTS ‘family_genus‘;
CREATE TABLE ‘family_genus‘ (
   ‘FAMILY_ID‘ int(11) NOT NULL,
   ‘GENUS_ID‘ int(11) NOT NULL,
  KEY ‘FK_Fa_Ge_Genus‘ (‘GENUS_ID‘) USING BTREE,
  KEY ‘FK_Fa_Ge_Family‘ (‘FAMILY_ID‘) USING BTREE,
  CONSTRAINT ‘family_genus_ibfk_1‘
  FOREIGN KEY (‘FAMILY_ID‘) REFERENCES ‘family‘ (‘FAMILY_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT ‘family_genus_ibfk_2‘
  FOREIGN KEY (‘GENUS_ID‘) REFERENCES ‘genus‘ (‘GENUS_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

—— ————————————————————————
—— Table structure for fruitseeds
—— ————————————————————————
DROP TABLE IF EXISTS ‘fruitseeds‘;
CREATE TABLE ‘fruitseeds‘ (
   ‘FRUITSEED_ID‘ int(11) NOT NULL,
   ‘FRUITSEEDNOTE‘ varchar(200) NOT NULL,
   ‘SHORTNOTE‘ varchar(60) DEFAULT NULL,
  PRIMARY KEY (‘FRUITSEED_ID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

—— ————————————————————————
—— Table structure for genus
—— ————————————————————————
DROP TABLE IF EXISTS ‘genus‘;
CREATE TABLE ‘genus‘ (
   ‘GENUS_ID‘ int(11) NOT NULL,
   ‘GENUSNAME‘ varchar(200) NOT NULL,
   ‘SHORTNOTE‘ varchar(60) DEFAULT NULL,
  PRIMARY KEY (‘GENUS_ID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for geospatial
-- ----------------------------
DROP TABLE IF EXISTS `geospatial`;
CREATE TABLE `geospatial` (
  `Geo_ID` varchar(30) NOT NULL,
  `LEVEL` varchar(255) DEFAULT NULL,
  `GeoLocation` varchar(255) DEFAULT NULL,
  `LAT1` decimal(15,7) DEFAULT NULL,
  `LNG1` decimal(15,7) DEFAULT NULL,
  `LAT2` decimal(15,7) DEFAULT NULL,
  `LNG2` decimal(15,7) DEFAULT NULL,
  `PARENT_ID` varchar(30) NOT NULL,
  `ShortNote` varchar(255) DEFAULT NULL,
 PRIMARY KEY (`Geo_ID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for habit
-- ----------------------------
DROP TABLE IF EXISTS `habit`;
CREATE TABLE `habit` (
  `HABIT_ID` int(11) NOT NULL,
  `HABITNOTE` varchar(200) NOT NULL,
  `SHORTNOTE` varchar(60) DEFAULT NULL,
 PRIMARY KEY (`HABIT_ID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for haec
-- ----------------------------
DROP TABLE IF EXISTS `haec`;
CREATE TABLE `haec` (
  `HAEC_ID` int(11) NOT NULL,
  `HAECNOTE` varchar(200) NOT NULL,
  `SHORTNOTE` varchar(60) DEFAULT NULL,
 PRIMARY KEY (`HAEC_ID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for images
-- ----------------------------
```

```
DROP TABLE IF EXISTS `images`;
CREATE TABLE `images` (
  `IMAGE_ID` int(11) NOT NULL AUTO_INCREMENT,
  `IMAGEPATH` varchar(255) NOT NULL,
  `SPECIES_ID` int(11) NOT NULL,
  `SHORTNOTE` varchar(60) DEFAULT NULL,
  PRIMARY KEY (`IMAGE_ID`),
  KEY `FK_Images_Species` (`SPECIES_ID`) USING BTREE,
  CONSTRAINT `images_ibfk_1`
  FOREIGN KEY (`SPECIES_ID`) REFERENCES `species` (`SPECIES_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB AUTO_INCREMENT=7539 DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for infl
-- ----------------------------
DROP TABLE IF EXISTS `infl`;
CREATE TABLE `infl` (
  `INFL_ID` int(11) NOT NULL,
  `INFLNOTE` varchar(200) NOT NULL,
  `SHORTNOTE` varchar(60) DEFAULT NULL,
  PRIMARY KEY (`INFL_ID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for iucnredlist
-- ----------------------------
DROP TABLE IF EXISTS `iucnredlist`;
CREATE TABLE `iucnredlist` (
  `idx` varchar(8) DEFAULT NULL,
  `SpeciesName` varchar(200) DEFAULT NULL,
  `Author` varchar(100) DEFAULT NULL,
  `Plant_Risk_Status` varchar(200) DEFAULT NULL
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for leaves
-- ----------------------------
DROP TABLE IF EXISTS `leaves`;
CREATE TABLE `leaves` (
  `LEAVE_ID` int(11) NOT NULL,
  `LEAVENOTE` varchar(200) NOT NULL,
  `SHORTNOTE` varchar(60) DEFAULT NULL,
  PRIMARY KEY (`LEAVE_ID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for literature
-- ----------------------------
DROP TABLE IF EXISTS ‘literature‘;
CREATE TABLE ‘literature‘ (
  ‘LITERA_ID‘ int(11) NOT NULL,
  ‘LITERANOTE‘ text NOT NULL,
  ‘URLLINK‘ varbinary(150) DEFAULT NULL,
  ‘COUNTRY_CODE‘ varchar(16) DEFAULT NULL,
  ‘SHORTNOTE‘ varchar(60) DEFAULT NULL,
 PRIMARY KEY (‘LITERA_ID‘),
 KEY ‘FK_Literature_Country‘ (‘COUNTRY_CODE‘) USING BTREE
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for localnames
-- ----------------------------
DROP TABLE IF EXISTS ‘localnames‘;
CREATE TABLE ‘localnames‘ (
  ‘LOCALNAME_ID‘ int(11) NOT NULL,
  ‘LOCALNOTE‘ varchar(400) NOT NULL,
  ‘SHORTNOTE‘ varchar(60) DEFAULT NULL,
 PRIMARY KEY (‘LOCALNAME_ID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for location
-- ----------------------------
DROP TABLE IF EXISTS ‘location‘;
CREATE TABLE ‘location‘ (
  ‘LOCATION_ID‘ int(11) NOT NULL DEFAULT '0',
  ‘COUNTRY_CODE‘ varchar(16) NOT NULL,
  ‘STATE‘ varchar(200) NOT NULL,
  ‘CITY‘ varchar(200) NOT NULL,
  ‘PLACE‘ varchar(300) DEFAULT NULL,
  ‘SHORTNOTE‘ varchar(60) DEFAULT NULL,
 PRIMARY KEY (‘LOCATION_ID‘),
 KEY ‘FK_Location_Country‘ (‘COUNTRY_CODE‘) USING BTREE,
 CONSTRAINT ‘location_ibfk_1‘
 FOREIGN KEY (‘COUNTRY_CODE‘) REFERENCES ‘country‘ (‘COUNTRY_CODE‘)
 ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

```
-- ----------------------------
-- Table structure for plantdes
-- ----------------------------
DROP TABLE IF EXISTS ‘plantdes‘;
CREATE TABLE ‘plantdes‘ (
  ‘PLANTDES_ID‘ int(11) NOT NULL,
  ‘Description‘ longtext NOT NULL,
  ‘SHORTNOTE‘ varchar(30) DEFAULT NULL,
 PRIMARY KEY (‘PLANTDES_ID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for risk_criteria
-- ----------------------------
DROP TABLE IF EXISTS ‘risk_criteria‘;
CREATE TABLE ‘risk_criteria‘ (
  ‘CR_ID‘ int(11) NOT NULL,
  ‘CRITERIA‘ varchar(150) DEFAULT NULL,
 PRIMARY KEY (‘CR_ID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for risk_status
-- ----------------------------
DROP TABLE IF EXISTS ‘risk_status‘;
CREATE TABLE ‘risk_status‘ (
  ‘RSID‘ int(10) NOT NULL,
  ‘PREFIXCAT‘ varchar(15) DEFAULT NULL,
  ‘CATEGORY‘ varchar(150) DEFAULT NULL,
 PRIMARY KEY (‘RSID‘),
 KEY ‘RSID‘ (‘RSID‘,‘PREFIXCAT‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for rnu
-- ----------------------------
DROP TABLE IF EXISTS ‘rnu‘;
CREATE TABLE ‘rnu‘ (
  ‘RNU_ID‘ int(11) NOT NULL,
  ‘RNUNOTE‘ varchar(200) NOT NULL,
  ‘SHORTNOTE‘ varchar(60) DEFAULT NULL,
 PRIMARY KEY (‘RNU_ID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

