University of Southampton

FACULTY OF MEDICINE, HEALTH AND LIFE SCIENCES

School of Biological Sciences

Data Quality Concepts and Techniques Applied to Taxonomic Databases

by

Eduardo Couto Dalcin

Thesis for the degree of Doctor of Philosophy

February 2005

University of Southampton

FACULTY OF MEDICINE, HEALTH AND LIFE SCIENCES

School of Biological Sciences

Doctor of Philosophy

Abstract

Data Quality Concepts and Techniques Applied to Taxonomic Databases

by

Eduardo Couto Dalcin

The thesis investigates the application of concepts and techniques of data quality in taxonomic databases to enhance the quality of information services and systems in taxonomy. Taxonomic data are arranged and introduced in Taxonomic Data Domains in order to establish a standard and a working framework to support the proposed Taxonomic Data Quality Dimensions, as a specialised application of conventional Data Quality Dimensions in the Taxonomic Data Quality Domains.

The thesis presents a discussion about improving data quality in taxonomic databases, considering conventional Data Cleansing techniques and applying generic data content error patterns to taxonomic data. Techniques of taxonomic error detection are explored, with special attention to scientific name spelling errors.

The spelling error problem is scrutinized through spelling error detecting techniques and algorithms. Spelling error detection algorithms are described and analysed. In order to evaluate the applicability and efficiency of different spelling error detection algorithms,

a suite of experimental spelling error detection tools was developed and a set of experiments was performed, using a sample of five different taxonomic databases. The results of the experiments are analysed from the algorithm and from the database point of view.

Database quality assessment procedures and metrics are discussed in the context of taxonomic databases and the previously introduced concepts of Taxonomic Data Domains and Taxonomic Data Quality Dimensions.

Four questions related to Taxonomic Database Quality are discussed, followed by conclusions and recommendations involving information system design and implementation and the processes involved in taxonomic data management and information flow.

CONTENTS

List of Tables

List of Figures

List of Appendices

List of Acronyms

API	Application Program Interface
ASCII	American Standard Code for Information Interchange
ATF	Alice Transfer Format
CDB	Convention on Biological Diversity
CHM	Clearing-house Mechanism
CITES	Convention on International Trade in Endangered Species of Wild Fauna and Flora
CNIP	Centro Nordestino de Informações sobre Plantas
CSV	Comma-Separated Variables
DBF	Database file (dBASE II/III+/IV, FoxPro, dBFast, DataBoss, etc. compatible database file)
DBMS	Database Management System
DNA	Deoxyribonucleic Acid
GBIF	Global Biodiversity Information Facility
GTI	Global Taxonomic Initiative
HISPID	Herbarium Information Standards and Protocols for Interchange Data
ICBN	International Code of Botanical Nomenclature
ICNB	International Code of Nomenclature of Bacteria
ICVCN	International Code of Virus Classification and Nomenclature
ICZN	International Code of Zoological Nomenclature
ILDIS	International Legume Database & Information Service
IPNI	International Plant Names Index
ITF	International Transfer Format
ITIS	Integrated Taxonomic Information System
KDD	Knowledge Discovery in Database
LD	Levenshtein Distance
LITCHI	Logic-based Integration of Taxonomic Conflicts in Heterogeneous Information Systems
NOAA	National Oceanic and Atmospheric Administration
NODC	National Oceanic Data Centre
OCR	Optical Character Recognition
PC	Personal Computer
PRECIS	National Herbarium Pretoria Computerised Information Service
TDQD	Taxonomic Data Quality Dimensions
TDWG	Taxonomic Database Working Group
UNEP	United Nations Environmental Program
URL	Uniform Resource Locator

1. General introduction

Taxonomic databases – databases that store information about biological entities: species and other taxa - have been developed to address curatorial management processes, taxonomic and scientific needs and, more recently, in response to the growing demand for a global plant information system.

In the last three decades considerable progress has been achieved, following advances in information technology, in modelling the complexity of taxonomic information and defining data standards and protocols which make it possible to interchange data and connect different databases. In some respects, biologists and taxonomists themselves have become more aware of the benefits of better data management.

In contrast, relatively little attention has been given to the data quality aspects of these databases. Without data quality policies, standards and strategies, taxonomic databases cannot fulfil their objectives.

1.1. Problem elucidation and delimitation

Since data became an important resource in the business world, data quality issues have become the object of studies and discussion. Some of the consensus about poor quality of data involves its impact on decision making, credibility of data, satisfaction of users, high costs of database management and the effective use and value of data (Redman, 1996).

There is a consensus as well that data quality is a multidimensional concept which involves, even at a simplified level, data modelling, data management, data presentation and the data values themselves (Redman, 1996). On the other hand, there is no agreement about the dimensions of data quality (Embury, 2001).

Regardless of the multidimensional aspect of data quality, this thesis will emphasize the data quality dimensions that are most pertinent to data values.

The literature on data quality shows that, unless extraordinary efforts have been made, we should expect data (field) error rates[1] of approximately 1-5% (Redman, 1998). A data field is in error if, for a variety of possible reasons, it does not correctly describe a property of the real world, resulting in a reduction in data quality. This poor data quality can lead to a number of potential impacts.

There is no reason to imagine that taxonomic databases are immune to this "poor data quality phenomenon". The impact of poor quality data in taxonomic databases may be seen in different ways, including:

- Pervasiveness of poor data

  Taxonomic information is the core of many other areas of related biological studies. Agriculture, plant breeding, genetic resources, biotechnology, pharmacognosy, forestry and ecology are some examples of sciences that rely on

[1] Error rate = number of incorrect fields/total number of fields

taxonomic data. Poor taxonomic data can “contaminate” related study areas. Some examples of the impact of contamination by poor data:

- Poor biodiversity knowledge and assessment
- Erroneous decision-making and conservation strategies

- Troublesome data and collection management

Taxonomic databases that support curatorial activities can be affected by poor data quality in several ways. As data quality dimensions encompass more than just the aspect of the accuracy of the data field (which itself is a source of enough trouble in any database), dimensions such as timeliness, the concise and consistent representation of the biological facts, accessibility and completeness have a significant impact on curatorial databases and collection management. For example, incorrect information about loans can cause loss of material; incorrect record classification for a specimen may compromise the organization of the collection for future taxonomical updates, etc.

- Difficult data integration and database merging

Merging taxonomic databases from different sources has become increasingly important in the last decade, in order to create regional, national or global species checklists. Data integration has become more and more important to achieve institutional goals, involving database merging from different scientists, projects and institutions.

In addition to the complexity of interpreting nomenclatural data, poor data quality affects both the data merging process and data merging results. For example, a misspelled scientific name may prevent the recognition and merging of data from two sources about the same species, so that two apparently different species appear in the merged result.

- Scientific and institutional reputation

In fulfilment of their mission to understand, explain, quantify and evaluate world biodiversity, scientists and scientific institutions are recognized as scientific information providers. This recognition is based on their ability to provide reliable information to the general public, decision-makers and colleagues. Ambiguous, confused, incomplete, contradictory and erroneous information, published in reports obtained as a result of searching their databases, can affect their reputation as information providers and scientific authorities.

At this point, some questions arise: How "good" is the information provided by a database? How can the "bad data" be detected? How can the data quality of a taxonomic database be improved? These are the issues which this thesis will address.

To understand the data quality aspects and their impact on taxonomic databases, it is essential to establish a precise delimitation and definition of taxonomic data (*taxonomic data domains*) in the context of data quality and data quality dimensions. After the definition of "Taxonomic Data Quality Dimensions", as a product of the interaction of

taxonomic data domains and the data quality dimensions, *taxonomic database quality metrics* can be proposed.

The improvement of data quality has two basic approaches: error detection (and correction) and error prevention (Embury, 2001). Error prevention is closely related to the processes of data acquisition and data entry, naturally prone to errors. Much effort can be given to improving these processes, with respect to reduction in entry error, but the fact often remains that errors in large data sets are common (Maletic and Marcus, 2000).

For existing data sets, the logical solution to this problem is to attempt to cleanse the data in some way. In most cases, to explore the data set for possible problems and endeavour to correct the errors “by hand” is completely out of the question given the amount of person hours involved (Maletic and Marcus, 2000).

In a taxonomic database, a regional checklist for example, this kind of task is too complex because it would involve, in most cases, a large number of specialists, each one responsible for scrutinizing their own area of knowledge.

A manual process of data cleansing is also laborious, time consuming, and itself prone to errors (Maletic and Marcus, 2000). Useful and powerful tools that automate or greatly assist in the data cleansing process are necessary. Because of the specialized nature and complexity of the representation and modelling of taxonomic knowledge, dealing with subjective biological concepts, such as synonymy, a combination of data cleansing concepts and taxonomic data quality dimensions is a requisite to achieve

efficient error detection algorithms and tools and error prevention techniques and procedures.

1.2. Aims

The general aims for this thesis are:

A. To propose a set of Taxonomic Database Quality Metrics;
B. To investigate and extend error detection algorithms and tools to a taxonomic context;
C. To propose error prevention and correction techniques to be applied in taxonomic databases.

These aims correspond with the products shown in green circles, and lettered A-C, in Figure 1.

1.3. Thesis plan

An introduction to Taxonomic Databases and Data Quality is given in chapters 1, 2 and 4. The general aims will be achieved by:

1. The definition of *Taxonomic Data Domains* based on the available taxonomic databases and literature (chapter 3);
2. The proposition of a set of *Taxonomic Data Quality Dimensions*, based on adaptation of data quality dimensions to a taxonomic context (chapter 5);
3. The definition of *Taxonomic Error Patterns*, based on existing general error pattern definitions, adapted to the context of the *Taxonomic Data Quality Dimensions* (chapter 6);

4. Testing and adapting existing data cleansing techniques in the taxonomic database context (chapter 7).
5. Taxonomic data quality metrics are discussed in chapter 8 and the use of these and other techniques for error prevention and correction are discussed in chapter 9 and used to formulate a series of recommendations in chapter 10.

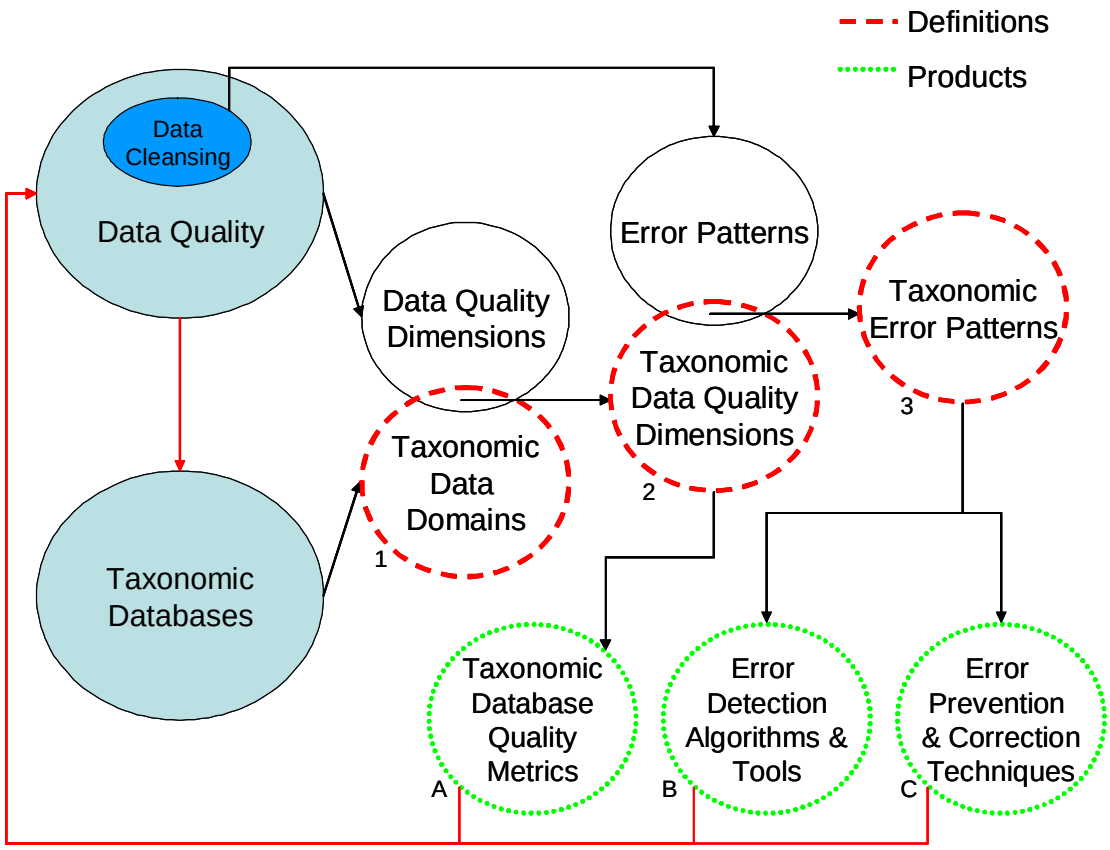


Figure 1- Graphical representation of the Thesis plan

The graphical representation of the thesis plan above identifies the new concepts and definitions 1-3 proposed in this thesis, in red circles, and the resulting practical products, in green circles, as result of the practical application of these concepts.

Summarizing Figure 1, this thesis will define the concept of *Taxonomic Data Domains* (1). The overlap of this concept with *Data Quality Dimensions* will give a new

definition – *Taxonomic Data Quality Dimensions* (2), which can be applied with the concept of *Error Patterns*, to define the concept of *Taxonomic Error Patterns* (3). Based on the new concept of *Taxonomic Data Quality Dimensions*, products and recommendations will be presented in order to promote the *Taxonomic Database Quality Metrics* (A). From the new concept of *Taxonomic Error Patterns*, both *Error Detection Algorithms and Tools* (B) and *Error Prevention and Correction Techniques* (C) can be defined and, in the last instance, applied (in the red arrow) to improve *Data Quality* in *Taxonomic Databases*.

2. Taxonomic databases

2.1. Overview

Taxonomic databases have become, in recent years, a fundamental tool to support the science of giving names to living organisms. A much more significant role has been to provide a framework for organizing biological data and as a tool for the storage, retrieval and, more recently, publishing of biological information. The literature shows that the utilization of “electronic data processing” techniques to organize, analyze and manage taxonomic data began around the nineteen-sixties when numerical taxonomy (Sokal, 1963; Sokal and Sneath, 1966), the organization of national floras (Perring, 1963), institutional scientific collections (Crovello, 1967; MacDonald, 1966; Soper and Perring, 1967; Squires, 1966) and scientific nomenclature (Silva, 1966; Stafleu, 1966; Wood Jr. *et al*., 1963) were examples of the first efforts to build structured tables and matrices of taxonomic data.

The necessity to achieve a better management of collections in museums and herbariums motivated the first efforts in the direction of developing tools based on databases. Despite the difficulties, the advantages of applying automated systems to carry out repetitive curatorial procedures, as for example, label printing, were attractive enough to start the process of computerization of these collections. At this time taxonomic databases were basically institutional initiatives or based on floristic survey projects. Most of these initiatives were based on minicomputers or mainframes, using “time-sharing” operating systems. These databases were known as “curatorial databases” since their main function was to facilitate the handling and management of

scientific collections (Crovello, 1967; MacDonald, 1966; Soper and Perring, 1967; Squires, 1966).

In the seventies, the evolution of database systems for handling collections in museums and herbariums continued (Argus and Sheard, 1972; Beschel and Soper, 1970; Brenan and Williams, 1975; Hall, 1972; 1974; Vit *et al*., 1977; Wetmore, 1979) and was followed by new and important techniques for the development of taxonomy. Numerical taxonomy, the elaboration of algorithms for generating diagnostic keys, and demand for the automated identification of specimens led to the creation of new applications for more modern, powerful and accessible computers (Sneath and Sokal, 1973). At this time we notice, by literature review, the creation of databases to handle specimen descriptive characteristics (Morse *et al*., 1971).

Data on morphology was, at this time, a target for species descriptive databases in which the objective was to supply data to the different applications able to analyze these data, to produce, in an automated form, descriptions and diagnostic keys, and to classify species automatically (Dallwitz, 1980; Jardine and Sibson, 1977; Johnston, 1980; Pankhurst, 1974; 1975; Sneath and Sokal, 1973). Although some of these applications used data organized in simple structured text files or electronic spreadsheets, instead of the use of a database, we can consider these files as the embryo of present day descriptive species databases. Later on, these structured text files were generated by exports from descriptive monographic databases (Adey *et al*., 1984; Allkin and Bisby, 1988; Pankhurst, 1988).

We can say that descriptive databases of specimens, rather than species, were a natural evolution of the curatorial databases. Herbarium labels represent the basic certification of the biological fact: the existence of a determined specimen. However, in the great majority of cases, the data contained in these labels are insufficient to characterize, and consequently to identify and to classify, the specimen.

Curatorial databases, responsible for storage and management information about specimens, are “factual” databases. The collection of the specimen, represented by the label annotations, is a fact without any dependence on considerations or opinions. In the same way, descriptive databases of specimens, which contain morphometric information, for example, are “factual” as well. However, lost label data cannot be replaced by examination of the specimen.

In the seventies, the organization of data about specimens, especially their geographical occurrence, frequently the consequence and product of the curatorial databases, led to the development of the first spatial or geo-codified databases. These databases have the capacity to represent, associated with graphical and mapping applications, through the creation of maps and graphs, the spatial distribution of the taxa, either for political divisions, as countries or continents, or for different ecosystems or floristic regions (Adams, 1974).

The organization of descriptive data about species had a very clear taxonomic objective: the identification and classification of the species into taxonomic groups. After all, the observation and measurement of morphological and structural characteristics of living organisms is the basis of taxonomic work. Specimens with the same characteristics are

congregated in hierarchically structured and related groups. The main products of these databases are diagnostic keys, taxonomic descriptions and models of classification. The results of these analyses lead, of course, to a demand for other types of databases, such as species descriptive databases and nomenclatural databases.

The literature shows, at this time, general descriptions and discussions about the utilization of computers to handle taxonomic data (Krauss, 1973; Morse, 1974; Shelter, 1974; Watson, 1971).

One of the most significant facts for the development of taxonomic databases, still in the seventies, occurred in the science of information: the development of the relational data model (Codd, 1970). The relational data model, with the evolution of database tools – database management systems (DBMS) - and the development of the personal computer, at the beginning of the eighties, definitively amalgamated the science of information with taxonomy, stimulating the development of a profusion of studies, articles and initiatives in the area of taxonomic databases in the following decade (Allkin, 1988; Allkin and Bisby, 1988; Barron, 1984; Bisby, 1984a; b; 1988; Freeston, 1984; Heywood, 1984; Pankhurst, 1988).

The eighties, stimulated by the popularity of personal computers and by the rise of "friendly" database management systems, provided innumerable taxonomic database initiatives. The ability to create a taxonomic database was available to everyone. Many of these initiatives had been developed by researchers themselves (Adey *et al*., 1984).

However, it became clear that the ease of access to the tools and the object of study in itself was no guarantee for the development of systems that could represent the real world of the complexity, and frequently the subjectivity, of taxonomic information. As the curatorial databases become more and more mature, the difficulties in creating taxonomic databases with information about species, especially the ones that intended to deal with nomenclature and classification, became ever clearer. Diverse “types” and “subtypes” of taxonomic databases appeared, especially those that were “species oriented”.

Species-oriented taxonomic databases have the taxon as the principal entity, represented by its main identification: the scientific name. In contrast to specimen-oriented databases, species-oriented databases represent generalizations based on observations made from a limited number of specimens available at herbariums and museums. Compilations of data about specimens are generalized for the taxonomic groups to which they belong (Adey *et al*., 1984; Allkin, 1984a).

Species-oriented databases can assume extremely simple forms, such as a “checklist” for one defined region; more complex forms, such as nomenclatural databases, have details about the accepted or valid names, synonyms, authorities and references, for example; and descriptive databases about species have, associated with the species names, information on vernacular names, geographic distribution, morphology, uses, chemical components, chromosome number etc.

It is important to notice that, from a data quality point of view, the information about a species is a compilation from the union of all the information available about specimens

which belong to this taxon; thus the reliability of the information is directly proportional to how representative is the group or number of specimens studied, which in turn is affected by the quality of the manual or automated procedures which placed the specimens in the taxon. Obviously, it is impossible for a taxonomist to observe and measure all the existing individuals of one determined taxon, in the past, present and future. Thus, taxonomic information is inherently of a lower degree of reliability than actual specimen observations, although for well-studied groups this difference is often negligible (Morse, 1974).

The 1980's brought great advances in taxonomic databases. Questions previously neglected were presented and discussed. The vision of the user, the need to transmit taxonomic knowledge efficiently and effectively, the difficulties in establishing an efficient data model and the need for "taxonomically intelligent database programs" (Allkin, 1988) were raised (Allkin, 1980; 1987; 1988; 1989; Allkin and White, 1989; Barron, 1984; Bisby, 1984b; 1988; Gómez-Pompa and Nevling Jr, 1988; IUCN-BGCS, 1987; Pankhurst, 1988).

Since more and more taxonomic databases were appearing, both institutional and individual, concern about sharing data was rising. At this moment the need to establish data standards and communication protocols was obvious, in order to make possible data sharing between different databases (Allkin, 1987; Allkin and White, 1989; Croft, 1989; IUCN-BGCS, 1987).

In September 1985 the first meeting of the "Taxonomic Database Working Group" (TDWG) took place. According to its constitution, the TDWG "is a not-for-profit

scientific and educational association formed to establish international collaboration among biological database projects so as to promote the wider and more effective dissemination of information about the world's heritage of biological organisms for the benefit of the world at large".

At that time TDWG was able to bring together researchers involved in the development and management of taxonomic databases, promoting meetings and publication of specific works and giving special emphasis to the definition of standards by the creation of specialized working groups.

Within the scope of data standards and communication protocols it is important also to note the establishment of the "Herbarium Information Standards and Protocols for Interchange Data" – HISPID – as a standard format for the interchange of electronic herbarium specimen information. HISPID has been developed by a committee of representatives from all major Australian herbaria. This interchange standard was first published in 1989, with a revised version published in 1993 (Croft, 1989; Whalwn, 1993). In 1987 the Botanic Gardens Conservation International published the International Transfer Format (ITF), a protocol to interchange data about live collections in botanical gardens (IUCN-BGCS, 1987).

The 1980's saw the appearance or consolidation of "project-driven" databases and "software packages" designed to meet the demand for organization, management and analysis of taxonomic data (Allkin and Winfield, 1990; Bosbach *et al.*, 1990; Johnston, 1980; Pankhurst, 1986; Partridge *et al.*, 1986; Reznicek and Estabrook, 1986; Vogellehner and Speck, 1988).

As examples of “project-driven” taxonomic databases we can cite the International Legume Database and Information Service (ILDIS), Flora de Veracruz, Flora of Mount Kinabalu, TROPICOS and PRECIS, among others (Allkin, 1986; Beaman and Regalado Jr., 1989; Bisby, 1986; Gibbs-Russell and Arnold, 1989; Gibbs-Russell and Gonsalves, 1984; Gómez-Pompa *et al*., 1984; Gómez-Pompa *et al*., 1985; Gómez-Pompa and Nevling Jr, 1988; Moreno and Allkin, 1981; 1988; Morin and Gomon, 1993).

Taxonomic databases had also acquired in this decade the status of institutional databases, in the sense that the scientific institutions had recognized the strategic importance of the information for their mission, and the balance between the cost and the benefits of adopting computer systems became more favourable (Crovello *et al*., 1984; Mackinder, 1984).

The beginning of the 1990’s greatly changed the context in which taxonomic databases were being discussed. In response to the threat to the earth’s biological resources, the United Nations Environment Program (UNEP) convened a group in 1988 to explore the need for an international convention on biological diversity. By 1991, the working group of experts on biological diversity had become known as the Intergovernmental Negotiating Committee, and its work culminated in May 1992 with the Nairobi Conference for the Adoption of the Agreed Text on the Convention on Biological Diversity (CBD) (Duff, 1997; Juma, 1997).

The Convention was open for signature by all States and regional economic organizations at the Rio de Janeiro United Nations Conference on Environment and

Development in June 1992 (“The Earth Summit”), and remained open for a year thereafter. During that time, it received 168 signatures, and by December 1993 the Convention entered into force with ratification from almost 20% of those signatories.

The world community’s growing commitment to sustainable development inspired the CBD, and it now represents a dramatic step forward in the conservation of biological diversity and the equitable sharing of benefits arising from the use of genetic resources.

The CBD includes in its scope, two aspects that have a strong relationships with taxonomic databases: The Clearing-house Mechanism (http://www.biodiv.org/chm/) and the Global Taxonomy Initiative (http://www.biodiv.org/programmes/cross-cutting/taxonomy/).

The general aim of the Clearing-house Mechanism (CHM) is to promote both national and international technical and scientific cooperation concerning the exchange of information on biodiversity. Article 17 of CBD provides a broad information mandate to facilitate the exchange of technical, scientific and socio-economic research, specialized knowledge, indigenous and traditional knowledge, training programmes, and the repatriation of knowledge.

The Global Taxonomic Initiative (GTI) is concerned with removing or ameliorating the so-called “taxonomic impediment” – that is, the knowledge gaps in our taxonomic system, the shortage of trained taxonomists and curators, and the impact these deficiencies have on our ability to conserve, use and share the benefits of our biological diversity.

With the CBD, and other international agreements such as CITES and "Agenda 21", all aspects of taxonomic information – sources, generation, maintenance, management, repatriation, sharing and accessibility - received global attention and consequent international political support and priority, in the 1990's.

This global focus and demand for biodiversity information, together with the phenomenal development of the Internet and the World Wide Web, launched taxonomic databases into a new era and, especially, gave them much greater visibility (Beach *et al*., 1993; Bisby, 2000; Burley *et al*., 1997; Canhos *et al*., 1997; Carling and Harrison, 1997; Green, 1994; Hagedorn and Rambold, 2000; Wilson, 2001). Taxonomic databases, in different formats and presentations, became available on the World Wide Web and the consumers of biodiversity information multiplied in unprecedented variety. "Global Species Databases" and "Federated Architecture" became new words in the vocabulary.

However, global analysis and assessment need a global information system, capable of interconnecting different taxonomic databases around the world to deliver a concise, precise and objective answer to a given question. By the end of the 1990's, this "global biodiversity information system" had become one of the major challenges for biologists and computer scientists (Bisby, 1993; 2000; Bisby and Brandt, 1999; Everard, 1993; Green, 1994; Sugden and Pennisi, 2000; Wilson, 2000).

International organizations and consortiums, such as the Global Biodiversity Information Facility (GBIF), and different approaches to the compilation of a global species checklist, such as Species 2000 (http://www.sp2000.org) and the "All Species Foundation" (http://www.all-species.org) came on to the scene.

The technical challenge to achieve this level of interoperability is not restricted to data standards, data models, networking and interface issues. Merging or linking heterogeneous biological databases which may be based on differing taxonomic treatments arose as a practical problem and a target for projects such as LITCHI (Brandt, 1999; Embury *et al*., 2001; Embury *et al*., 1999; Jones *et al*., 2001; Sutherland *et al*., 2000).

Almost three decades of intellectual effort, associated with technological advances in the database scenario, gave to taxonomic database models new approaches in order to attempt a better representation of the complexity of classifications (Berendsohn, 1995; 1997; Berendsohn *et al*., 1999; Graham *et al*., 2002; Pullan *et al*., 2000; Raguenaud *et al*., 1999; Raguenaud *et al*., 2002; Zhong *et al*., 1996).

At the beginning of the new century, taxonomic databases are becoming a pivot of the "Holy Grail" of global biodiversity knowledge and sustainable management. Taxonomic information breaks through the barrier of the specialized world of scientific taxonomy to achieve public visibility and social demand via the World Wide Web (Burley *et al*., 1997).

3. Taxonomic data domains

In order to define a workable framework for establishing a link with data quality concepts in the next chapter, the concept of “data domains” will be explored.

The concept of a “domain” can be understood as “an area under one rule”. Taxonomic data domains will be defined to group together and delimit different taxonomic data types which relate to common sets of data quality concepts and rules. Examples of the different types of domains will be discussed below.

The *data domain* can be expressed as a set of data elements that are related to one another. It should be clear that classifying different data types in the same data domain is for the purpose of treating taxonomic data in the context of data quality and not as a data modelling exercise. The difference between *data model* and *data domain* is that the first has a straightforward relationship with the concept, logical design and implementation of a database and its structure, while *data domain* is a concept to aggregate data elements in order to relate with *data quality dimensions*.

3.1. Nomenclatural data domain

The classification of living things is called systematics, or taxonomy, and should reflect the evolutionary trees (phylogenetic trees) of the different organisms. Taxonomy combines organisms into hierarchical groups called taxa, while systematics seeks their relationships. The dominant system is called Linnaean taxonomy, which includes ranks and binomial nomenclature. How organisms are named is governed by international

agreements such as the International Code of Botanical Nomenclature (ICBN), the International Code of Zoological Nomenclature (ICZN), and the International Code of Nomenclature of Bacteria (ICNB). A fourth Draft BioCode was published in 1997 (http://www.rom.on.ca/biodiversity/biocode/biocode1997.html) in an attempt to standardize naming in the three areas, but it does not appear to have yet been formally adopted. The International Code of Virus Classification and Nomenclature (ICVCN) remains, until now, outside the BioCode.