```
-- ----------------------------
-- Table structure for r_s_source
-- ----------------------------
DROP TABLE IF EXISTS ‘r_s_source‘;
CREATE TABLE ‘r_s_source‘ (
   ‘RSSID‘ int(10) NOT NULL,
   ‘LABEL‘ varchar(200) DEFAULT NULL,
   ‘VALID_YEAR‘ varchar(10) DEFAULT NULL,
  PRIMARY KEY (‘RSSID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for seeds
-- ----------------------------
DROP TABLE IF EXISTS ‘seeds‘;
CREATE TABLE ‘seeds‘ (
   ‘SEED_ID‘ int(11) NOT NULL,
   ‘SEEDNOTE‘ varchar(200) NOT NULL,
   ‘SHORTNOTE‘ varchar(60) DEFAULT NULL,
  PRIMARY KEY (‘SEED_ID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species
-- ----------------------------
DROP TABLE IF EXISTS ‘species‘;
CREATE TABLE ‘species‘ (
   ‘SPECIES_ID‘ int(11) NOT NULL,
   ‘BARCODE‘ varchar(32) NOT NULL,
   ‘SPECIESNAME‘ varchar(300) NOT NULL,
   ‘FAMILY_ID‘ int(11) NOT NULL,
   ‘RISK_STATUS_ID‘ int(11) DEFAULT NULL,
   ‘RISK_CRITERIA_ID‘ int(11) DEFAULT NULL,
   ‘RLVERSION‘ varchar(15) DEFAULT NULL,
   ‘R_S_SOURCE_ID‘ int(11) DEFAULT NULL,
   ‘SHORTNOTE‘ varchar(200) DEFAULT NULL,
  PRIMARY KEY (‘SPECIES_ID‘),
  KEY ‘FK_Species_Family‘ (‘FAMILY_ID‘) USING BTREE,
  KEY ‘FK_RISK_STATUS_ID‘ (‘RISK_STATUS_ID‘),
  KEY ‘FK_RISK_CRITERIA_ID‘ (‘RISK_CRITERIA_ID‘),
  KEY ‘FK_R_S_SOURCE_ID‘ (‘R_S_SOURCE_ID‘),
  CONSTRAINT ‘FK_RISK_CRITERIA_ID‘
  FOREIGN KEY (‘RISK_CRITERIA_ID‘) REFERENCES ‘risk_criteria‘ (‘CR_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT ‘FK_RISK_STATUS_ID‘
  FOREIGN KEY (‘RISK_STATUS_ID‘) REFERENCES ‘risk_status‘ (‘RSID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT ‘FK_R_S_SOURCE_ID‘
  FOREIGN KEY (‘R_S_SOURCE_ID‘) REFERENCES ‘r_s_source‘ (‘RSSID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT ‘species_ibfk_1‘
  FOREIGN KEY (‘FAMILY_ID‘) REFERENCES ‘family‘ (‘FAMILY_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_authors
-- ----------------------------
DROP TABLE IF EXISTS ‘species_authors‘;
CREATE TABLE ‘species_authors‘ (
  ‘SPECIES_ID‘ int(11) NOT NULL,
  ‘AUTHOR_ID‘ int(11) NOT NULL,
  KEY ‘FK_Species_Authors_Authors‘ (‘AUTHOR_ID‘) USING BTREE,
  KEY ‘FK_Species_Authors_Species‘ (‘SPECIES_ID‘) USING BTREE,
  CONSTRAINT ‘species_authors_ibfk_1‘
  FOREIGN KEY (‘AUTHOR_ID‘) REFERENCES ‘authors‘ (‘AUTHOR_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT ‘species_authors_ibfk_2‘
  FOREIGN KEY (‘SPECIES_ID‘) REFERENCES ‘species‘ (‘SPECIES_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_basionym
-- ----------------------------
DROP TABLE IF EXISTS ‘species_basionym‘;
CREATE TABLE ‘species_basionym‘ (
  ‘SPECIES_ID‘ int(11) NOT NULL,
  ‘BASIONYM_ID‘ int(11) NOT NULL,
  KEY ‘FK_species_basionym_species‘ (‘SPECIES_ID‘) USING BTREE,
  KEY ‘FK_species_basionym_basionym‘ (‘BASIONYM_ID‘) USING BTREE,
  CONSTRAINT ‘species_basionym_ibfk_1‘
  FOREIGN KEY (‘BASIONYM_ID‘) REFERENCES ‘basionym‘ (‘BASIONYM_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT ‘species_basionym_ibfk_2‘
  FOREIGN KEY (‘SPECIES_ID‘) REFERENCES ‘species‘ (‘SPECIES_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION
```

```
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_bbt
-- ----------------------------
DROP TABLE IF EXISTS `species_bbt`;
CREATE TABLE `species_bbt` (
  `SPECIES_ID` int(11) NOT NULL,
  `BBT_ID` int(11) NOT NULL,
  KEY `FK_Species_Branch_Branchlet_Twig_Species` (`SPECIES_ID`)
  USING BTREE,
  KEY `FK_Species_Branch_Branchlet_Twig_Branch_Branchlet_Twig` (`BBT_ID`)
  USING BTREE,
  CONSTRAINT `species_bbt_ibfk_1`
  FOREIGN KEY (`BBT_ID`) REFERENCES `bbt` (`BBT_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `species_bbt_ibfk_2`
  FOREIGN KEY (`SPECIES_ID`)
  REFERENCES `species` (`SPECIES_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_characters
-- ----------------------------
DROP TABLE IF EXISTS `species_characters`;
CREATE TABLE `species_characters` (
  `SPECIES_ID` int(11) NOT NULL,
  `CHARACTER_ID` int(11) NOT NULL,
  KEY `FK_Species_Characters_Species` (`SPECIES_ID`)
  USING BTREE,
  KEY `FK_Species_Characters_Characters` (`CHARACTER_ID`)
  USING BTREE,
  CONSTRAINT `species_characters_ibfk_1`
  FOREIGN KEY (`CHARACTER_ID`)
  REFERENCES `characters` (`CHARACTER_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `species_characters_ibfk_2`
  FOREIGN KEY (`SPECIES_ID`)
  REFERENCES `species` (`SPECIES_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_collectors
-- ----------------------------
DROP TABLE IF EXISTS `species_collectors`;
CREATE TABLE `species_collectors` (
  `SPECIES_ID` int(11) NOT NULL,
  `COLLECTOR_ID` int(11) NOT NULL,
  KEY `FK_Species_Collectors_Collectors` (`COLLECTOR_ID`) USING BTREE,
  KEY `FK_Species_Collectors_Species` (`SPECIES_ID`) USING BTREE,
  CONSTRAINT `species_collectors_ibfk_1`
  FOREIGN KEY (`COLLECTOR_ID`) REFERENCES `collectors` (`COLLECTOR_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `species_collectors_ibfk_2`
  FOREIGN KEY (`SPECIES_ID`) REFERENCES `species` (`SPECIES_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_commonname
-- ----------------------------
DROP TABLE IF EXISTS `species_commonname`;
CREATE TABLE `species_commonname` (
  `SPECIES_ID` int(11) NOT NULL,
  `COMMON_ID` int(11) NOT NULL,
  KEY `FK_Species_Commonname_Species` (`SPECIES_ID`) USING BTREE,
  KEY `FK_Species_Commonname_Commonname` (`COMMON_ID`) USING BTREE,
  CONSTRAINT `species_commonname_ibfk_1`
  FOREIGN KEY (`COMMON_ID`) REFERENCES `commonname` (`COMMON_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `species_commonname_ibfk_2`
  FOREIGN KEY (`SPECIES_ID`) REFERENCES `species` (`SPECIES_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_distribution
-- ----------------------------
DROP TABLE IF EXISTS `species_distribution`;
CREATE TABLE `species_distribution` (
  `SPECIES_ID` int(11) NOT NULL,
  `DISTRIB_ID` int(11) NOT NULL,
  KEY `FK_Species_Distribution_Distribution` (`DISTRIB_ID`) USING BTREE,
  KEY `FK_Species_Distribution_Species` (`SPECIES_ID`) USING BTREE,
  CONSTRAINT `species_distribution_ibfk_1`
  FOREIGN KEY (`DISTRIB_ID`) REFERENCES `distribution` (`DISTRIB_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `species_distribution_ibfk_2`
  FOREIGN KEY (`SPECIES_ID`) REFERENCES `species` (`SPECIES_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_exudates
-- ----------------------------
DROP TABLE IF EXISTS `species_exudates`;
CREATE TABLE `species_exudates` (
  `SPECIES_ID` int(11) NOT NULL,
  `EXUDATE_ID` int(11) NOT NULL,
  KEY `FK_Species_Exudates_Exudates` (`EXUDATE_ID`) USING BTREE,
  KEY `FK_Species_Exudates_Species` (`SPECIES_ID`) USING BTREE,
  CONSTRAINT `species_exudates_ibfk_1`
  FOREIGN KEY (`EXUDATE_ID`) REFERENCES `exudates` (`EXUDATE_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `species_exudates_ibfk_2`
  FOREIGN KEY (`SPECIES_ID`) REFERENCES `species` (`SPECIES_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_fruitseeds
-- ----------------------------
DROP TABLE IF EXISTS `species_fruitseeds`;
CREATE TABLE `species_fruitseeds` (
  `SPECIES_ID` int(11) NOT NULL,
  `FRUITSEED_ID` int(11) NOT NULL,
  KEY `FK_Species_Fruitseeds_Fruitseeds` (`FRUITSEED_ID`) USING BTREE,
  KEY `FK_Species_Fruitseeds_Species` (`SPECIES_ID`) USING BTREE,
  CONSTRAINT `species_fruitseeds_ibfk_1`
  FOREIGN KEY (`FRUITSEED_ID`) REFERENCES `fruitseeds` (`FRUITSEED_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `species_fruitseeds_ibfk_2`
  FOREIGN KEY (`SPECIES_ID`) REFERENCES `species` (`SPECIES_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_geospatial
-- ----------------------------
DROP TABLE IF EXISTS `species_geospatial`;
```