A nomenclatural data domain can be defined as a set of name elements, such as genus name, specific epithet and authority, related together by a pre-defined set of rules, based on the international codes cited above. For plant names, these name elements were represented by Bisby (1995).

One notable characteristic of the nomenclatural data domain, in the data quality context, is the ambiguity of the data type elements. Redman (1996) considered three types of data: labels, categories and quantities. *Labels* are data items created primarily for the purpose of naming or identifying an entity; *Categories* are data items that indicate specific categories within a classification scheme; and *quantities* are data items that are results of measurement in well-defined units or counts. Despite the binomial representation of a scientific name it behaves as a *label*, and its elements, with the exception of the specific epithet, can be considered as categories.

In the real world, the representation of scientific names in databases has a large variation in structure and taxonomic precision. These variations are related to the

objectives of the databases and the ability of these databases to represent the nomenclatural structure.

3.2. Classification data domain

Two major fields of systematics are *classification* and *nomenclature*. Classification is the process of establishing and defining systematic groups. The systematic groups so produced are known as taxa (singular, taxon). Nomenclature is the allocation of names to the taxa so produced. In carrying out their researches, systematists first complete their classificatory work. Only when they are sure they have achieved, on the basis of the information available, the best possible systematic arrangement of the organisms they have studied, do they begin to ascertain the correct names for the taxa they have established (Jeffrey, 1977).

The commonly used identifier for a taxon is its name, but the same name may be applied to different, non-congruent concepts of a taxon. The difference may range from outright misapplication of a name to slight disagreement over the circumscription of a taxon (i.e., its boundaries against other taxa) (Berendsohn, 1997).

Some of the most meticulous species-oriented taxonomic databases go beyond a simple list of scientific names. They store and handle additional data elements, such as status and synonyms, which are related to the taxonomic treatment of scientific names.

Despite the different approaches that have been taken in order to model the complexity of taxonomic classification (Berendsohn, 1995; Pullan *et al*., 2000; Raguenaud *et al*.,

2002; Zhong *et al*., 1996), the following common data elements related to classification information can be associated with the classification data domain:

- Name status (accepted, provisional, etc.)
- Nomenclatural status (synonym, homonym, etc.)
- Taxonomic reference citation (bibliographic reference: author, place and date of publication, etc.)
- Any other data elements that represent classification information

3.3. Field data domain

Berendsohn *et al*. (1999), define *field data* as information about the "who" and "where" of a collection or field record, descriptors resulting from field observations and the collection event itself. Field data elements are related to collections databases (herbarium and living collections) and databases that support floristic surveys, ecological and phytosociological studies.

In the context of data quality, sub-domain levels of the field data domain are desirable.

3.3.1. Spatial data sub-domain

The spatial data sub-domain refers to data elements, in the field data domain, which represent the geographic position of the observation or the collection event and its environmental characteristics. Some examples of *spatial sub-domain* are names of countries, regions or localities (*categories*), geographic coordinates (*quantities*) and general description of the site (*quantities*, *labels* and *categories*).

3.3.2. Curatorial data sub-domain

In the curatorial data sub-domain we can aggregate data elements which comprise heterogeneous but related sources, in the context of data quality. The curatorial sub-domain includes the following data elements:

- Identification events (i.e. determination author, determination date)
- Related collection management and curatorial activities data (i.e. loans, voucher ID, voucher localization)

3.3.3. Specimen descriptive data sub-domain

The specimen descriptive data sub-domain includes all data elements which are the result of a process of observation or analysis carried out on the specimen. Morphological, physiological and phenological data elements are examples of data in this sub-domain. In taxonomic databases, a specimen descriptive data domain is related to curatorial databases as a representation of the specimens' collectors and their notes. Specimen descriptive data sub-domains are present also in structured data, not necessarily in a database system, to be used in applications in order to generate, for example, cladistic analyses[2] and automatically generated descriptions or identification tools. Examples of specimen descriptive data are height (generally used for trees), flower colour, presence of fruits, measures of leaves, etc. The collector name, collection date and expedition data belongs to this sub-domain as well.

[2] A cladistic analysis attempts to reconstruct the evolutionary history of a group of organisms

3.4. Species descriptive data domain

In the species descriptive data domain we group all descriptive data that results from compilation of descriptive data from the specimens of a given species. Species descriptive data are directly related to species oriented databases. The importance of species descriptive data was pointed out by Bisby (1988) as “The simplest and probably the commonest type of taxonomic communication is providing a user with information or further information about a taxon”. Examples of species descriptive data can be a morphological description, geographical distribution, and “common knowledge” data such as uses, life-form, vernacular names, habitat, etc, and may extend into other areas such as chemical, behavioural, or ecological data (Bisby, 1988).

4. Data quality

4.1. Data quality definitions and principles

Before beginning to establish a framework for studies in taxonomic data quality, it is necessary to introduce data quality concepts and definitions previously proposed. Despite the several different definitions, studies and approaches to *Data Quality* (DQ), what seems to be a consensus is that any quality concept can only be applied at the moment where the data are used for some purpose. A widely adopted definition of high-quality data is “data fit for use” (Strong *et al*., 1997).

Data in a database has no actual value (or even quality); it only has *potential* value. Data has *realized* value only when someone uses it to do something useful (English, 1999). The quality of data cannot be assessed independently of the people who use the data – data consumers (Strong *et al.*, 1997).

At this point, it should be clear that data quality issues go beyond the data values themselves, and one way to categorize these issues is as follows (Redman, 1998):

- Issues associated with data “views” (models of the real world captured in the data) such as relevance, granularity and level of detail;
- Issues associated with data values, such as accuracy, consistency, currency and completeness;
- Issues associated with the presentation of data, such as the appropriateness of the format, ease of interpretation and so forth;
- Other issues such as privacy, security and ownership.

Despite the obvious dependence between the data quality and data use, English (1999) classified data or information quality issues according to two different standpoints:

- **Inherent information quality** – The degree to which data accurately reflect the real-world object that the data represent or, simply stated, the data accuracy;
- **Pragmatic information quality** – The degree of usefulness and value the data has to support the enterprise or institution processes that enable enterprise or institution objectives to be accomplished. In essence, it is the degree of consumer satisfaction derived by the data consumers.

This classification recognizes that some aspects of data quality can be tackled without the dependence, and consequently subjectivity, of the data use. Because the study of all aspects of information quality in sufficient depth would be impossible, with the time and resources available, this thesis will therefore concentrate on inherent information quality.

In order to better understand and define data quality issues that affect taxonomic databases, we must understand the processes and activities involved in taxonomic data creation, management, usage and delivery. For the same reason, we must recognize the "actors" involved in this process.

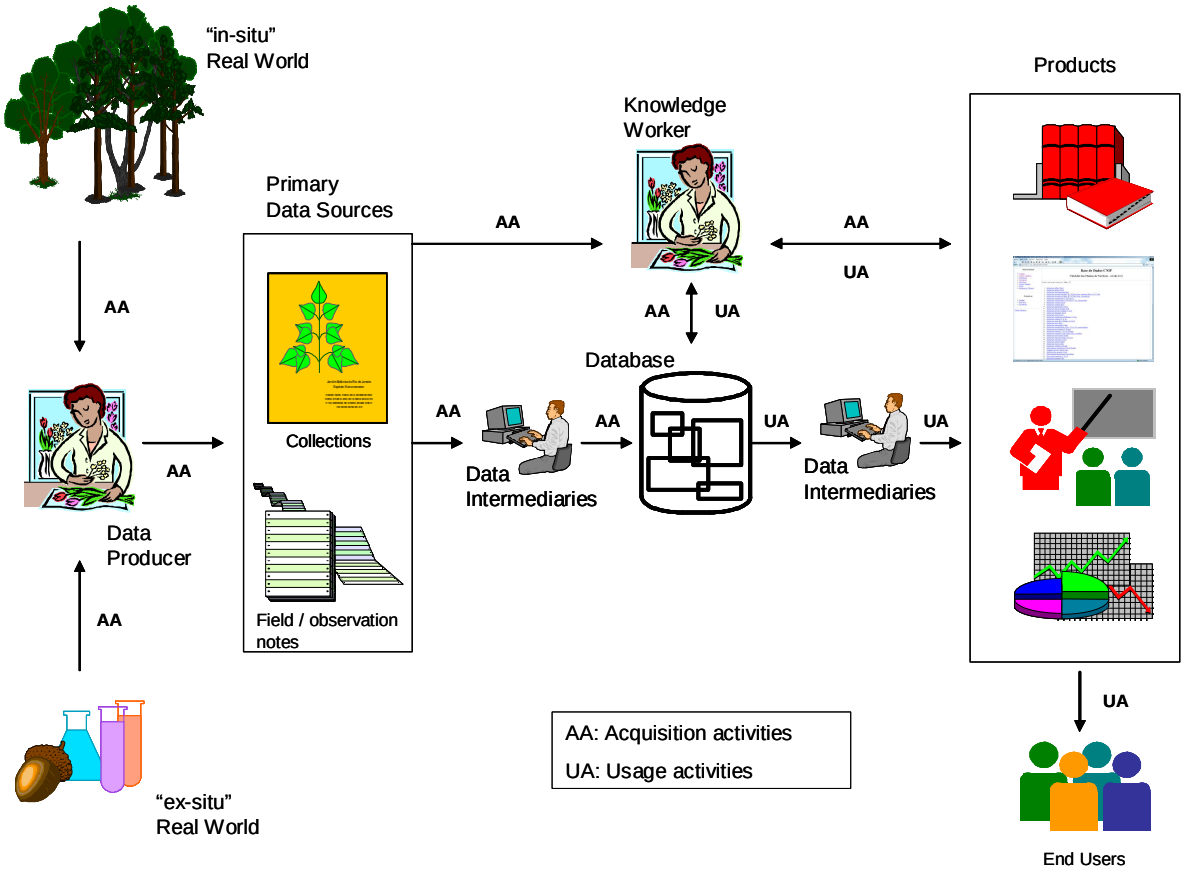


Figure 2 – Representation of the "taxonomic information chain"

For the purpose of this work, the figure above illustrates, at a basic level, the processes involved in taxonomic database creation and taxonomic information production and delivery. This set of processes is also known as the *Information Chain* (English, 1999; Redman, 1996), and can be related to the *life-cycle model* proposed by Redman (1996) which has *data storage* as the middle activity, related to *data acquisition* activities and *data usage* activities. Notice that the model presented above does not intend to detail all aspects and complexity of the taxonomic information chain. But the level of detail adopted is appropriate as a reference in the following parts of this Thesis.

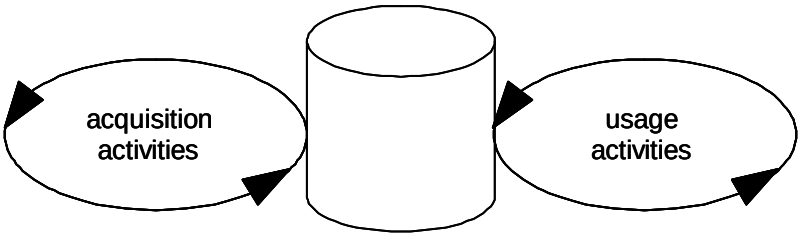


Figure 3 - Data Life Cycle Model (Redman, 1996)

The reason to adopt this conceptual model to illustrate the activities involved with taxonomic databases is because, despite its simplicity, it is perfectly applicable within the taxonomic database context and provides an understanding about how the activities are related with the database (data storage) itself. According to Redman's model, the following activities can be defined:

Acquisition Activities:

1. **Define a view** – According to the prevailing approach to data modelling, specification of a view is a required first step in dealing with data. A view defines the "part of the real world" to be captured in the data. One or more entity classes have to be specified, each defined by a set of attributes;
2. **Implement the view** – A view is merely a set of definitions. Thus, it must be implemented. This step should take into account restrictions and/or limitations imposed by the storage medium and by the data management system;
3. **Obtain values** – This step deals with acquiring specific values for the attributes of individual instances of the defined entity classes. Conceptually this activity is straightforward, but its importance and impact in data quality have been underestimated. Failure to take this into account lies at the root of many data quality issues. A incomplete list of the ways data values may be obtained includes measurement, surveys, observation, and copying from another source;

4. **Update records** – Here, data are stored in some medium. We use the term "update" in its broad sense, to include addition of a new record, deletion, and modification of existing records.

Usage Activities:

1. **Define a sub-view** – Typically a usage process will make use of only a small fraction of the data available. Thus, usage begins with a definition of the subset of data to be used. The terms used to describe the initial definition of the acquisition and usage activities are intentionally similar, as both aim at selecting a subset of something. But it is important to distinguish between the acquisition activities, which define a view of the real world and the usage activities, which define a user's view of a data collection;
2. **Retrieve** – As a rule, data are collected and stored for later use. Retrieval, understood as fetching data previously stored, is a necessary step toward that end;
3. **Manipulate** – Retrieved data serve as input to processing, such as classification, analysis, manipulation and synthesis. This step may be bypassed, as data are often retrieved for the sole purpose of being presented to the user;
4. **Present results** – In this activity, results of the retrieval and manipulation are passed on to the user. It is the form of the results, rather than their content, that is of interest here. The appropriate presentation form should be determined by a number of factors, including the nature of the results (numeric, text, graphics etc.), the medium (paper, computer screen etc.), and the user's preference

(people, applications software and optical readers all prefer different forms, for example);

5. **Use** – Here the result are finally used. The users of data are the final judge of the quality of the data used.

From the taxonomic information chain, we can describe and distinguish the actors involved in the processes as follows:

Data Producer – The data producers are those whose role is to collect data from the "real world". In the taxonomic context, we consider the real world to include any set of specimens, "in-situ" or "ex-situ" and the associated data (environmental, spatial, ecological, etc.). On the data producer depends the production of the data sources, the raw material of the taxonomic information. What distinguishes a data producer from a knowledge worker is that a data producer collects data from the real world and their records are (or should be) the ultimate evidence of biological facts and their relationship with the environment. Data producer is not an exclusive role. In most practical cases, taxonomists act as both data producers and knowledge workers. For example, a taxonomist who collects a specimen to produce a herbarium voucher is playing a data producer role. This same taxonomist, studying a set of herbarium vouchers to produce a taxonomic revision, is acting as a knowledge worker. Their role in the data quality context is crucial. They are the interface between the real world and its conceptual, logical and physical representation – the database.

Knowledge Worker – The knowledge workers are those that produce information. They access and compile primary data sources and other references (such as databases)

and information products in order to produce information material. They are, generally, technical users of data and information (taxonomists, other scientists, researchers, teachers, etc.) that require information to do their jobs. Knowledge workers have an important role in data quality issues because they are the judges of the inherent information quality. In a practical taxonomic database context, they are responsible for updating records[3]. Once again, it is not an exclusive role.

Data Intermediaries – Data intermediaries are those whose role is to transcribe data from one form into another. Data entry clerks, computer analysts, software developers, webmasters, desktop publisher designers and all sorts of workers that are involved in storage, management and presentation of taxonomic information are data intermediaries. The difference between data intermediaries and knowledge workers is that the result of the data intermediary's job is not the information itself but its framework.

End User – Knowledge acquisition is the principal aim of the end user and this knowledge is used to promote actions and changing behaviour. End users don't require taxonomic information to do their jobs. Although they are the judges of the pragmatic information quality, they can't assess the inherent information quality aspects. The general public are the most common end users.

4.2. Data quality dimensions

Data quality, as presented in the literature, is a multidimensional concept (Wand and Wang, 1996). In the data quality research area, a number of data quality dimensions

[3] See definition of "update records" in "acquisition activities"

have been identified, although there is a lack of consensus, both on what constitutes a set of a "good" data quality dimensions, and on what an appropriate definition is for each. In fact, even a relatively obvious dimension such as *accuracy*, does not have a well established definition (Wang *et al*., 1995b).

One of the most comprehensive definitions of data quality dimensions is expressed as a collection of 15 dimensions, arranged in 4 categories (Strong *et al*., 1997).

DQ Category	DQ Dimensions
Intrinsic	Accuracy, Objectivity, Believability, Reputation
Accessibility	Accessibility, Access security
Contextual	Relevancy, Value-added, Timeliness, Completeness, Amount of Data
Representational	Interpretability, Ease of understanding, Concise representation, Consistent representation

Table 1 - Data Quality categories and dimensions (Strong *et al.*, 1997)

Because all these definitions and classifications of data quality issues were formulated from the business management and enterprise point of view, the first principle of this work is that to understand and tackle data quality issues in taxonomic databases, we must define specific data quality dimensions, in regard to the particular characteristics of taxonomic data. These specific quality dimensions, we designate *Taxonomic Data Quality Dimensions* (TDQD), which will be developed in the next chapter.

5. Taxonomic data quality dimensions

The Taxonomic Data Quality Dimensions will be formulated by applying and adapting existing data quality dimension definitions to *Taxonomic Data Domains*, previously defined and described in the third chapter of this thesis (Figure 4).

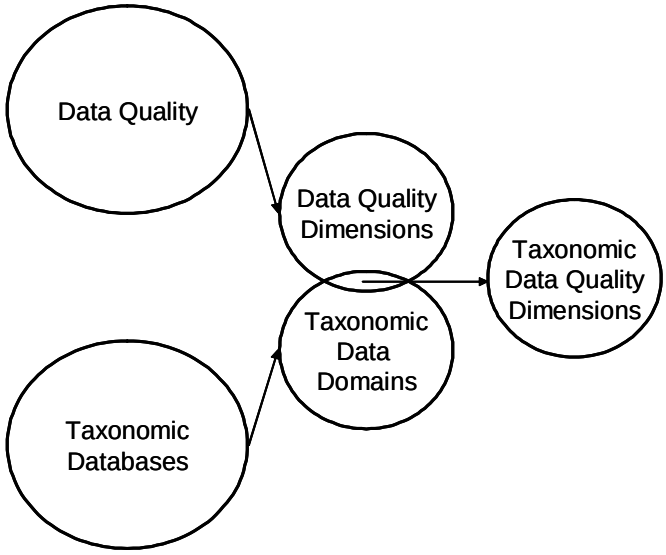


Figure 4 – Definition of the process of formulating Taxonomic Data Quality Dimensions

5.1. Accuracy

The accuracy dimension, also known as *correctness*, *soundness*, *validity* and *freedom-from-error*, or even *precision*, is regarded in most data quality studies as a key dimension (Wand and Wang, 1996). For these different names, *accuracy* is defined as "the extent to which data is correct and reliable" (Pipino *et al*., 2002) and "whether the data available are the true values" (Motro and Rakov, 1998).

Veregin (1998) defined accuracy as the inverse of error, and *error* as a discrepancy between the encoded and actual (real world) value of a particular attribute for a given

entity. For Redman (1996), accuracy of a *datum* <*e,a,v*> refers to the nearness of the value *v* to some value *v'* in the attribute domain, which is considered as the correct values (or one of the correct values) for the entity *e* and the attribute *a*. In some cases, *v'* is referred to as the standard. If the datum's value *v* coincides with the correct value *v'*, the datum is said to be correct.

Regardless of how it is defined, as the degree of true representation of the real world, accuracy can have different impacts over the different taxonomic data domains and, in the taxonomic information chain, taxonomic data accuracy relies mostly on the acquisition activities.

The field data domain is represented by data elements which are representations of a biological fact, recorded based on observations and measurements of specimens. The accuracy of the field data domain relies on the accuracy which the *data producer* achieves in his observation, interpretation, measurement and recording of the "real world", in this case the specimen and its environment.

In the spatial data sub-domain, especially the data elements which represent the collection site, the accurate representation of the collection site has a significant impact on the composition of information about, for example, taxon distribution patterns. However, the quality of the spatial sub-domain can be affected by another dimension, strongly related with accuracy: *precision*, whose impact on quality is discussed below.

Nowadays, there is a consensus between data producers, when engaged in *acquisition activities* in the "in-situ" real world, to rely on GPS equipment for accuracy and

precision of the collection site data and, consequently, geographic coordinates. However, for historical data in herbariums and museums, the same level of precision and, sometimes, the accuracy of the spatial data domain are simply unachievable.

Redman (1996) defines precision in a context called *level of detail* which refers to the quantity of data to be included and how precise those data must be. Two issues are involved in the *level of detail* dimension: *granularity of attributes* and *precision of domains* [4].

When analysing data from the specimen descriptive data sub-domain, accuracy is a key quality domain for achieving the quality of the output in the species descriptive data domain and some quality dimensions associated with the classification data domain. Once again, the degree of *precision* of the specimen descriptive data elements can have a significant impact on *data usage activities*, especially those carried out by the *knowledge workers*.

Although the data elements of the field data domain are "factual" and their *accuracy* represents the conformity of the recorded data with the "real world", this reality may be unverifiable (e.g. historical data), impractical to observe (e.g. too costly) and/or perceived rather than real (e.g. subjective evaluation such as colour, abundance, etc.).

As species descriptive data elements always represent generalizations based on specimen descriptive data elements, the accuracy of species descriptive data elements relies on three aspects:

[4] "*Domains"* are used here in a database context, which means the set of possible values for a given attribute.

- The numbers of specimens observed
- The accuracy of the specimen descriptive data for each specimen observed
- The "accuracy" of identification of specimens

The first aspect cited above may involve other quality dimensions such as *accessibility* (no access to all available specimens) and *currency* (the data about all the specimens was not available at the desired time).

Even if the classification and the descriptive data about all the observed specimens are perfectly accurate, it is virtually impossible for a taxonomist to access "all" the specimens of a given taxon. Thus, as pointed out by Morse (1974), "taxonomic information (species descriptive data domain) is inherently of lower degree of reliability than specimen observations (specimen descriptive data domain)".

In the nomenclatural data domain, *accuracy* can be expressed by the conformity of the name elements with the rules and standards that define its domain (the nomenclatural "real world").

In addition, an accurate nomenclatural data domain can be expressed in a reverse form, as an accurate label for a given biological entity (species or specimen) which makes it possible to associate, select and access accurate information for that biological entity from different data sources, especially for the species descriptive data domain.

In the context of the field data domain, the taxonomic accuracy data quality dimension represents the degree of conformity between the data and the biological fact, for a given specimen.

Associated with the classification data domain, Stribling *et al*. (2003), in a paper discussing data quality issues to be considered when conducting taxonomic analyses for biological assessments, consider 3 potential relationships between taxonomic accuracy and precision. The 1st, and the goal of taxonomic work, is that the data are both accurate and precise. The 2nd is that data are precise but not accurate. This relationship represents a bias that might have resulted from misinterpretation of a dichotomous key or morphological structure, or use of invalid nomenclature. The 3rd potential relationship represents an undesirable scenario where data are both inaccurate and imprecise.

Figure 5 – Relationship between precision and accuracy for the classification data domain

In Figure 5 above, taken from Stribling *et al*. (2003), "T" represents the "analytical truth" and, for taxonomy, the analytical truth is: 1) the most currently accepted taxonomic literature, 2) a reference collection, preferably verified by appropriate taxonomic specialists, or 3) type material (e.g., holotype).

5.2. Believability

Believability is the extent to which data are regarded as true and credible. Among other factors, it may reflect an individual's assessment of the credibility of the data source, comparison to a commonly accepted standard, and previous experience (Pipino *et al*., 2002). Believability has a strong relation with another cited dimension: *reputation* – the extent to which data is highly regarded in terms of its source or content (Pipino *et al*., 2002).

As pointed out by Wang *et al*. (1995a) in their hierarchic representation of data quality dimensions (Figure 6), *believability* has a direct relationship with other dimensions, such as *accuracy*, *consistency*, and *completeness*. However, in a taxonomic "world", *believability* and another close related dimension, *reputation* can have specific meanings and impact on data quality aspects.

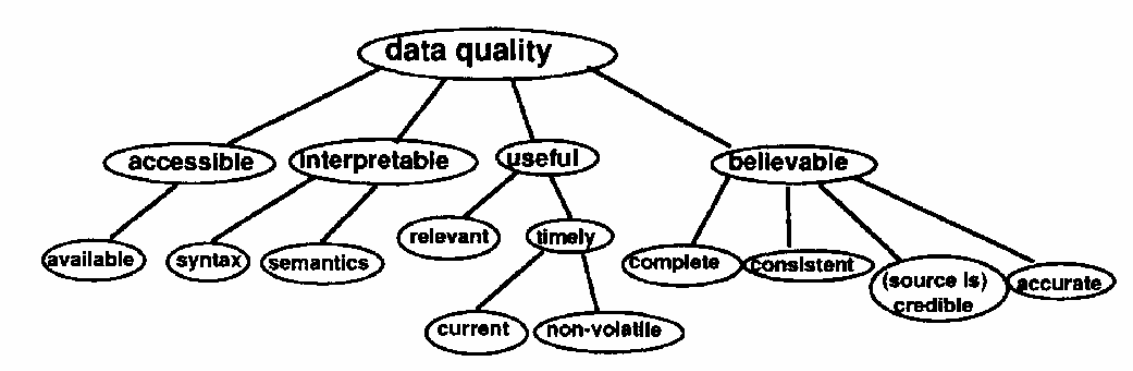


Figure 6 – Hierarchy of data quality dimensions (Wang *et al.*, 1995a)

In a taxonomic context, *believability* may have two different meanings: the believability of data sources and the believability (reputation) of knowledge workers.

Taxonomic databases are created from primary data source records, other databases and other references (for example, monographic revisions), named as *products* in the

taxonomic information chain (see Figure 2). Obviously, the believability or credibility of this database relies on the believability of these primary data sources, and the subjective reputation of the knowledge workers as authors of the products.

The believability of the subjective judgment of a knowledge worker in, for example, compiling different data sources to produce a monographic revision is related to pragmatic information quality and is out of the scope of this Thesis. However, as specimens are the basic operational taxonomic units, the definition of a taxon should ideally include reference to all specimens used to form its concept and thus allow for re-examination of the taxonomist's conclusions (Berendsohn, 1995).

Thus, related to the data content, the focus of this Thesis, the *believability quality dimension* of a taxonomic database can be related to the capability to trace back to the original source of an attribute value, ultimately the specimen observation or voucher.

5.3. Completeness

The completeness dimension can be viewed from many perspectives. At the most abstract level, one can define the concept of the schema completeness, which is the degree to which entities and attributes are not missing from the schema. At the data level, one can define column completeness as a function of the missing values in a column of a table (Pipino *et al*., 2002). For Wand and Wang (1996), completeness is the ability of an information system to represent every meaningful state of the represented real world system.

Veregin (1998) defines completeness as “a lack of errors of omission in a database” and describes two kinds of completeness: *data completeness*, as a measurable error of omission observed between the database and the specification; and *model completeness*, as the agreement between the database specification and the “abstract universe” that is that part of the real world for which data are required for a particular database application.

Motro and Rakov’s (1998) definition for completeness is “whether all the data are available”, regarding the database terminology where data completeness refers to both the completeness of files (no records are missing), and to the completeness of records (all fields are known for each record).

The *completeness* quality dimension in a taxonomic database context has impact and peculiarities in all taxonomic data domains.

In the nomenclatural data domain, the record completeness represents the presence of all name elements needed to compose the scientific name as expressed by the nomenclatural rules.

In the same data domain, the file completeness is related to the context of the database. A regional checklist should have all scientific names related to that specific area. In the same way, a taxonomic database on a specific taxonomic group should store all the scientific names related to that specific taxonomic group.

The most common issue concerning *completeness* in the classification data domain, besides the record completeness of the classification attributes, is related to synonyms. For example, a "good quality" *nomenclatural database* [5] should have all other possible scientific names associated with the "accepted" name. Or, in a *descriptive database* [6], the user should be able to access all the information related to a specific taxon, based on any given name which could be relevant.

In the species descriptive data domain, *completeness* issues are always present. According to the data model and especially the level of granularity chosen for each descriptive attribute, achieving a reasonable degree of *completeness* will very often be a problem because, for species attributes such as vernacular names and uses, they may not exist or the information may not have been compiled.

As far as an individual datum is concerned, only two situations are possible: either a value is assigned to the attribute in question or not. In the latter null case, a special element of an attribute's domain can be assigned as the attribute's value. Depending on whether the attribute is mandatory, optional, or inapplicable, null can mean different things. If the attribute is mandatory, a non null value is expected. The null value is interpreted as "value unknown", the classical interpretation of database theory, and the datum is incomplete (Redman, 1996).

Next, consider an optional attribute such as "shape of leaf" of a species. Here, the null value can mean three different things: The species has leaves but the shape is unknown, the species has no leaves, or we don't know if the species has leaves at all. In the first

[5] Database of scientific names and their status and relationship
[6] Database of descriptive characteristics of species

case, the attribute is applicable to the entity, but its value is unknown. The datum is incomplete. In the second case, the attribute is not applicable due to *character dependence*[7] and the null value is the correct value. The datum is complete. Finally, in the third case, we have no knowledge about the applicability of the attribute and hence of its completeness.