```
CREATE TABLE ‘species_geospatial‘ (
  ‘Species_ID‘ int(20) NOT NULL,
  ‘Geo_ID‘ varchar(30) DEFAULT NULL
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_habit
-- ----------------------------
DROP TABLE IF EXISTS ‘species_habit‘;
CREATE TABLE ‘species_habit‘ (
  ‘SPECIES_ID‘ int(11) NOT NULL,
  ‘HABIT_ID‘ int(11) NOT NULL,
 KEY ‘FK_Species_Habit_Species‘ (‘SPECIES_ID‘) USING BTREE,
 KEY ‘FK_Species_Habit_Habit‘ (‘HABIT_ID‘) USING BTREE,
 CONSTRAINT ‘species_habit_ibfk_1‘
 FOREIGN KEY (‘HABIT_ID‘) REFERENCES ‘habit‘ (‘HABIT_ID‘)
 ON DELETE NO ACTION ON UPDATE NO ACTION,
 CONSTRAINT ‘species_habit_ibfk_2‘
 FOREIGN KEY (‘SPECIES_ID‘) REFERENCES ‘species‘ (‘SPECIES_ID‘)
 ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_haec
-- ----------------------------
DROP TABLE IF EXISTS ‘species_haec‘;
CREATE TABLE ‘species_haec‘ (
  ‘SPECIES_ID‘ int(11) NOT NULL,
  ‘HAEC_ID‘ int(11) NOT NULL,
 KEY ‘FK_Species_Habit_Ecology_Habit_Ecology‘ (‘HAEC_ID‘) USING BTREE,
 KEY ‘FK_Species_Habit_Ecology_Species‘ (‘SPECIES_ID‘) USING BTREE,
 CONSTRAINT ‘species_haec_ibfk_1‘
 FOREIGN KEY (‘HAEC_ID‘) REFERENCES ‘haec‘ (‘HAEC_ID‘)
 ON DELETE NO ACTION ON UPDATE NO ACTION,
 CONSTRAINT ‘species_haec_ibfk_2‘
 FOREIGN KEY (‘SPECIES_ID‘) REFERENCES ‘species‘ (‘SPECIES_ID‘)
 ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_infl
-- ----------------------------
DROP TABLE IF EXISTS ‘species_infl‘;
CREATE TABLE ‘species_infl‘ (
  ‘SPECIES_ID‘ int(11) NOT NULL,
  ‘INFL_ID‘ int(11) NOT NULL,
  KEY ‘FK_Species_Inflorescence_Flower_Inflorescence_Flower‘ (‘INFL_ID‘)
  USING BTREE,
  KEY ‘FK_Species_Inflorescence_Flower_Species‘ (‘SPECIES_ID‘)
  USING BTREE,
  CONSTRAINT ‘species_infl_ibfk_1‘
  FOREIGN KEY (‘INFL_ID‘) REFERENCES ‘infl‘ (‘INFL_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT ‘species_infl_ibfk_2‘
  FOREIGN KEY (‘SPECIES_ID‘)
  REFERENCES ‘species‘ (‘SPECIES_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_leaves
-- ----------------------------
DROP TABLE IF EXISTS ‘species_leaves‘;
CREATE TABLE ‘species_leaves‘ (
  ‘SPECIES_ID‘ int(11) NOT NULL,
  ‘LEAVE_ID‘ int(11) NOT NULL,
  KEY ‘FK_Species_Leaves_Leaves‘ (‘LEAVE_ID‘) USING BTREE,
  KEY ‘FK_Species_Leaves_Species‘ (‘SPECIES_ID‘) USING BTREE,
  CONSTRAINT ‘species_leaves_ibfk_1‘
  FOREIGN KEY (‘LEAVE_ID‘) REFERENCES ‘leaves‘ (‘LEAVE_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT ‘species_leaves_ibfk_2‘
  FOREIGN KEY (‘SPECIES_ID‘) REFERENCES ‘species‘ (‘SPECIES_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_literature
-- ----------------------------
DROP TABLE IF EXISTS ‘species_literature‘;
CREATE TABLE ‘species_literature‘ (
  ‘SPECIES_ID‘ int(11) NOT NULL,
  ‘LITERA_ID‘ int(11) NOT NULL,
  KEY ‘FK_Species_Literature_Species‘ (‘SPECIES_ID‘)
  USING BTREE,
  KEY ‘FK_Species_Literature_Literature‘ (‘LITERA_ID‘)
  USING BTREE,
  CONSTRAINT ‘species_literature_ibfk_1‘
  FOREIGN KEY (‘LITERA_ID‘)
  REFERENCES ‘literature‘ (‘LITERA_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT ‘species_literature_ibfk_2‘
  FOREIGN KEY (‘SPECIES_ID‘)
  REFERENCES ‘species‘ (‘SPECIES_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_localnames
-- ----------------------------
DROP TABLE IF EXISTS ‘species_localnames‘;
CREATE TABLE ‘species_localnames‘ (
  ‘SPECIES_ID‘ int(11) NOT NULL,
  ‘LOCALNAME_ID‘ int(11) NOT NULL,
  KEY ‘FK_Species_LocalNames_LocalNames‘ (‘LOCALNAME_ID‘)
  USING BTREE,
  KEY ‘FK_Species_LocalNames_Species‘ (‘SPECIES_ID‘)
  USING BTREE,
  CONSTRAINT ‘species_localnames_ibfk_1‘
  FOREIGN KEY (‘LOCALNAME_ID‘) REFERENCES ‘localnames‘ (‘LOCALNAME_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT ‘species_localnames_ibfk_2‘
  FOREIGN KEY (‘SPECIES_ID‘)
  REFERENCES ‘species‘ (‘SPECIES_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_location
-- ----------------------------
DROP TABLE IF EXISTS ‘species_location‘;
CREATE TABLE ‘species_location‘ (
  ‘SPECIES_ID‘ int(11) NOT NULL,
  ‘LOCATION_ID‘ int(11) NOT NULL,
  KEY ‘FK_Species_Location_Location‘ (‘LOCATION_ID‘)
  USING BTREE,
  KEY ‘FK_Species_Location_Species‘ (‘SPECIES_ID‘)
  USING BTREE,
  CONSTRAINT ‘species_location_ibfk_1‘
  FOREIGN KEY (‘LOCATION_ID‘)
  REFERENCES ‘location‘ (‘LOCATION_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `species_location_ibfk_2`
  FOREIGN KEY (`SPECIES_ID`)
  REFERENCES `species` (`SPECIES_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_plantdes
-- ----------------------------
DROP TABLE IF EXISTS `species_plantdes`;
CREATE TABLE `species_plantdes` (
  `SPECIES_ID` int(11) NOT NULL,
  `PLANTDES_ID` int(11) NOT NULL,
  KEY `FK_Species_PlantDes_PlantDes` (`PLANTDES_ID`) USING BTREE,
  KEY `FK_Species_PlantDes_Species` (`SPECIES_ID`) USING BTREE,
  CONSTRAINT `species_plantdes_ibfk_1`
  FOREIGN KEY (`PLANTDES_ID`) REFERENCES `plantdes` (`PLANTDES_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `species_plantdes_ibfk_2`
  FOREIGN KEY (`SPECIES_ID`)
  REFERENCES `species` (`SPECIES_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_rnu
-- ----------------------------
DROP TABLE IF EXISTS `species_rnu`;
CREATE TABLE `species_rnu` (
  `SPECIES_ID` int(11) NOT NULL,
  `RNU_ID` int(11) NOT NULL,
  KEY `FK_Species_Remark_Note_Use_Remark_Note_Use` (`RNU_ID`)
  USING BTREE,
  KEY `FK_Species_Remark_Note_Use_Species` (`SPECIES_ID`)
  USING BTREE,
  CONSTRAINT `species_rnu_ibfk_1`
  FOREIGN KEY (`RNU_ID`)
  REFERENCES `rnu` (`RNU_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `species_rnu_ibfk_2`
  FOREIGN KEY (`SPECIES_ID`)
  REFERENCES `species` (`SPECIES_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

```
-- ----------------------------
-- Table structure for species_seeds
-- ----------------------------
DROP TABLE IF EXISTS ‘species_seeds‘;
CREATE TABLE ‘species_seeds‘ (
  ‘SPECIES_ID‘ int(11) NOT NULL,
  ‘SEED_ID‘ int(11) NOT NULL,
  KEY ‘FK_Species_Seeds_Seeds‘ (‘SEED_ID‘)
  USING BTREE,
  KEY ‘FK_Species_Seeds_Species‘ (‘SPECIES_ID‘)
  USING BTREE,
  CONSTRAINT ‘species_seeds_ibfk_1‘
  FOREIGN KEY (‘SEED_ID‘)
  REFERENCES ‘seeds‘ (‘SEED_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT ‘species_seeds_ibfk_2‘
  FOREIGN KEY (‘SPECIES_ID‘)
  REFERENCES ‘species‘ (‘SPECIES_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_specimens
-- ----------------------------
DROP TABLE IF EXISTS ‘species_specimens‘;
CREATE TABLE ‘species_specimens‘ (
  ‘SPECIES_ID‘ int(11) NOT NULL,
  ‘SPECIMEN_ID‘ int(11) NOT NULL,
  KEY ‘FK_Species_Specimens_Specimens‘ (‘SPECIMEN_ID‘)
  USING BTREE,
  KEY ‘FK_Species_Specimens_Species‘ (‘SPECIES_ID‘)
  USING BTREE,
  CONSTRAINT ‘species_specimens_ibfk_1‘
  FOREIGN KEY (‘SPECIES_ID‘)
  REFERENCES ‘species‘ (‘SPECIES_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT ‘species_specimens_ibfk_2‘
  FOREIGN KEY (‘SPECIMEN_ID‘)
  REFERENCES ‘specimens‘ (‘SPECIMEN_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_spnotes
-- ----------------------------
DROP TABLE IF EXISTS `species_spnotes`;
CREATE TABLE `species_spnotes` (
  `SPECIES_ID` int(11) NOT NULL,
  `SPNOTE_ID` int(11) NOT NULL,
  KEY `FK_Species_SpeciesNotes_SpeciesNotes` (`SPNOTE_ID`)
  USING BTREE,
  KEY `FK_Species_SpeciesNotes_Species` (`SPECIES_ID`)
  USING BTREE,
  CONSTRAINT `species_spnotes_ibfk_1`
  FOREIGN KEY (`SPECIES_ID`) REFERENCES `species` (`SPECIES_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `species_spnotes_ibfk_2`
  FOREIGN KEY (`SPNOTE_ID`)