As *completeness* represents, in short, all the possible representations of the real world, we can define *completeness* in the following different forms, related to the different taxonomic data domains:

- Nomenclatural Completeness represents, at the record level, the presence of all name elements necessary to label a taxon. At the file level, nomenclatural completeness represents all possible names, given a context, generally a taxonomic context (a list of names for a specific taxonomic group) or a spatial context (a list of names for a specific region).

- Classification Completeness represents, at the file level, all possible names related to an "accepted" name for a given taxon.

- Species Descriptive Completeness represents, at the record level, all available descriptive information on a given taxon, according to the defined attributes.

For example, a hypothetical "complete" descriptive database of the Legumes of Brazil should have:

[7] The concept of *character dependency* has been explained in Pankhurst (1993).

- All scientific names of species of legumes which occur in Brazil (nomenclatural file completeness).
- Each scientific name should have all the essential name elements, necessary to distinguish one name from another (nomenclatural record completeness).
- Each scientific name should be associated with all other possibly related names (classification file completeness).
- Each scientific name should have complete descriptive information (descriptive file completeness).
- Every field of descriptive information for each scientific name should have data (descriptive record completeness).

In a field data domain, the completeness data quality domain can be applied at the record level, which means that all fields for a given specimen should contain data. At the file level, the completeness of a field data domain can be represented as a sufficient amount of data to describe the biological facts.

5.4. Consistency

In the literature, consistency refers to several very diverse aspects of data. In particular, to values of data, to the representation of data, and to physical representation of data (Wand and Wang, 1996). Veregin (1998) defines consistency as “the absence of apparent contradictions in a database”.

Redman (1996) recognizes two aspects of *consistency*: *semantic consistency*, that the expression and interpretation of data should be clear, unambiguous and consistent; and

structural consistency, by which entity types and attributes should have the same basic structure wherever possible. Pipino *et al*. (2002) define *consistent representation* as the extent to which data is present in the same format.

Inconsistency can be related to the classification data domain if two or more names are considered as “accepted” to represent the same taxon at the same time. However, in a concept-oriented database where several taxonomic views may be stored simultaneously, such a situation would not be regarded as an inconsistency.

Inconsistency issues can arise with any two related data items. For example, if a species descriptor is given as “habit = herbaceous” and “uses = wood”, the data is inconsistent. This kind of inconsistency is very common in the spatial data sub-domain as, for example, latitude and longitude coordinates that point to places that do not represent the same place described textually.

Inconsistent representation of the same attribute is also a *consistency* problem and will be present also in taxonomic databases with poor attribute domain definitions and lack of data standards. For example, “*habit = shrub*” and “*habit = bush*”.

Inconsistent data values will often arise when two or more collections of data overlap in their entities and attributes. This is particularly true when compiling classification data domains and species descriptor data domains from different data sources, with different authors. For instance, in the classification data domain, two different databases could have the same scientific name as accepted, in the first one, and synonym, in the second one. This kind of inconsistency, treated as “conflicts”, is the main subject of projects

such as LITCHI (Embury *et al*., 2001; Embury *et al*., 1999; Jones *et al*., 2001). For descriptor data domain, one database can show that one species is native while the other database will assert that it is an introduced species in the same region.

In a nomenclatural data domain, inconsistent representation of a scientific name is a very common problem, often related to spelling errors. For example, “Vicia faba” and “Vycia faba”. Moreover, consistency in the nomenclatural data domain means that a scientific name has just one correct representation (no redundant representation by two or more different scientific names with different spelling forms, unless it is the intention of the database to record this, as in some nomenclatural databases).

Consistency in the classification data domain means that, for a given classification, every scientific name has a classification status, with just one “accepted” name for each taxon, and all the “non-accepted” names should be related to at least one accepted name. In addition, the classification itself needs to be consistent, e.g. two species in the same genus can’t be in different families.

Consistency in the field data domain and species descriptive data domain refers to a consistent representation of the descriptive attributes and the absence of conflict between related attributes, e.g. “*locality = Rio de Janeiro*” with “*country = England*”; “*habit = herb*” with “*use = wood*”.

To summarize, a simple set of rules for defining the consistency of a record in a taxonomic database can be defined as:

Consistency exists when a taxonomic database record satisfies the following rules:

- all the nomenclatural data elements include just one valid scientific name
- all the classification data elements represent just one specific taxon
- all the field data elements represent just one single observation or collection event
- all the spatial data elements represent just one valid location
- all curatorial data elements represent just one valid specimen
- all specimen and species descriptive data are:
    - not in conflict by character dependence
    - not in conflict by their descriptive characteristics (morphological, ecological, physiological etc.)
    - standardized in their meanings

5.5. Flexibility

Levitin and Redman (1995) introduce *flexibility* as a quality dimension under a conceptual view of *reaction to change*. *Flexibility* they understand as the capacity for a view to change in order to accommodate new demands.

Although flexibility is not a notable data quality dimension (Wand and Wang, 1996), it is a notable and desirable characteristic of taxonomic databases. As pointed out by Pullan *et al*. (2000), the same organism may at times be classified according to different taxonomic opinions and subsequently have several alternative names. However, almost all present taxonomic database implementations are designed to handle only a single

taxonomic view. The usual approach to handling taxonomic data has been to use names as identifiers of taxon concepts, with statements regarding the taxonomic status of a taxon assigned to the name. This unrealistically forces the adoption of a single consensus classification (Pullan *et al*., 2000).

A few efforts have been made to overcome these limitations, based on more appropriate data models (Berendsohn, 1995; 1997; Berendsohn *et al*., 1999; Pullan *et al*., 2000; Raguenaud *et al*., 1999; Zhong *et al*., 1996) and software implementations (Raguenaud *et al*., 2002).

Despite the focus of this thesis on inherent information quality – based on data values – and for the reasons discussed above, we consider it appropriate to include a flexibility data quality dimension, in the framework proposed in this thesis, as the capability of a taxonomic database, in a model or implementation level, to handle different classifications of the same group of specimens, accommodate new attributes and new data types, for example, images, sounds etc., and add new values to existing descriptors.

5.6. Relevance

Relevance is defined in data quality research papers as “the extent to which data is applicable and helpful for the task at hand” (Pipino *et al*., 2002), as “the degree to which the provided information satisfies user needs” (Naumann, 2001), and as the extent to which “the view should provide data needed by the application” (Redman, 1996).

One practical example of relevance can be found at the database website of the Northeast Brazil Plant Checklist of the Plant Information Centre of Northeast Brazil – CNIP (www.cnip.org.br). In its first version, the database was published with the main page as an index of the botanical families (Figure 7). Despite the great relevance for taxonomists, a group of important *end users*, such as agricultural technicians and small farmers' associations couldn't see any relevance at all, until the second version was published with the main page indexed by vernacular name (Figure 8).

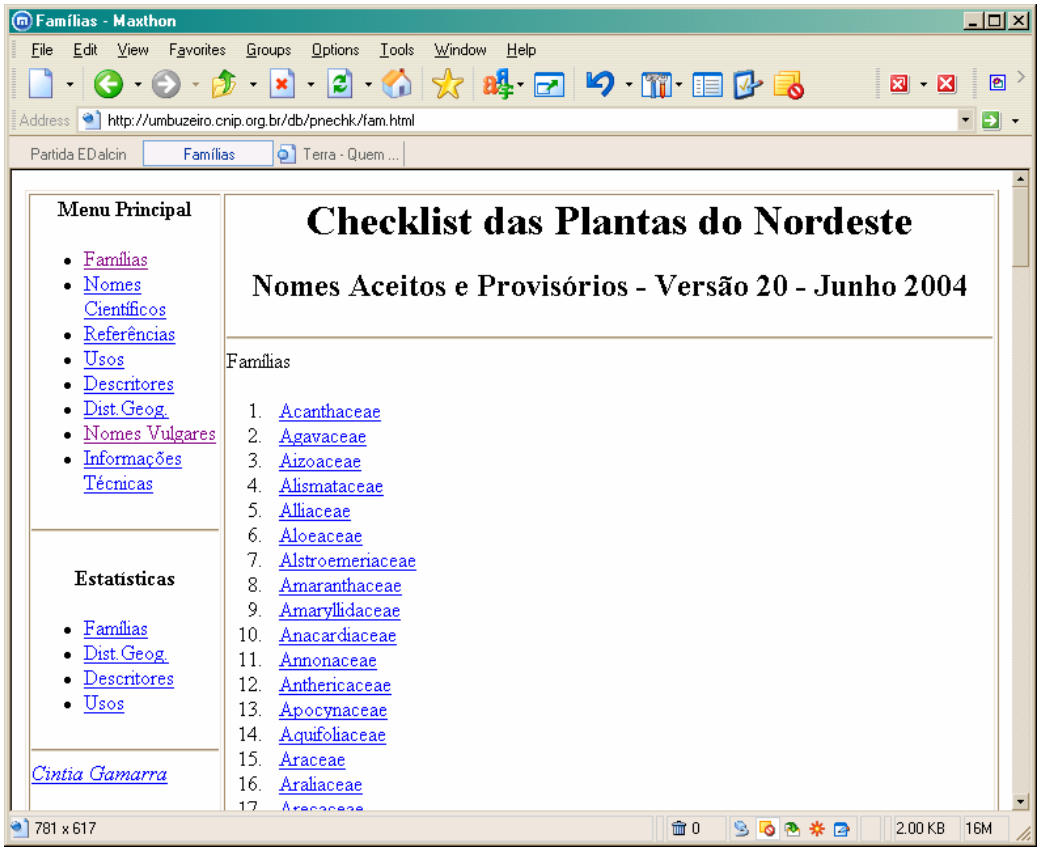


Figure 7 – Database of Northeast Brazil Checklist indexed by family

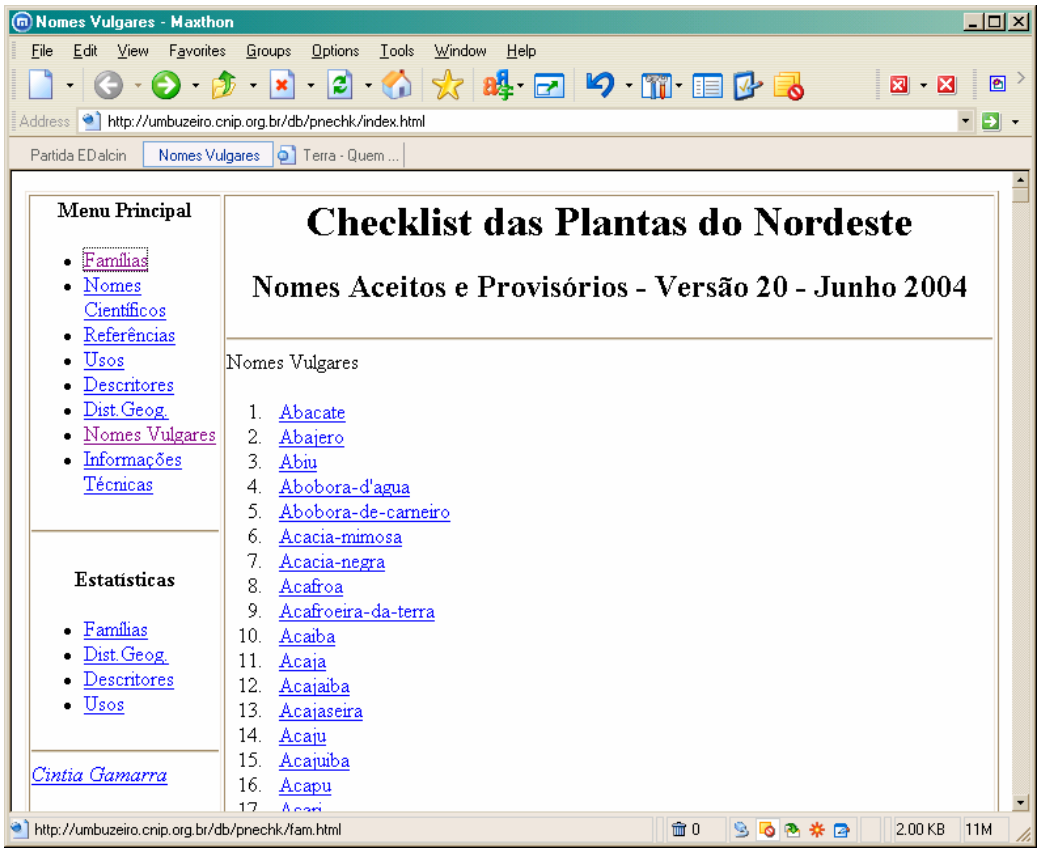


Figure 8 – Database of Northeast Brazil Checklist indexed by vernacular name

Relevance is related, without doubt, to pragmatic information quality and can have a significant impact on what is called by Bisby (1984b) “The Taxonomic Information System”. In his papers, Bisby (1984b; 1988) has pointed out the absence of studies and surveys about who are “the users” of taxonomic information and the wide range of “tasks” and communication flows involving taxonomic information.

Although the specificity of the relevance concept is based on its direct dependence with “the task in hand” and “the user needs”, the relevance of taxonomic information can often be related to the classification data domain and descriptive data domain.

Generally, the classification data domain seems to be highly relevant to taxonomists, a very specific and well delimited group of knowledge users, while general descriptive

information, called *the consumer's target data* by Bisby (1984b) seems to be more relevant to end users.

To achieve a high relevance data quality or, in other words, to tackle issues related to the relevancy data quality domain, taxonomic databases should be built on a clear understanding of the data consumers, their needs and their tasks. In other words, they should be *consumer-focussed* databases.

5.7. Timeliness

Timeliness reflects how up-to-date the data is with respect to the task it is used for (Pipino *et al*., 2002).

Timeliness has been defined in terms of whether the data is out of date and availability of output on time. A closely related concept is *currency*, which is interpreted as the time a data item has been stored (Wand and Wang, 1996).

Timeliness is affected by three factors: How fast the information system state is updated after the real world system changes (system currency); the rate of change of the real world system (volatility); and the time the data are actually used (Wand and Wang, 1996).

Considering taxonomic data domains, the classification data domain is the main data domain which is affected by "volatility" and, consequently, is the main data domain affected by timeliness issues.

The spatial data domain, as evidence of the real world at a specific time, is not volatile at all. However, considering a herbarium voucher, the scientific name given to that botanical sample is indeed volatile, for example when its identification changes. This volatility will be reflected at the field level, and at the nomenclatural data domain, but it is important to notice that once published and correctly represented (the accuracy dimension), one scientific name is volatile only in its classification aspect.

The curatorial data domain is volatile as well but only in the data elements which are related to the collection management data (such as loans and voucher storage data, for example) and identification events, as a result of the volatility of classification.

The species descriptive data domain is volatile in the sense that a new set of raw data (specimens) could result in adjustment or addition to the generalized data which make up the species information. This new set of raw data could arise from new collections or specimen observations or changes in classification of existing specimens.

The concepts discussed in this chapter and previous chapters can be summarized in the following table (Table 2).

		Taxonomic Data Domains					
	Domain	Nomenclatural	Classification	Field			Species Descriptive
	Sub-domain			Spatial	Curatorial	Specimen Descriptive	
Taxonomic Data Quality Dimensions	Accuracy	X	X	X		X	X
	Believability		X				
	Completeness	X	X	X	X	X	X
	Consistency	X	X	X		X	X
	Flexibility		X				X
	Relevance		X				X
	Timeliness		X		X		X

Table 2 – Summary of Taxonomic Data Quality Dimensions, Taxonomic Data Domains and their relationships explored as examples of inherent data quality in this thesis

6. Improving data accuracy in taxonomic databases

The quality of a real-world data set depends on a number of issues (English, 1999; Redman, 1996; Wang *et al*., 1995b), but the source of the data is the crucial factor. Data entry and acquisition are inherently prone to errors, both simple and complex. Much effort can be given to improving this front-end process, with respect to reduction in entry errors, but the fact often remains that errors in large data sets are common (Marcus *et al*., 2001).

Most experts agree that the general principles of quality management as applied to products can also be applied to data. This suggests there should be two basic approaches to the improvement of data quality, namely: defect detection (and correction) and defect prevention (Embury, 2001).

Defect prevention is considered to be far superior to defect detection, since detection is often costly and cannot guarantee to be 100% successful at any stage. It is discussed further in section 9.2. However, defect detection has a particularly important role to play when dealing with legacy applications, for which large volumes of incorrect data already exist (Embury, 2001). In this case, existing data defects can be identified, and eventually removed, via a process known as *data cleansing*.

6.1. Data cleansing

Data cleansing, also called *data cleaning* or *scrubbing*, deals with detecting and removing errors and inconsistencies from data in order to improve its quality (Rahm and Do, 2000).

For English (1999), the process of cleansing is to improve the quality of data within the existing data structures. This means standardizing non-standard data values and domains, filling in missing data, correcting incorrect data, and consolidating duplicate occurrences.

Aspects and techniques of data cleansing have been studied in the context of *Knowledge Discovery in Databases* (KDD), also known as *data mining*, where data cleansing is part of a series of interactive steps leading from raw data collections to some form of new knowledge (Figure 9).

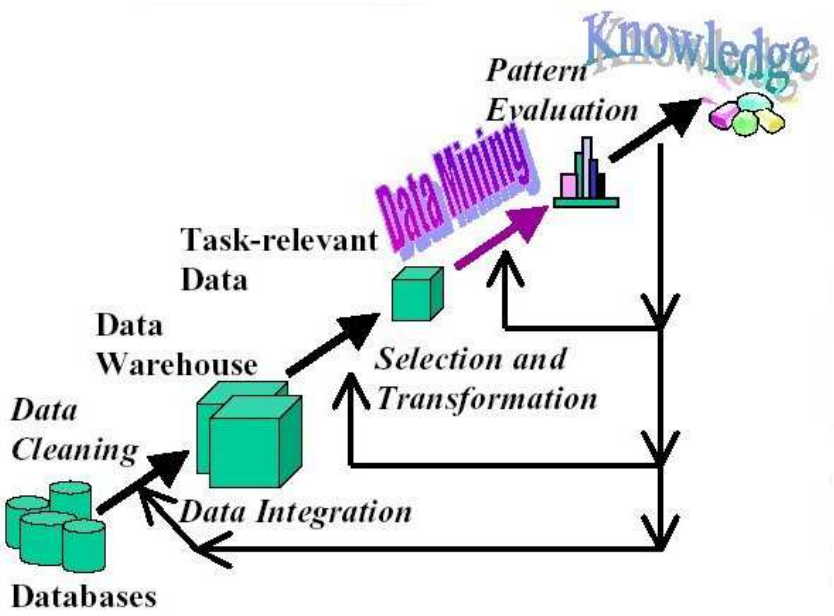


Figure 9 – Steps of the knowledge discovery process (Zaïane, 1999)

The general framework for data cleansing is described by Maletic and Marcus (1999) as follows:

- Define and determine error types
- Search and identify error instances
- Correct the errors uncovered

Thus, it is appropriate to say that the focus of this thesis will be on the task “define and determine error types” (section 6.2) and on the task “search and identify error instances” (section 6.3). Correction techniques and strategies are part of the discussion, in section 9.1.

In the context of taxonomic databases, the process of data cleansing will be related to the taxonomic data domains and what will be called *taxonomic error patterns* in order to explore and emphasise the different aspects and limitations of error detection and correction in each taxonomic data domain.

6.2. Taxonomic error patterns

In order to establish a framework to support the conceptual aspects and practical work and experiments of this Thesis, a set of *Taxonomic Error Patterns* are defined, based on the general data content defect (or error) patterns proposed by English (1999). These generic error patterns will be related to the *taxonomic data domains* where they assume different meanings in different data domains.

6.2.1. Domain value redundancy

> ***"Domain value redundancy*** *– No standardized data values, or synonym values in which two or more values or codes mean the same thing." (English, 1999)*

This kind of error pattern could have different meanings in the taxonomic context. First, it could be exemplified in the species descriptive data domain, when we have two different records for the same species: "*habit = shrub*" and "*habit = bush*", or "*flower colour = carmine*" and "*flower colour = crimson*".

This domain redundancy can be very typical in a descriptive data domain which lacks standards or in a badly controlled process of data compilation from different sources and from specimen descriptive data domains.

The specimen descriptive data domain is, in the same manner, vulnerable to *domain value redundancy* when considering, for example, herbarium databases. However, as the raw data generated is usually obtained from different data producers (specimen collectors), the domain value redundancy in the specimen descriptive data domain is expected and, at a reasonable level, accepted.

Domain value redundancy can easily happen by mistake and may be the most common and deleterious error pattern in taxonomic databases, for example when we have two different spellings of the scientific name for the taxon. However, as we will see in the following data content error pattern definitions, this fits better in the *duplicate occurrence* error pattern.

6.2.2. Missing data values

> ***"Missing data values –*** *Data that should have a value is missing. This includes both required fields and fields not required to be entered at data capture, but are needed in downstream processing." (English, 1999)*

Missing data values could have different meanings and impacts in different taxonomic data domains, but are always related with the *completeness* and *accuracy/precision* Taxonomic Data Quality Dimensions.

Considering a species descriptive database; a missing data value in a descriptive data domain could have one of the following meanings:

1. The value of the given attribute exists but is unknown;
2. The value of the given attribute doesn't exist;
3. The value of the given attribute is inapplicable; and,
4. The value of the given attribute exists, is known and applicable but the data field was missed in the data entry/conversion/migration process.

In the first case, an empty field for an *average height* attribute is a good example. Despite the fact that almost every species, in the real world, must have an *average height*, for that specific entity (species) the data value was not compiled or observed.

In the second case, an empty field for the *vernacular name* attribute is a good example since it is a real world fact that not every species has a vernacular name.

In the third case, we can have a *character dependency*, where an attribute has its value in a dependency (usually morphological) relationship with another attribute. Examples could be *flower colour* for a fern species record or a *spine size* attribute for a species record without spines.

In the fourth case we have an effective error and a missing data value.

Despite the complexity of the descriptive data in taxonomic databases, well documented in the taxonomic literature (Allkin, 1984b; Allkin *et al*., 1992; White *et al*., 1993), the "missing value error pattern" will be considered here as a generic error pattern, illustrated by the examples above.

In the spatial data sub-domain, missing values are very common in historical records in herbarium databases, especially in the *latitude* and *longitude* fields.

In the nomenclatural data domain, a missing value of a name element, usually represented by the absence of a nomenclatural rank, could represent an incomplete classification of the entity (specimen or species) and is a very common "error pattern" in taxonomic databases.

In some cases, the absence of scientific name elements, including authority name elements, could be caused by a bad conversion process, with the objective of atomizing the scientific name string to be fitted in a different data structure. However, the absence of infraspecific name elements does not prevent a record being considered as complete and correct.

The general rule is that the missing value error pattern will have a direct relationship with the data model and the structure of the database or, in other words, the number of "fields" or attributes which were defined to describe the "real world". The greater the number of attributes, the higher the chance of missing values issues.

6.2.3. Incorrect data values

> ***"Incorrect data values –*** *These may be caused by transposition of key-strokes, entering data in the wrong place, misunderstanding of the meaning of the data captured, or forced values due to fields requiring a value not known to the information producer or data intermediary." (English, 1999)*

Incorrect data values are the most obvious and well known error patterns and can affect every data value in every taxonomic data domain, with no exceptions. Incorrect data values are usually associated with poor data gathering or interpretation, data migration, conversion, transmission or input issues.

Incorrect data values can be of two types:

- The value is not valid for the given attribute

  This type of incorrect data value can be typically exemplified by values outside the possible range for dates, morphometric measures or character states and can be easily detected by automated procedures.

- The value is valid but not a correct representation of the given attribute

This type of incorrect data value generally cannot be detected by automated procedures and requires close investigation by *knowledge workers* to be detected and corrected.

6.2.4. Nonatomic data values

> ***"Nonatomic data values –*** *Data fields may be misdefined as nonatomic or multiple facts may be entered in a single field." (English, 1999)*

The nonatomic data values problem arises as a result of a poor data structure and imprecise data modelling. This error pattern became very common when knowledge workers (generally taxonomists) had access to "user friendly database applications" and started to create their own database information systems. In fact, the most common situation among taxonomists, in the absence of institutional support and with a lack of specific applications, is to use a spreadsheet or "flat file" to organize information about species or specimens. The result, very often, is a structure like this:

Family	Species	Vernacular Name
MIMOSACEAE	*Anadenanthera colubrina* (Vell.) Brenan var. *cebil* (Griseb.) Altschul	angico, angico-jacare
ANACARDIACEAE	*Schinopsis brasiliensis* Engl.	barauna

Table 3 – Nonatomic data values examples

In the table above, we consider the multiple values for vernacular name as an example of the same error pattern of the typical nonatomic values, shown in the species field.

In a database management context, this kind of structure represents a real "nightmare" for the database administration staff and has a significant negative impact on efforts to improve data quality and processes such as database merging.

However, nonatomic data values will be present in better (but still not sufficiently good) data structures and are very common in situations like these:

Family	Genus	Species	Subspecies	Author
LEGUMINOSAE	Swartzia	simplex	var. grandiflora	(Raddi) Cowan

Table 4 – Nonatomic data values example

This situation reflects the absence of a specific field to accommodate the *rank* of the infra-specific name element or epithet and thus will result in a nonatomic data value problem.

Descriptive data domains will often be vulnerable to nonatomic data values when faced with recording complex features, such as morphological structures or colour patterns.

6.2.5. Domain schizophrenia

> ***"Domain schizophrenia*** *– Fields may be used for different purposes depending on a specific requirement or purpose." (English, 1999)*

Caused mostly by a poor data model, one of the most frequent expressions of *domain schizophrenia* in taxonomic databases is related to the nomenclatural data domain. Incomplete or temporary attempts to classify a taxon by a taxonomist could result in

textual scientific name representations which don't satisfy nomenclatural rules and will be reflected in the database content.

Considering the following examples:

Family	Genus	Species
LAURACEAE	Ocotea	?
LAURACEAE	Ocotea	puberula?
LAURACEAE	Ocotea	= Ocotea puberula
LAURACEAE	undetermined	
LAURACEAE	Ocotea	To be verified
LAURACEAE	Ocotea	sp.nv.
LAURACEAE	Ocotea	sp.1

Table 5 – Domain schizophrenia examples

The examples above can be very common in taxonomic databases and illustrate the utilization of fields for purposes that they were not designed for. The reasons for these representations could be doubt on the part of the taxonomist, an unfinished identification process, lack of expertise to identify the taxon, or could represent an acceptable level of precision in ecological studies.

Thus, because in the taxonomic context the *domain schizophrenia* error pattern can be related with the *embedded meaning data value* error pattern proposed by English (1999), we are grouping these two concepts. As an example of an *embedded meaning* we can take the first record in Table 5, where the question marks "?" has the *embedded meaning* of an incomplete classification. Another example of *embedded meaning* can be seen in Figure 10, in the Northeast of Brazil Repatriation Database, maintained at the Royal Botanic Gardens, Kew, and available on the Internet (http://www.rbgkew.org.uk/brazilrepat/SearchApp), where a question mark "?" is present together with the scientific name.

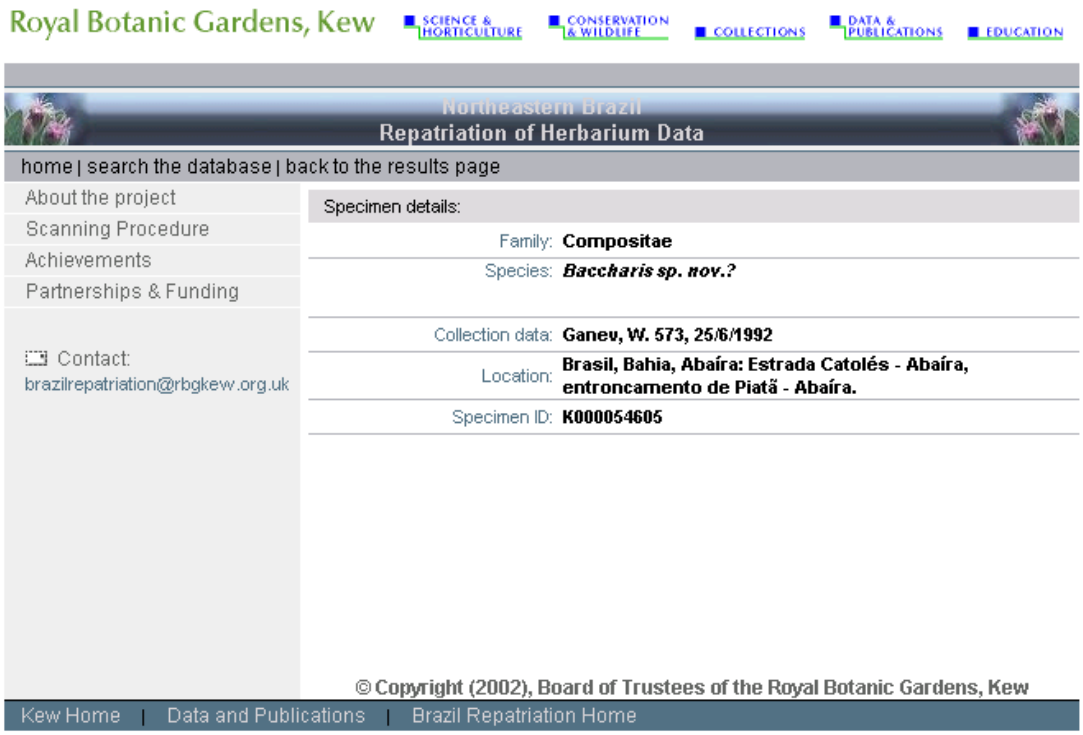


Figure 10 – *Embedded meaning* example in Kew's North-eastern Brazil Repatriation Database

6.2.6. Duplicate occurrences

> ***"Duplicate occurrences –*** *Multiple records that represent one single real world entity."*
>
> *(English, 1999)*

The seven example records given for the *domain schizophrenia* error pattern above could be regarded as examples of the duplicate occurrence of records which are intended to represent a single taxon. However, because of the lack of a complete or correct representation of the name elements, and the *embedded meaning* of doubt and unfinished work, we are not considering it as a typical case of duplicate occurrence.