  REFERENCES `spnotes` (`SPNOTE_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_synonyms
-- ----------------------------
DROP TABLE IF EXISTS `species_synonyms`;
CREATE TABLE `species_synonyms` (
  `SPECIES_ID` int(11) NOT NULL,
  `SYNONYM_ID` int(11) NOT NULL,
  KEY `FK_species_synonyms_species` (`SPECIES_ID`) USING BTREE,
  KEY `FK_species_synonyms_synonyms` (`SYNONYM_ID`) USING BTREE,
  CONSTRAINT `species_synonyms_ibfk_1`
  FOREIGN KEY (`SPECIES_ID`) REFERENCES `species` (`SPECIES_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `species_synonyms_ibfk_2`
  FOREIGN KEY (`SYNONYM_ID`) REFERENCES `synonyms` (`SYNONYM_ID`)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_treatment
-- ----------------------------
DROP TABLE IF EXISTS `species_treatment`;
CREATE TABLE `species_treatment` (
  `SPECIES_ID` int(11) NOT NULL,
  `TREATMENT_ID` int(11) NOT NULL,
  KEY `FK_Species_Treatment_Treatment` (`TREATMENT_ID`) USING BTREE,
  KEY ‘FK_Species_Treatment_Species‘ (‘SPECIES_ID‘) USING BTREE,
  CONSTRAINT ‘species_treatment_ibfk_1‘
  FOREIGN KEY (‘SPECIES_ID‘) REFERENCES ‘species‘ (‘SPECIES_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT ‘species_treatment_ibfk_2‘
  FOREIGN KEY (‘TREATMENT_ID‘) REFERENCES ‘treatment‘ (‘TREATMENT_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for species_trunkbark
-- ----------------------------
DROP TABLE IF EXISTS ‘species_trunkbark‘;
CREATE TABLE ‘species_trunkbark‘ (
  ‘SPECIES_ID‘ int(11) NOT NULL,
  ‘TRUNKBARK_ID‘ int(11) NOT NULL,
  KEY ‘FK_Species_Trunk_Bark_Species‘ (‘SPECIES_ID‘) USING BTREE,
  KEY ‘FK_Species_Trunk_Bark_Trunk_Bark‘ (‘TRUNKBARK_ID‘) USING BTREE,
  CONSTRAINT ‘species_trunkbark_ibfk_1‘
  FOREIGN KEY (‘SPECIES_ID‘) REFERENCES ‘species‘ (‘SPECIES_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT ‘species_trunkbark_ibfk_2‘
  FOREIGN KEY (‘TRUNKBARK_ID‘) REFERENCES ‘trunkbark‘ (‘TRUNKBARK_ID‘)
  ON DELETE NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for specimens
-- ----------------------------
DROP TABLE IF EXISTS ‘specimens‘;
CREATE TABLE ‘specimens‘ (
  ‘SPECIMEN_ID‘ int(11) NOT NULL,
  ‘SPECIMENNOTE‘ varchar(200) NOT NULL,
  ‘SHORTNOTE‘ varchar(60) DEFAULT NULL,
  PRIMARY KEY (‘SPECIMEN_ID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for spnotes
-- ----------------------------
DROP TABLE IF EXISTS ‘spnotes‘;
CREATE TABLE ‘spnotes‘ (
  ‘SPNOTE_ID‘ int(11) NOT NULL,
  ‘SPNOTES‘ varchar(400) NOT NULL,
  `SHORTNOTE` varchar(60) DEFAULT NULL,
  PRIMARY KEY (`SPNOTE_ID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for synonyms
-- ----------------------------
DROP TABLE IF EXISTS `synonyms`;
CREATE TABLE `synonyms` (
  `SYNONYM_ID` int(11) NOT NULL,
  `SYNONYMNOTE` varchar(200) NOT NULL,
  `SHORTNOTE` varchar(60) DEFAULT NULL,
  PRIMARY KEY (`SYNONYM_ID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for tmp652
-- ----------------------------
DROP TABLE IF EXISTS `tmp652`;
CREATE TABLE `tmp652` (
  `SPID` int(11) NOT NULL,
  PRIMARY KEY (`SPID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for treatment
-- ----------------------------
DROP TABLE IF EXISTS `treatment`;
CREATE TABLE `treatment` (
  `TREATMENT_ID` int(11) NOT NULL,
  `DISEASESTATE` varchar(200) NOT NULL,
  `SYMPTOM` varchar(200) DEFAULT NULL,
  `PREPARATION` varchar(400) DEFAULT NULL,
  `TREATMENTNOTE` varchar(510) DEFAULT NULL,
  `SHORTNOTE` varchar(60) DEFAULT NULL,
  PRIMARY KEY (`TREATMENT_ID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- Table structure for trunkbark
-- ----------------------------
DROP TABLE IF EXISTS `trunkbark`;
CREATE TABLE `trunkbark` (
  `TRUNKBARK_ID` int(11) NOT NULL,
  `TRUNKBARKNOTE` varchar(200) NOT NULL,
  `SHORTNOTE` varchar(60) DEFAULT NULL,
  PRIMARY KEY (`TRUNKBARK_ID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------
-- View structure for species_geospatial-w652
-- ----------------------------
DROP VIEW IF EXISTS `species_geospatial-w652`;
CREATE ALGORITHM=UNDEFINED DEFINER=`root`@`localhost`
SQL SECURITY DEFINER VIEW `species_geospatial-w652`
AS select `vsp`.`SPECIES_ID`
AS `SPECIES_ID`,`vsp`.`SPECIESNAME`
AS `SPECIESNAME`,`vgeo`.`GeoLocation`
AS `GeoLocation`,`vgeo`.`LEVEL`
AS `LEVEL`,`vgeo`.`LAT1`
AS `LAT1`,`vgeo`.`LNG1` AS `LNG1`,`vgeo`.`LAT2` AS `LAT2`,
`vgeo`.`LNG2` AS `LNG2`,`vgeo`.`Parent`
AS `Parent`
from ((`species` `vsp`
join `species_geospatial` `vsg`
on((`vsp`.`SPECIES_ID` = `vsg`.`Species_ID`)))
join `vgeospatial` `vgeo`
on((`vgeo`.`Geo_ID` = `vsg`.`Geo_ID`)))
where `vsp`.`SPECIES_ID` in (select `tmp652`.`SPID` from `tmp652`)
order by `vsp`.`SPECIES_ID` ;

-- ----------------------------
-- View structure for spneeds
-- ----------------------------
DROP VIEW IF EXISTS `spneeds`;
CREATE ALGORITHM=UNDEFINED DEFINER=`root`@`localhost`
VIEW `spneeds`
AS Select vsp.SPECIES_ID, vsp.BARCODE, vsp.SPECIESNAME,
vfm.FAMILYNAME, vau.AUTHORNAME,vsp.SHORTNOTE
From species as vsp
INNER JOIN  species_authors as vsa on vsp.SPECIES_ID = vsa.SPECIES_ID
INNER JOIN  `authors` as vau on vsa.AUTHOR_ID = vau.AUTHOR_ID
INNER JOIN  family as vfm on vsp.FAMILY_ID = vfm.FAMILY_ID
ORDER BY vsp.SPECIES_ID ;

-- ----------------------------
-- View structure for vgeospatial
-- ----------------------------
```

```
DROP VIEW IF EXISTS `vgeospatial`;
CREATE ALGORITHM=UNDEFINED DEFINER=`root`@`localhost`
SQL SECURITY DEFINER VIEW `vgeospatial` AS ;

-- ----------------------------
-- View structure for vsp_rl167
-- ----------------------------
DROP VIEW IF EXISTS `vsp_rl167`;
CREATE ALGORITHM=UNDEFINED DEFINER=`root`@`localhost`
SQL SECURITY DEFINER VIEW `vsp_rl167`
AS SELECT sp.SPECIES_ID, sp.SPECIESNAME,fm.FAMILYNAME, rs.PREFIXCAT,
 rs.CATEGORY,rc.CRITERIA, sp.RLVERSION, rss.LABEL,rss.VALID_YEAR,
GROUP_CONCAT(CONCAT(dis.DISTRIBNOTE) SEPARATOR '; ')
 as Distribution_Notes
FROM species sp

LEFT JOIN RISK_STATUS rs on sp.RISK_STATUS_ID=rs.RSID
LEFT JOIN risk_criteria rc on sp.RISK_CRITERIA_ID =rc.CR_ID
LEFT JOIN r_s_source rss on sp.R_S_SOURCE_ID=rss.RSSID
LEFT JOIN family fm on sp.FAMILY_ID =fm.FAMILY_ID
INNER JOIN  species_distribution spd on sp.SPECIES_ID =spd.SPECIES_ID
-- LEFT JOIN  species_distribution spd on sp.SPECIES_ID =spd.SPECIES_ID
LEFT JOIN distribution dis on spd.DISTRIB_ID =dis.DISTRIB_ID


WHERE sp.RISK_STATUS_ID is NOT NULL
GROUP BY sp.SPECIES_ID ;

-- ----------------------------
-- View structure for vt217
-- ----------------------------
DROP VIEW IF EXISTS `vt217`;
CREATE ALGORITHM=UNDEFINED DEFINER=`root`@`localhost`
SQL SECURITY DEFINER VIEW `vt217`
AS SELECT sp.SPECIES_ID, sp.SPECIESNAME, RS.CATEGORY
FROM species sp, risk_status RS
 Where sp.RISK_STATUS_ID=rs.RSID
 and sp.SPECIES_ID in (SELECT tmp652.spid from tmp652) ;

-- ----------------------------
-- View structure for vwsp_leaves
-- ----------------------------
DROP VIEW IF EXISTS `vwsp_leaves`;
CREATE ALGORITHM=UNDEFINED DEFINER=`root`@`localhost`
```

```
VIEW `vwsp_leaves` AS (Select vs.SPECIES_ID, vs.BARCODE,
vs.SPECIESNAME, vf.FAMILYNAME, va.AUTHORNAME,
GROUP_CONCAT(vl.LEAVENOTE SEPARATOR ', ') as LEAVE_NOTE
From species as vs
INNER JOIN  species_leaves as vsl on vs.SPECIES_ID = vsl.SPECIES_ID
INNER JOIN  `leaves` as vl on vsl.LEAVE_ID = vl.LEAVE_ID
INNER JOIN  species_authors as vsa on vs.SPECIES_ID= vsa.SPECIES_ID
INNER JOIN  `authors` as va on vsa.AUTHOR_ID = va.AUTHOR_ID
INNER JOIN  family as vf on vs.FAMILY_ID = vf.FAMILY_ID
GROUP BY vs.SPECIES_ID) ;
```

Bibliography

[2013a] *Air Pollution.* http://www.sciencedaily.com/articles/a/air_pollution.htm, Accessed June 2013. (Cited on page 11.)

[2011b] *Biodiversity Informatics and Co-Operation in Taxonomy for Interactive shared Knowledge Base (BIOTIK).* http://www.biotik.org, Accessed September 2011. (Cited on pages 2, 32 and 38.)

[2013c] *Biological Sequences.* http://www.ncbi.nlm.nih.gov/IEB/ToolBox/SDKDOCS/BIOSEQ.HTML, Accessed July 2013. (Cited on page 63.)

[2013d] *Biology Online.* http://www.biology-online.org, Accessed July 2013. (Cited on page 64.)

[2013e] *Botanical Research And Herbarium Management System (BRAHMS).* http://herbaria.plants.ox.ac.uk/bol/, Accessed January 2013. (Cited on pages 2, 32 and 33.)

[2014f] http://wiki.openstreetmap.org/wiki/Bounding_box, Accessed April 2014. (Cited on page 38.)

[2013g] *Climate Change and Human Health.* http://www.who.int/globalchange/en/, Accessed February 2013. (Cited on pages 7, 10 and 12.)

[2013h] *Earth Carbon Dioxide (CO2).* http://co2now.org, Accessed September 2013. (Cited on pages iii and 11.)

[2013i] *General situation of world fish stocks.* http://www.fao.org/newsroom/common/ecg/1000505/en/stocks.pdf, 2013. (Cited on page 13.)

[2014j] *GeoCAT: Geospatial Conservation Assessment Tool.* http://geocat.kew.org/, Accessed April 2014. (Cited on page 38.)

[2013k] *Global Biodiversity Outlook 3.* http://www.cbd.int/gbo3, Accessed January 2013. (Cited on pages iii, 1, 7, 8, 9, 10, 11 and 12.)

[2013l] *Global Health Observatory (GHO), Outdoor air pollution.* http://www.who.int/gho/phe/outdoor_air_pollution/en/index.html, Accessed June 2013. (Cited on page 11.)

[2013m] *International Trade and Invasive Alien Species.* http://www.cbd.int/invasive, Accessed June 2013. (Cited on page 12.)