One of the most typical expressions of *duplicate occurrence* in taxonomic databases occurs in the nomenclatural data domain, as a result of a spelling error in the scientific name. As a core entity in a taxonomic database, a "whole new taxon", represented by its scientific name, can be easily created from a genuine taxon by addition (omission, transposition or substitution) of one single letter.

Consider the following example, a fragment of a database of marine species from New Zealand:

Genus	Species
Kolestoneura	novaezelandiae
Kolostoneura	novaezelandiae
Kolostoneura	novaezealandiae
Harpecia	spinosissima
Harpecia	spinossissima
Harpecia	spinossisima
Astraea	heliotrophum
Astraea	heliotropium
Astraea	heliotropum
Astraea	heliotrophon
Astraea	heliotrophum

Table 6 – Duplicate occurrences example

As a database created without any kind of data entry control, we can assume with a high degree of confidence that the eleven scientific names presented are intended to represent only three different taxa. However, in databases with the intention of documenting the existence of valid (published) misspelled names, e.g. a nomenclatural database, then these names might not be considered as duplicate occurrences at all.

The misspelling error in scientific names, considered here as the main data quality problem in taxonomic databases, related to data content and the inherent data quality context, will be treated separately in the following chapter.

6.2.7. Inconsistent data values

> ***"Inconsistent data values –*** *Unmanaged data stored in redundant databases often gets updated inconsistently; that is, data may be updated in one database, but not the others. Or, it may even be updated inconsistently in the different databases." (English, 1999)*

Following the English (1999) definition, inconsistent data values refers to different and redundant databases. This situation could arise in an institutional context, when different and unrelated databases have different data management procedures and are part of different taxonomic information chains.

For example, consider an institution which has a herbarium database and a different specimen oriented database to handle accession and management of a live collection. The lack of synchronization between both nomenclatural "core" tables will become obvious if the herbarium databases have a significant update, based on a recent monographic revision for a specific taxonomic group.

As shown in the examples given in the consistency data quality dimension, inconsistencies could be found in data values which belong to related descriptive attributes. However, they will be treated, following the English (1999) definition, as examples of the incorrect data values error pattern.

6.2.8. Information quality contamination

> ***"Information quality contamination –*** *The result of deriving inaccurate data by combining accurate data with inaccurate data." (English, 1999)*

In a compilation and generalization of specimen descriptive data, the species descriptive data domain can be affected by information quality contamination. Information quality contamination could be, in fact, the cause of other error patterns such as the contamination of a species checklist by misspelling a scientific name on a herbarium voucher.

Information quality contamination is very often present on summarized reports or tables when, for example, just one defect value in the *height* field will contaminate the *average size* result.

6.3. Taxonomic error detection approaches and techniques

In addition to these generic data content error patterns, proposed by English (1999) and related here to taxonomic database issues, I propose three specific taxonomic error patterns that will be explored in the context of error detection.

The focus in these three error patterns on definition and error detection techniques are based on previous experience of the author, and they are considered as responsible for a significant impact on data quality in taxonomic databases.

Nomenclatural Structural and Scientific Name Spelling errors have major importance in taxonomic databases because they affect the principal identifier of the taxonomic information. Database domain errors will be presented as a conceptual framework which can be used to detect inconsistencies between data values and the database's view of the real world.

6.3.1. Nomenclatural structural errors

As pointed out in Chapter Two, nomenclatural data domains, as far as plant names are concerned, are governed by articles of the International Code of Botanical Nomenclature (Greuter *et al*., 2000). Based on these rules, and taking account of the changes in successive editions and the occasional exceptions caused by the failure of the rules, we can establish a set of procedures in order to check the accuracy of a scientific name in a taxonomic database.

For illustrative purposes, we consider the following name elements in a taxonomic database:

LEGUMINOSAE	*Vicia*	*faba*	L.
Family [8]	*Genus name*	*Species epithet*	*Species author string*
F	*G*	*S*	*SA*

Table 7 – Name elements in scientific names

[8] Although the concept of *full name* adopted as standard (Bisby, 1995), does not include the *family* name element, we consider this name element as generally present in most taxonomic databases, thus with a potential impact on taxonomic data quality over the database as a whole.

Levels of validity

The scientific name can be checked following three levels of validity:

- Structural level
- Nomenclatural level
- Classification level

These levels represent ordered steps of a validation process designed to achieve the full quality of scientific names in a taxonomic database.

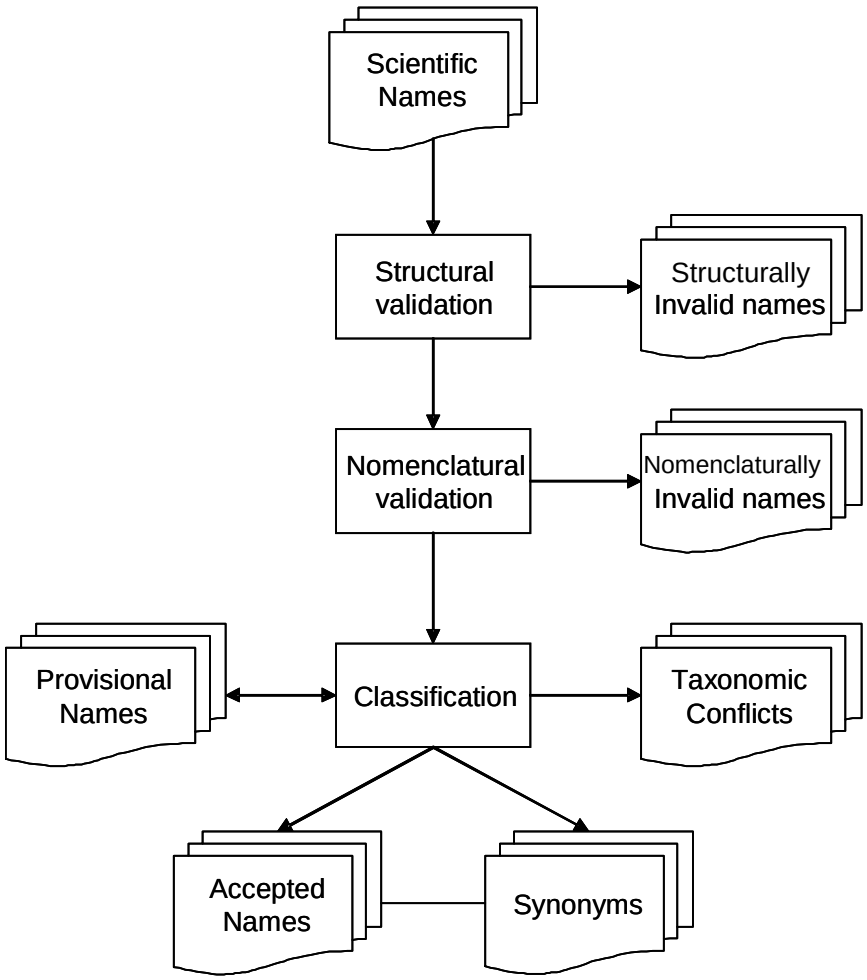


Figure 11 – Validation process for scientific names in taxonomic databases

As a set of specific rules that can be applied to each name element independently, structural validation can be applied without the participation of *knowledge workers* (taxonomists), as an automated process by *data intermediaries* or as a part of data entry checks or interfaces.

The structurally invalid names, detected as the result of the structural validation process, can usually be corrected only by accessing the original data sources or, alternatively, purged.

The nomenclatural validation process aims to authenticate the relationship between the name elements, regarding their hierarchical relationship. Typical examples of nomenclaturally invalid names are a *genus* related to an incorrect *family*, *species* related to the wrong *genus* or names without a valid authority. In essence, nomenclatural validation certifies that a scientific name was validly published and actually exists in the scientific community as a valid name.

The classification process intends to verify the relationship between the scientific names in the database, regarding their *status* as "correct" names, synonyms, misapplied names, orthographic variants and so on. The classification validation process in a taxonomic database context does not intend to evaluate subjective judgements, but to detect taxonomic conflicts between scientific names and their application as taxon labels. Aspects of taxonomic conflicts in taxonomic databases have been studied in the context of the LITCHI project (Sutherland *et al*., 2000) and, as a pragmatic aspect of data quality, the correct use of a scientific name as an identifier of a taxon will not be treated in this thesis.

Structural level

Validity at the structural level can be checked by comparing the scientific name elements with the nomenclatural rules established by the International Code of Botanical Nomenclature (Greuter *et al.*, 2000) or by the International Code of Zoological Nomenclature (ICZN. *et al.*, 1999). For illustrative purposes, the following examples will consider a plant names database.

Following the example given above in Table 7, we define:

Scientific name (SN) =Family (F) + Genus (G) + Specific epithet (S) + Species Authority (SA), in this order.

To establish the validity of a scientific name, the following example rules can be applied:

SN is valid (va) when:

a) $F <> [\varnothing] \wedge G <> [\varnothing] \wedge S <> [\varnothing] \wedge SA <> [\varnothing]$

b) $F = (va) \wedge G=(va) \wedge S=(va) \wedge SA=(va)$

Where [∅] denotes the empty set, i.e. a null string

F = (va) when**:**

a) *F* string elements $\subset$ [A..Z] [9]

b) The last two string elements of *F* = "AE" [10]

[9] If the convention is to store family names in upper-case letters.
[10] International Code of Botanical Nomenclature (Tokyo Code), Art. 18.1 e 18.5

G = (va) when:

a) *G* string elements $\subset$ [A..Z] $\vee$ [a..z] $\vee$ $\supset$ [-] [11] (hyphen)

b) The first *G* string element is upper case [12]

c) All the *G* string elements, except the first one are in lower case [13]

d) The first *G* string element $\wedge$ the last *G* string element $\neq$ [-]

S = (va) when:

a) *S* string elements $\subset$ [a..z] $\vee$ $\supset$ [-] [14] (hyphen)

b) The first *S* string element $\wedge$ the last *S* string element $\neq$ [-]

c) All *S* string elements are in lower case [15]

d) $S \neq G$ [16]

SA = (va) when:

a) *SA* string elements $\subset$ [a..z] $\vee$ [A..Z] [17]

b) *SA* string elements $\supset$ [&] [18] $\vee$ [.] $\vee$ [ ] $\vee$ [‘] $\vee$ [( ] $\vee$ [ )]

c) If [( ] is present in *SA*, then [ )] must be present in *SA*, in a position > [( ] position.

d) [ )] is not the last character of *SA*.

Those examples can be adapted for different levels of *accuracy* demanded by different databases purposes. Although the TDWG standard publication *Plant Names in Botanical Databases* (Bisby, 1995) does not consider *Genus Authority* as mandatory,

[11] International Code of Botanical Nomenclature (Tokyo Code), Art. 20.3
[12] International Code of Botanical Nomenclature (Tokyo Code), Art. 20.1
[13] International Code of Botanical Nomenclature (Tokyo Code), Art. 20.1
[14] International Code of Botanical Nomenclature (Tokyo Code), Art. 23.1
[15] International Code of Botanical Nomenclature (Tokyo Code), Art. 60F.1
[16] International Code of Botanical Nomenclature (Tokyo Code), Art. 23.4
[17] Including specific language characters and accentuation
[18] International Code of Botanical Nomenclature (Tokyo Code), Art. 46C.1

specific monographic databases can define this name element as mandatory, consequently considering invalid any records without this name element.

Nomenclatural level

Nomenclatural validation consists of authenticating a set of name elements as a valid scientific name. Two approaches can be taken to automate detection of nomenclatural errors, both based on a technique called *database bashing*, which involves comparison of data in two or more databases.

The first approach aims to validate the individual name elements as a recognized label for a taxonomic rank. Higher taxonomic ranks, such as family and genus can be checked against authoritative datasets, such as Brummitt (1992), in order to validate the names related to these ranks. In practice, only taxonomic ranks at the genus level or above can be checked against other datasets or dictionaries.

For species names, at species or infra specific level, for any taxonomic group, there are no electronic datasets (or even non electronic ones) which could be considered to have a sufficiently high level of quality or confidence to be used as a reference for the database bashing process. However, the *database bashing* concept admits the comparison between two or more unreliable databases. Records that are equivalent in both databases are considered to be reliable. Records that conflict are deemed unreliable.

With authority data, despite the existence of a data dictionary considered as a standard (Brummitt and Powell, 1992), the complexity of all the possible (and valid)

representations of authorities related to a taxonomic rank[19] makes an automated validation of authority data a very difficult task.

The second approach aims to validate the relationship between the taxonomic ranks in the scientific name. In essence, to verify if the given genus belongs to the appropriate family, the given species belongs to ditto genus, and so on. The validation of authority fields can be checked at this step, by their relationship to a given taxonomic rank.

Again, there are datasets or data dictionaries that could be used in order to compare the existing relationship between the different taxonomic ranks in a scientific name (Brummitt, 1992).

Summarizing, the process of detecting nomenclatural structural errors aims to establish an ordered set of procedures to validate a collection of strings as a valid scientific name. In the process, a series of outputs containing "suspicious data" are generated which should be evaluated by *knowledge workers*, corrected and then re-integrated into the process until the last validation process is successfully achieved.