[2013n] *Natural Products Information System (NAPIS).* http://whitepointsystems.com, Accessed February 2013. (Cited on pages 2, 32 and 33.)

[2013o] *Ontobee Web Portal.* http://www.ontobee.org, Accessed February 2013. (Cited on page 18.)

[2013p] *Society for Range Management.* http://www.rangelands.org, Accessed Febuary 2013. (Cited on page 10.)

[2013q] *The Convention on Biological Diversity (CBD).* http://www.cbd.int, Accessed September 2013. (Cited on pages 1 and 12.)

[2013r] *The Economics of Ecosystems and Biodiversity.* http://ec.europa.eu/environment/nature/biodiversity/economics/index_en.htm, Accessed March 2013. (Cited on page 10.)

[2014s] *The IUCN Red List of Threatened Species.* http://www.iucnredlist.org/, Accessed January 2014. (Cited on pages 2, 33 and 38.)

[2013t] *The Plant Ontology (PO).* http://plantontology.org, Accessed January 2013. (Cited on page 16.)

[2013u] *Types of Biodiversity.* http://www.aboutbioscience.org/topics/biodiversity, Accessed January 2013. (Cited on page 5.)

[2013v] *What is Biodiversity?* http://www.unep-wcmc.org/what-is-biodiversity_50.html, Accessed January 2013. (Cited on page 5.)

[2013w] *What is biodiversity?* http://iucn.org/what/biodiversity/about/, September 2013. (Cited on page 7.)

[2013x] *What threatens our biodiversity.* http://www.nhm.ac.uk/nature-online/biodiversity/what-is-threatening-biodiversity, Accessed February 2013. (Cited on page 11.)

[2004y] *World Population to 2300.* http://www.un.org/esa/population/publications/longrange2/WorldPop2300final.pdf, 2004. (Cited on pages iii and 12.)

[agi 2014] *Robustness Diagrams: An Agile Introduction.* http://www.agilemodeling.com/artifacts/robustnessDiagram.htm, Accessed November 2014. (Cited on page 56.)

[Agrawal 1994] R. Agrawal and R. Srikant. *Fast Algorithms for Mining Association Rules in Large Databases.* In 20th International Conference on Very Large Data Bases, pages 478–499. Morgan Kaufmann, Los Altos, CA, 1994. (Cited on page 46.)

[Bachman 2011] Steven Bachman, Justin Moat, Andrew W Hill, Javier de Torre and Ben Scott. *Supporting Red List threat assessments with GeoCAT: geospatial conservation assessment tool.* ZooKeys, no. 150, page 117, 2011. (Cited on pages 27 and 40.)

[Bassil 2012] Youssef Bassil. *A Data Warehouse Design for A Typical University Information System.* arXiv preprint arXiv:1212.2071, 2012. (Cited on page 31.)

[Benniamin 2008] A. Benniamin, V. Irudayaraj and V. S. Manickam. *How to Identify Rare and Endangered Ferns and Fern Allies.* Ethnobotanical Leaflets, vol. 12, pages 108–117, March 2008. (Cited on page 38.)

[Blanco 2011] J. Blanco and H. Kheradmand. Climate change - socioeconomic effects. IntechOpen, 2011. (Cited on pages 10 and 12.)

[Cardinale 2012] B. Cardinale. *Impacts of Biodiversity Loss.* Science, vol. 336, no. 6081, pages 552–553, 2012. (Cited on pages 7, 8 and 10.)

[Casalegno 2011] S. Casalegno. Global warming impacts - case studies on the economy, human health, and on urban and natural environments. InTech, 2011. (Cited on pages 10 and 11.)

[Chawathe 1994] Sudarshan Chawathe, Hector Garcia-Molina, Joachim Hammer, Kelly Ireland, Yannis Papakonstantinou, Jeffrey Ullman and Jennifer Widom. *The TSIMMIS project: Integration of heterogenous information sources.* 1994. (Cited on page 27.)

[COHAB 2010] COHAB. *The Importance of Biodiversity to Human Health.* Biodiversity and Global Health, vol. 1, October 2010. (Cited on page 10.)

[Conruyt 2012] N. Conruyt, D. Grosser and R. Vignes-Lebbe. *Knowledge Discovery for Biodiversity: From Data Mining to Sign Management.* In R. Seppelt, A.A. Voinov, S. Lange and D. Bankamp, editeurs, Proceedings of the 6th International Congress on Environmental Modelling and Software Managing Resources of a Limited Planet, Leipzig, Germany, July 2012. International Environmental Modelling and Software Society (iEMSs). (Cited on page 21.)

[Cox 2010] Barry Cox and Peter D. Moore. Biogeography: An ecological and evolutionary approach. John Wiley & Sons, 2010. (Cited on page 24.)

[Cruz 2005] Isabel F. Cruz and Huiyong Xiao. *The Role of Ontologies in Data Integration.* Journal of Engineering Intelligent Systems, vol. 13, pages 245–252, 2005. (Cited on pages 17 and 27.)

[Daconta 2003] Michael C Daconta, Leo J Obrst and Kevin T Smith. The semantic web: a guide to the future of xml, web services, and knowledge management. John Wiley & Sons, 2003. (Cited on page 16.)

[Dansereau 1957] Pierre Dansereau. Biogeography; an ecological perspective. Ronald Press Co., New York, USA, 1957. (Cited on page 24.)

[De Craene 2009] Louis Ronse De Craene and Livia Wanntorp. *Floral development and anatomy of Salvadoraceae.* Annals of botany, page mcp170, 2009. (Cited on page 52.)

[del Monte 1985] M del Monte and Ottavio Vittori. *Air pollution and stone decay: the case of Venice*. Endeavour, vol. 9, no. 3, pages 117 – 122, 1985. (Cited on page 8.)

[Doan 2012] AnHai Doan, Alon Halevy and Zachary Ives. Principles of data integration. Elsevier, 2012. (Cited on page 26.)

[Domenig 2000] Ruxandra Domenig and Klaus R Dittrich. *A query based approach for integrating heterogeneous data sources*. In Proceedings of the ninth international conference on Information and knowledge management, pages 453–460. ACM, 2000. (Cited on page 27.)

[Donoho 2010] Steve Donoho. *Link analysis*. In Data Mining and Knowledge Discovery Handbook, pages 355–368. Springer, 2010. (Cited on page 23.)

[EBI 2013] EBI. *Integrating Biodiversity into Environmental and Social Impact Assessment Processes*. Environmental and Social Impact Assessment, Feburay 2013. (Cited on pages 7 and 10.)

[Eldredge 2002] N. Eldredge. Life on earth: An encyclopedia of biodiversity, ecology, and evolution, volume 1 of *Life on Earth*. ABC-CLIO, 2002. (Cited on pages 1 and 7.)

[Everett 1944] CJ Everett. *Closure operators and Galois theory in lattices*. Transactions of the American Mathematical Society, vol. 55, no. 3, pages 514–525, 1944. (Cited on page 25.)

[Faith 2014] Daniel P. Faith and Laura J. Pollock. *Phylogenetic Diversity and the Sustainable Use of Biodiversity*. In Luciano M. Verdade, Maria Carolina Lyra-Jorge and Carlos I. Piña, editeurs, Applied Ecology and Human Dimensions in Biological Conservation, pages 35–52. Springer, 2014. (Cited on page 24.)

[Fritsch 2011] Peter W. Fritsch and Catherine M. Bush. *A New Species of Gaultheria (Ericaceae) from Mount Kinabalu, Borneo, Malaysia*. Novon: A Journal for Botanical Nomenclature, vol. 21, no. 3, pages 338–342, September 2011. (Cited on page 51.)

[Ganter 1999] Bernhard Ganter and Rudolf Wille. Formal concept analysis - mathematical foundations. Springer, 1999. (Cited on page 25.)

[Gaston 1992] Kevin J. Gaston and Robert M. May. *Taxonomy of taxonomists*. Nature, vol. 356, pages 281–282, 1992. (Cited on page 24.)

[Gaston 2004] Kevin J. Gaston and John I. Spicer. *Biodiversity: An Introduction (Second Edition)*. Oryx, vol. 38, pages 465–465, October 2004. (Cited on page 10.)

[Gotelli 2012] Nicholas J. Gotelli, Aaron M. Ellison and Bryan A. Ballif. *Environmental proteomics, biodiversity statistics and food-web structure.* Trends in Ecology & Evolution, vol. 27, no. 8, pages 436–442, 2012. (Cited on page 24.)

[Grillo 2011] O. Grillo and G. Venora, editeurs. Biological diversity and sustainable resources use. InTech, 2011. (Cited on pages 1, 7 and 13.)

[Groombridge 2002] Brian Groombridge and Martin Jenkins. World atlas of biodiversity: Earth's living resources in the 21st century. UNEP-WCMC, 2002. (Cited on page 5.)

[Guarino 1997] N. Guarino. *Understanding, Building, and Using Ontologies: A Commentary to "Using Explicit Ontologies in KBS Development", by van Heijst, Schreiber, and Wielinga.* International Journal of Human and Computer Studies, no. 46, pages 293–310, 1997. (Cited on page 16.)

[Guarino 1998] N. Guarino. *Formal Ontology and Information Systems.* In N. Guarino, editeur, Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS), pages 3–15, Trento, Italy, 1998. IOS Press. (Cited on page 15.)

[Hall 2009] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten. *The WEKA Data Mining Software: An Update.* SIGKDD Explorations, vol. 11, no. 1, 2009. (Cited on page 22.)

[Han 2011] J. Han, M. Kamber and J. Pei. Data mining: Concepts and techniques. Morgan Kaufmann Publishers Inc., San Francisco, USA, 3rd édition, 2011. (Cited on pages 2, 18 and 20.)

[Hitzler 2009] Pascal Hitzler, Markus Krötzsch and Sebastian Rudolph. Foundations of semantic web technologies. Chapman & Hall/CRC, 2009. (Cited on page 16.)

[Hochachka 2007] W. M. Hochachka, R. Caruana, D. Fink, A. Munson, M. Riedewald, D. Sorokina and S. Kellings. *Data-Mining Discovery of Pattern and Process in Ecological Systems.* The Journal of Wildlife Management, vol. 71, no. 7, pages 2427–2437, 2007. (Cited on pages 2 and 22.)

[Inmon 2005] William H Inmon. Building the data warehouse. John wiley & sons, 2005. (Cited on page 29.)

[Inthasone 2014] Somsack Inthasone, Nicolas Pasquier, Andrea GB Tettamanzi and Célia da Costa Pereira. *The BioKET Biodiversity Data Warehouse: Data and Knowledge Integration and Extraction.* In Advances in Intelligent Data Analysis XIII, pages 131–142. Springer, 2014. (Cited on page 25.)

[Jaiswal 2005] Pankaj Jaiswal, Shulamit Avraham, Katica Ilic, Elizabeth A. Kellogg, Susan McCouch, Anuradha Pujar, Leonore Reiser, Seung Y. Rhee, Martin M.

Sachs, Mary Schaeffer, Lincoln Stein, Peter Stevens, Leszek Vincent, Doreen Ware and Felipe Zapata. *Plant Ontology (PO): A Controlled Vocabulary of Plant Structures and Growth Stages: Research Articles.* Comp. Funct. Genomics, vol. 6, no. 7-8, pages 388–397, October 2005. (Cited on page 17.)

[Jøker 2000] Dorthe Jøker*et al. Hopea odorata Roxb.* Seed Leaflet-Danida Forest Seed Centre, no. 49, 2000. (Cited on page 51.)

[Karahalil 2005] U. Karahalil and S. Keles. *The Effects of Biodiversity Concerns on Economic Profits of Timber in Forest Management.* In Proceedings of the 7th Balkan Conference on Operational Research (BACOR), Constanta, Romania, May 2005. (Cited on page 10.)

[Khallaf 2011] Mohamed K. Khallaf. The impact of air pollution on health, economy, environment and agricultural sources. IntechOpen, 2011. (Cited on pages 7, 8, 10 and 11.)

[Kimball 2002] Ralph Kimball, Margy Ross*et al. The data warehouse toolkit: the complete guide to dimensional modelling.* US: John Wiley & Sons, 2002. (Cited on page 29.)

[Knapp 2004] Sandra Knapp, Lynn Bohs, Michael Nee and David M. Spooner. *Solanaceae – a model for linking genomics with biodiversity.* Comparative and Functional Genomics, vol. 5, no. 3, pages 285–291, 2004. (Cited on page 23.)

[Kuenne 2007] Christian Kuenne, Ivo Grosse, Inge Matthies, Uwe Scholz, Tatjana Sretenovic-Rajicic, Nils Stein, Andreas Stephanik, Burkhard Steuernagel and Stephan Weise. *Using data warehouse technology in crop plant bioinformatics.* J Integr Bioinf, vol. 4, no. 88, page 10, 2007. (Cited on pages 26 and 31.)

[Kwuida 2014] Léonard Kwuida, Rokia Missaoui, Abdélilah Balamane and Jean Vaillancourt. *Generalized pattern extraction from concept lattices.* Annals of Mathematics and Artificial Intelligence, pages 1–18, 2014. (Cited on page 17.)

[Lameed 2012] Gbolagade Akeem Lameed, editeur. Biodiversity enrichment in a diverse world. InTech, 2012. (Cited on page 11.)

[Lindquist 2012] E. J. Lindquist, R. D. Annunzio, A. Derrand, K. Macdicken, F. Achard, R. Beuchle, A. Brink, H. D. Eva, P. Mayaux, J. San-Miguel-Ayanz and H. J. Stibig. *Global forest land-use change 1990-2005.* Food and Agriculture Organization of the United Nations and European Commission Joint Research Center, 2012. (Cited on page 10.)

[MA 2005] Millennium Ecosystem Assessment MA. *Ecosystems and human well-being: biodiversity synthesis.* Washington, DC: World Resources Institute, 2005. (Cited on pages 1, 7 and 10.)

[Magurran 2011] A.E. Magurran and B.J. McGill. Biological diversity: Frontiers in measurement and assessment. Oxford biology. OUP Oxford, 2011. (Cited on page 7.)

[Magurran 2013] A.E. Magurran. Measuring biological diversity. Wiley, 2013. (Cited on page 7.)

[Malinowski 2009] Elzbieta Malinowski and Esteban Zimányi. *Advanced Data Warehouse Design.* 2009. (Cited on pages iii and 29.)

[Marbán 2009] O. Marbán, G. Mariscal and J. Segovia. *A Data Mining & Knowledge Discovery Process Model.* In Julio Ponce and Adem Karahoca, editeurs, Data Mining and Knowledge Discovery in Real Life Applications. InTech, Vienna, Austria, 2009. (Cited on pages 3 and 19.)

[Mariscal 2010] G. Mariscal, Ó. Marbán and C. Fernández. *A Survey of Data Mining and Knowledge Discovery Process Models and Methodologies.* The Knowledge Engineering Review, vol. 25, no. 2, pages 137–166, May 2010. (Cited on pages 3 and 19.)

[Mason 1950] Herbert L. Mason. *Taxonomy, Systematic Botany and Biosystematics.* Madroño, vol. 10, no. 7, July 1950. (Cited on page 24.)

[Midgley 2012] G.F. Midgley. *Biodiversity and Ecosystem Function.* Science, vol. 335, no. 6065, pages 174–175, 2012. (Cited on pages 1 and 7.)

[Mondal 2012] Kartick Chandra Mondal, Nicolas Pasquier, Anirban Mukhopadhyay, Ujjwal Maulik and Sanghamitra Bandyopadhyay. *A New Approach for Association Rule Mining and Bi-clustering Using Formal Concept Analysis.* In Petra Perner, editeur, Machine Learning and Data Mining in Pattern Recognition - 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings, volume 7376 of *Lecture Notes in Computer Science*, pages 86–101. Springer, 2012. (Cited on pages 3 and 46.)

[Nijkamp 2008] P. Nijkamp, G. Vindigni and P.A.L.D. Nunes. *Economic Valuation of Biodiversity: A Comparative Study.* Ecological Economics, vol. 67, no. 2, pages 217–231, 2008. (Cited on page 10.)

[Obrst 2003] Leo Obrst. *Ontologies for Semantically Interoperable Systems.* In Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM), pages 366–369. ACM, 2003. (Cited on pages 16 and 36.)

[O'Sullivan 2010] B. O'Sullivan, S. Keady, E. Keane, S. Irwin and J. O'Halloran. *Data Mining for Biodiversity Prediction in Forests.* In Proceedings of the 19th European Conference on Artificial Intelligence (ECAI), pages 289–294. IOS Press, 2010. (Cited on page 22.)

[Parr 2012] Cynthia S. Parr, Robert Guralnick, Nico Cellinese and Roderic D. M. Page. *Evolutionary informatics: Unifying knowledge about the diversity of life*. Trends in Ecology & Evolution, vol. 27, no. 2, pages 94–103, February 2012. (Cited on page 54.)

[Paterson 2004] T. Paterson, J.B. Kennedy, M.R. Pullan, A. Cannon, K. Armstrong, M.F. Watson, C. Raguenaud, S.M. McDonald and G. Russell. *A Universal Character Model and Ontology of Defined Terms for Taxonomic Description*. In Erhard Rahm, editeur, Data Integration in the Life Sciences, volume 2994 of *Lecture Notes in Computer Science*, pages 63–78. Springer Berlin Heidelberg, 2004. (Cited on page 22.)

[Pennisi 2000] Elizabeth Pennisi. *Taxonomic Revival*. Science, vol. 289, no. 5488, pages 2306–2308, 2000. (Cited on page 24.)

[Perrings 2010] C. Perrings. *Biodiversity, Ecosystem Services, and Climate Change - The Economic Problem*. Environmental Economics Series, November 2010. (Cited on pages 10 and 12.)

[Peters 2009] C. Peters, D. Peters and J.H. Cota-Sánchez. *Data Mining and Mapping of Herbarium Specimens using Geographic Information Systems: A Look at the Biodiversity Informatics Project of the W. P. Fraser Herbarium (SASK)*. `http://www.herbarium.usask.ca/research/Data%20Mining,%20CBA%202009.pdf`, 2009. (Cited on pages 22, 23 and 38.)

[Pipe 1995] R.K Pipe, J.A Coles, M.E Thomas, V.U Fossato and A.L Pulsford. *Evidence for environmentally derived immunomodulation in mussels from the Venice Lagoon*. Aquatic Toxicology, vol. 32, no. 1, pages 59–73, 1995. (Cited on page 8.)

[Plant 2010] Claudia Plant and Christian Bohm. Database technology for life sciences and medicine, volume 6. World Scientific, 2010. (Cited on page 16.)

[Popy 2009] S. Popy. *Définition des enjeux relatifs à la biodiversité en Languedoc-Roussillon*. Synthesis of the stakes related to biodiversity in Languedoc-Roussillon, pages 3–20, 2009. (Cited on page 5.)

[Porter 2008] C. L. Porter. Taxonomy of flowering plants. Blackburn Press, 2008. (Cited on page 24.)

[Raguenaud 2001] C. Raguenaud, M. Graham and J.B. Kennedy. *Two Approaches to Representing Multiple Overlapping Classifications: A Comparison*. In Proceedings of the 13th International Conference on Scientific and Statistical Database Management (SSDBM), pages 239–244, Fairfax, USA, July 2001. IEEE Computer Society. (Cited on page 23.)

[Rahangdale 2014] SR Rahangdale and Sanjaykumar R Rahangdale. *Plant species composition on two rock outcrops from the northern Western Ghats, Maharashtra, India.* Journal of Threatened Taxa, vol. 6, no. 4, pages 5593–5612, 2014. (Cited on page 51.)

[Ratnam 2014] W Ratnam. *Hopea odorata Roxb.* `http://www.worldagroforestry.org/treedb/AFTPDFS/Hopea_odorata.pdf`, Accessed March 2014. (Cited on page 51.)

[Sala 2003] Osvaldo E Sala. *(Almost) All About Biodiversity.* Science, vol. 299, no. 5612, pages 1521–1521, 2003. (Cited on page 5.)

[SCBD 1992] SCBD. *Convention on Biological Diversity.* Montreal, Canada, 1992. `http://www.cbd.int`. (Cited on page 24.)

[Seaborne 2008] Andy Seaborne, Geetha Manjunath, Chris Bizer, John Breslin, Souripriya Das, Ian Davis, Steve Harris, Kingsley Idehen, Olivier Corby, Kjetil Kjernsmo*et al.* *SPARQL/Update: A language for updating RDF graphs.* W3C Member Submission, vol. 15, 2008. (Cited on page 16.)

[Shah 2011] A. Shah. Why is biodiversity important? who cares? Global Issues, April 2011. (Cited on pages 1 and 7.)

[Silva 2013] Raquel A Silva and et al. West. *Global premature mortality due to anthropogenic outdoor air pollution and the contribution of past climate change.* Environmental Research Letters, vol. 8, no. 3, page 034005, 2013. (Cited on page 11.)

[Smith 2007] B. Smith, M. Ashburner, C. Rosse, C. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R.H. Scheuermann, N. Shah, P.L. Whetzel and S. Lewis. *The Open Biological and Biomedical Ontologies (OBO): Coordinated evolution of ontologies to support biomedical data integration.* `http://www.obofoundry.org`, November 2007. (Cited on page 18.)

[Spehn 2009] Eva M. Spehn and Christian Korner, editeurs. Data mining for global trends in mountain biodiversity. CRC Press, 2009. (Cited on pages 2, 22 and 23.)

[Staudinger 2012] Michelle D. Staudinger, N. B. Grimm, A. Staudt, S. L. Carter, F. S. Stuart, P. Kareiva, M. Ruckelshaus and B. A. Stein. *Impacts of Climate Change on Biodiversity, Ecosystems, and Ecosystem Services.* `http://assessment.globalchange.gov`, July 2012. (Cited on page 11.)

[Swingland 2001] Ian R. Swingland. *Biodiversity, Definition of.* Encyclopedia of Biodiversity, vol. 1, pages 377–391, 2001. (Cited on page 5.)

[Talent 2012] J.A. Talent. Earth and life: Global biodiversity, extinction intervals and biogeographic perturbations through time. International Year of Planet Earth. Springer, 2012. (Cited on pages 1, 5 and 7.)

[Thornton 2011] Daniel H. Thornton, Lyn C. Branch and Melvin E. Sunquist. *The relative influence of habitat loss and fragmentation: Do tropical mammals meet the temperate paradigm?* Ecological Applications, vol. 21, no. 6, pages 2324–2333, August 2011. (Cited on page 11.)

[Van Huis 2013] Arnold Van Huis, Joost Van Itterbeeck, Harmke Klunder, Esther Mertens, Afton Halloran, Giulia Muir and Paul Vantomme. Edible insects: Future prospects for food and feed security. United Nations Food and Agriculture Organization (FAO), 2013. (Cited on pages 12 and 13.)

[Van So 2000] Nguyen Van So. *The potential of local tree species to accelerate natural forest succession on marginal grasslands in Southern Vietnam.* Forest Restoration for Wildlife Conservation. Chiang Mai University, Thailand, pages 135–148, 2000. (Cited on page 51.)

[Wache 2001] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann and S. Hübner. *Ontology-Based Integration of Information - A Survey of Existing Approaches.* In Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI). Workshop on Ontologies and Information Sharing, 2001. (Cited on page 27.)

[Whetzel 2011] P.L. Whetzel, N.F. Noy, N.H. Shah, P.R. Alexander, C. Nyulas, T. Tudorache and M.A. Musen. *BioPortal: Enhanced Functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications.* Nucleic Acids Res., no. 39, pages W541–5, July 2011. (Cited on pages 18, 62, 63, 64 and 65.)

[Whetzel 2013] P.L. Whetzel, N.F. Noy, N.H. Shah, P.R. Alexander, C. Nyulas, T. Tudorache and M.A. Musen. *What are Ontologies?* `http://www.bioontology.org/learning-about-ontologies`, Accessed March 2013. (Cited on page 15.)

[Wilson 1992] Edward. O. Wilson. The diversity of life. Questions of science. Belknap Press of Harvard University Press, 1992. (Cited on page 54.)

[Xiang 2011] Z. Xiang, C. Mungall, A. Ruttenberg and Y. He. *Ontobee: A Linked Data Server and Browser for Ontology Terms.* In International Conference on Biomedical Ontologies (ICBO), pages 279–281, University at Buffalo, USA, July 2011. (Cited on page 18.)

[Yang 2006] Q. Yang and X. Wu. *10 Challenging Problems in Data Mining Research.* International Journal of Information Technology & Decision Making, vol. 5, no. 4, pages 597–604, 2006. (Cited on page 20.)

[Zannetti 1977] P. Zannetti, P. Melli and E. Runca. *Meteorological factors affecting SO2 pollution levels in Venice.* Atmospheric Environment, vol. 11, no. 7, pages 605–616, 1977. (Cited on page 8.)