6.3.2. Database domain exception

Taxonomic databases are intended to represent a subset of the real world. These subsets can be defined as a "domain" of the database. For the taxonomic data quality approach, we assume every taxonomic database will contain at least one of the following domains:

- Taxonomic domain

[19] International Code of Botanical Nomenclature (Tokyo Code), Arts. 46-50

- Spatial domain
- Descriptive domain

We can illustrate these database domains with an example of a hypothetical database of "Woody Leguminosae of Brazil":

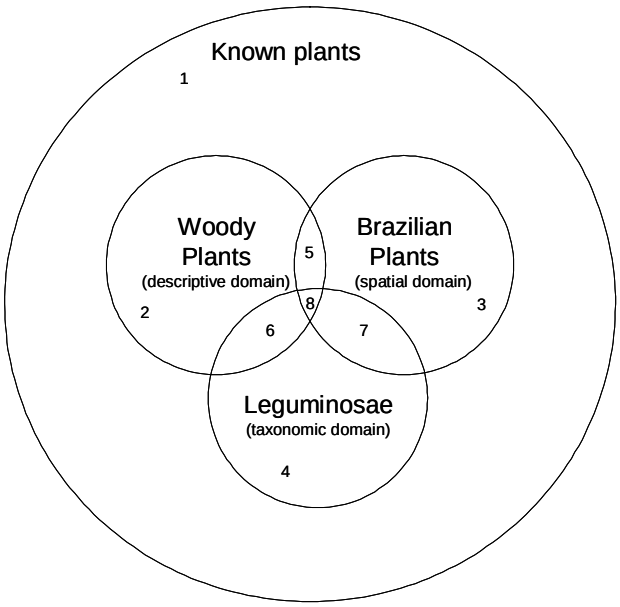


Figure 12 – Database Domain example: Woody Leguminosae of Brazil database

In this particular database, the only valid records are those which are found in region no.8 of the above figure.

Thus, the process for detecting database domain exceptions consists in identifying the records which fall in any other region outside the intersection of all the domains of a given taxonomic database.

Experiment in spatial domain exception detection

A practical example of spatial domain exception detection can be described with a real case as follows:

The Plant Information Centre for the Northeast, in Brazil, received a database of specimens recorded in the northeast region of Brazil from a specific herbarium. This database consists of a single table. All the records have latitude and longitude coordinates and in order to check if all specimens were originally recorded from the northeast region, an application was developed to check these records. The application was developed using xBASE compatible language and executed under Microsoft Visual FoxPro 7.0 ©.

The application was set with four coordinates to delimit a rectangle enclosing the North East region of Brazil: 01° 02' 30" S (*latmin*), 18° 20' 07" S (*latmax*), 34° 47' 30" W (*longmin*) and 48° 45' 24" W (*longmax*).

In sequence, the application checked all specimen coordinates against the referential coordinates for the northeast of Brazil, according to the following algorithm:

For each specimen record.

Given latitude and longitude.

If specimen longitude > longmax $\vee$ *specimen longitude < longmin*

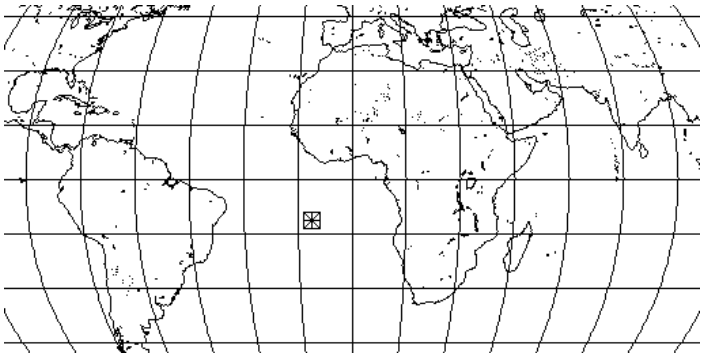
$\vee$ *specimen latitude > latmax* $\vee$ *specimen latitude < latmin*

then (the record lies outside the permitted region)

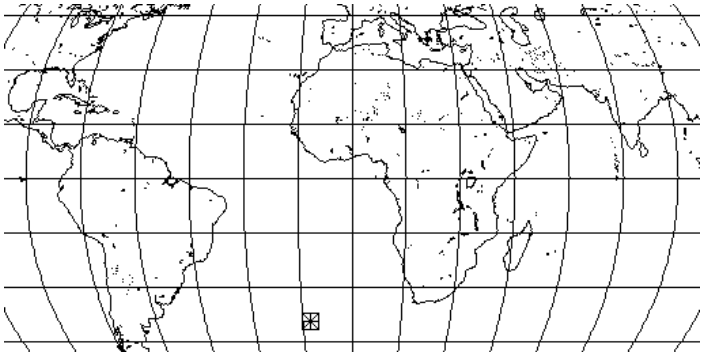
plot the spatial representation to specimen.

From 425 specimen records, two records were detected outside the spatial domain:

record# k000013118
LAT 11^{o} 33’ S – LONG 11^{o} 37’ W



record# k000013312
LAT 39^{o} 42’ S – LONG 13^{o} 30’ W



For record# k000013118, we assume the LONG value was erroneously transcribed with a latitude value, probably belonging to another specimen. For record# k000013312 we assume a swap between the latitude and the longitude value occurred.

In essence, a “domain” represents a set of all possible values of a database (records) or of an attribute (values). Therefore, detecting values outside the attribute domain is the basic task to attempt to improve data quality.

6.3.3. Scientific name spelling errors

Computers handle textual information – such as scientific names - as a string of characters, consisting mostly of a sequence of letters. However, although all scientific names are strings of characters, unfortunately not every string of characters is a scientific name.

A scientific name is a label - a textual representation - for a biological entity: the taxon. In taxonomic databases, this particular string of characters usually represents the main key to store, retrieve and access all the information related to a particular taxon or "entity". As pointed out by Bisby (1988), "most taxon-based information systems provide data-retrieval for a taxon starting from the taxon name". Thus, if this sequence of letters is incorrect, the data can't be retrieved, correct data will be linked with a wrong entity or a duplication of records that aim to represent the same entity will occur. The process of merging two or more different taxonomic databases is particularly vulnerable to duplication of records.

Built mainly by manual data entry or, more recently and on a small scale, by optical capture devices, taxonomic databases are particularly susceptible to data entry mistakes because of the nature of scientific names. The Latin language is a foreign language for everyone and the textual representation of the phonetic sound of Latin names could be easily misrepresented.

Spelling errors can be introduced in many ways; the following three are probably most important (Peterson, 1980).

- Unfamiliarity with the authors' names – Such errors can lead to consistent misspellings and are probably related to differences between how a word sounds and its actual spelling;
- Typographical errors on typing – These are less consistent but perhaps more predictable, since they are related to the position of keys on the keyboard and probably result from errors in finger movements.
- Transmission, conversion and storage errors – These are related to the specific encoding and transmission mechanisms, for example, optical character recognition (OCR) used in data entry.

Spelling error patterns vary greatly depending on the application task. For example, transcription-typing errors tend to reflect typewriter keyboard adjacencies, e.g., the substitution of *b* for *n*. In contrast, errors introduced by optical-character recognisers are more likely to be based on confusions due to feature similarities between letters, e.g., the substitution of *D* for *O*. Despite the difference of patterns, spelling errors can be of three different types (Kukich, 1992):

- Typographic errors
- Cognitive errors
- Phonetic errors

In the case of typographic errors (e.g., the -> teh, spell -> speel), it is assumed that the writer or typist knows the correct spelling but simply makes a motor coordination slip. The source of cognitive errors (e.g., receive -> recieve, conspiracy -> conspiricy) is presumed to be a misconception or a lack of knowledge on the part of the writer or

typist. Phonetic errors (e.g., abyss -> abiss) are a special class of cognitive errors in which the writer substitutes a phonetically correct but orthographically incorrect sequence of letters for the intended word.

Damerau (1964) found that approximately 80% of all misspelled words contained a single instance of one of the following four error types: insertion, deletion, substitution, and transposition. Misspellings that fall into this large class are often referred to as single error misspellings; misspellings that contain more than one such error have been dubbed multi-error misspellings.

In the following chapter of this thesis, experiments to detect spelling errors will be set up, investigating the error rates in taxonomic databases and the performance of different string similarity algorithms in detecting these errors.

7. Experiments in detecting spelling errors in scientific names

7.1. Introduction

Automatic spelling error detection and correction have been the subject of research since the early 1960's (Alberga, 1967; Damerau, 1964; Kukich, 1992) and have been related to three main problems: spelling checkers in word processor systems, retrieval of proper names in databases and, more recently, in word recognition in OCR processes and "pen-based interface" devices.

The spelling error detection problem can be defined as:

> Given an alphabet $\sum$ and a dictionary *D* consisting of strings in $\sum^*$, a string *s* is considered a spelling error if $s \in \sum^*$ and $s \notin D$.

This definition establishes a working framework which will be used in this thesis, considering the string *s* as a scientific name.

Spelling error detection in taxonomic databases can be a straightforward task when it involves scientific names that represent taxonomic hierarchies such as Family and Genus names. In this case, reference tables could be used as a dictionary, as explained in the case of structural nomenclatural error detection (Section 6.3.1). In contrast, species names, as a binomial combination of genus name and specific epithet, need a

different approach because, in most practical situations, a reliable dictionary of species names with the same set of domains as a database to be checked is not available.

7.2. Purpose of the tests

The process of detecting scientific name spelling errors in taxonomic databases presented in this thesis aims to minimize "one of the more tedious tasks of taxonomists: to check and update long lists of species names" (Froese, 1997), in order to improve the overall data quality.

Thus, the objective of the spelling error detection process is to detect and isolate potential spelling errors in scientific names[20], using similarity algorithms in order to identify pairs of scientific names which have a high degree of similarity to each other but are not exactly the same.

The process of spelling error detection presented in this thesis can be described as follows:

For each scientific name (SNa)

Compare with all other scientific names (SNb) in the database

If $SNa \approx SNb \wedge SNa \neq SNb$ *then*

List SNa, SNb

---

[20] The "scientific name" for the spelling error detection process is defined as a composition of the genus name and a specific epithet.

The process assumes that the existence of two scientific names with a “high degree of similarity” ($\approx$) but which are not exactly the same ($\neq$) might represent a spelling error.

In order to establish a reliable process, the performance of different algorithms to detect similarity between two scientific names will be investigated in these tests. In the experiments, we are classifying the error types into the following categories, with the following names:

- Extra or missing letter (insertion or deletion)
- Wrong letter (substitution)
- Transposition
- Multi-error

The error rates related to scientific name spelling errors in different taxonomic databases will be determined through the use of these different algorithms.

7.3. The Databases

7.3.1. The selection

Five databases were selected as raw material for the experiments:

Plant Databases:

1. **Brazilian Atlantic Rain Forest**

   This is a Species-oriented database developed and maintained by a research project at Rio de Janeiro Botanic Garden – Brazil. The Atlantic Rain Forest

Project Database is managed using the ALICE© Software (Allkin and Winfield, 1993).

2. **Brazilian Northeast checklist (Version 11)**

   This is a Species-oriented database developed and maintained by the Centro Nordestino de Informações sobre Plantas at the Federal University of Pernambuco – Brazil. The Plants of North-east Brazil Database is also managed using the ALICE© Software (Allkin and Winfield, 1993).

3. **International Legume Database & Information Service – ILDIS (South America Split, version 6)**

   The International Legume Database & Information Service (Bisby, 1986; Zarucchi, 1991; Zarucchi *et al*., 1993) used in this thesis is a dataset called "South America Split", version 6.0. It contains South America taxa for the family Leguminosae. The ILDIS Database is also managed using the ALICE© Software (Allkin and Winfield, 1993).

Animal Database:

4. **Marine Invertebrates**

   The Marine Invertebrates of New Zealand Database was offered by Dr Geoffrey B. Read, from the National Institute of Water & Atmosphere, Wellington, New Zealand. This database was built by importing "old free text", with no data quality control at data entry. The original format of the given database was a delimited ASCII text file.

Mixed Database:

5. **Species 2000 (Year 2002 Annual Checklist)**

   The Species 2000 Database (www.sp2000.org) is a compilation of the following databases:

   - Bacteria and Archaea from BIOS
   - Brown, green and red seaweeds from AlgaeBase
   - Amoebidales, Asellariales, Dimargaritales, Eccrinales, Harpellales, Kickxellales, Mucorales, Phyllachorales, Rhytismatales, Zoopagales, and, in part, Xylariales from CAB International's Species Fungorum
   - Fish (Classes Myxini, Cephalaspidomorphi, Elasmobranchii, Holocephali, Sarcopterygii, Actinopterygii) from FishBase
   - Birds, Isopods, Hydrozoan stony corals, turtles, crocodiles, and some groups of mammals, amphibians, molluscs, and crustaceans from ITIS
   - Scarabaeid beetles from World Scarabaeidae Database
   - Cephalopods from CephBase
   - Sea Anemones (Order Actiniaria) from Hexacorallians of the World
   - Phyla Cephalorhyncha, Chaetognatha, Ctenophora, Echiura, Gastrotricha, Hemichordata, Mesozoa, Phoronida, Sipuncula; classes Pogonophora, Pycnogonida, Scyphozoa; and orders Pennatulacea and Scleractinia from the UNESCO/IOC Register of Marine Organisms
   - Additional regional data from ITIS for groups not yet covered globally by Species 2000
   - Family Leguminosae (Fabaceae; the legumes) from ILDIS World Database of Legumes

- Families Casuarinaceae, Magnoliaceae and Irvingiaceae from IOPI Global Plant Checklist
- Order Potamogetonales (Cymadoceaceae, Posidoniaceae, Zosteraceae) and family Hydrocharitaceae (in Hydrocharitales) from the Seaweed 2000 Database

7.3.2. Data preparation

Accepted and provisional scientific names from the first three databases, originally managed by Alice Software©, were exported to a flat table using the application called ALEX, which is part of the Alice Software© application suite. The table generated by the ALEX application, with the accepted and provisional scientific names and compatible with the DBF format, is a part of a standard defined by Alice Software as a transfer format between Alice Databases and is called A.T.F. – Alice Transfer Format. Following this standard, the table with the accepted and provisional scientific names, called SAM.TDI, has the following structure:

Field Name	Type	Size
T_NO	Numeric	5
STATUS	Character	22
GENUS	Character	30
G_AUTHOR	Character	40
SPECIES	Character	30
S_AUTHOR	Character	40
RANK	Character	7
SUBSP	Character	30
SP_AUTHOR	Character	40
B_NO	Numeric	4

Table 8 – SAM.TDI table structure

As the ATF format permits repetition of scientific names, to allow more than one bibliographic reference to be associated with a scientific name through the key field

"B_NO", the tools described in sections 7.5.2 and 7.5.5 were set up to remove the redundancy of identical scientific names before comparing the similarity between the remaining unique names.

The New Zealand Marine Invertebrates Database was offered in a "comma delimited ASCII file", as the following example:

```
"omorii","Acartia","Arthropoda Crustacea",1269
"simplex","Acartia","Arthropoda Crustacea",1269
"teclae","Acartia","Arthropoda Crustacea",1269
"tranteri","Acartia","Arthropoda Crustacea",1269
"praerupta","Acasta","Arthropoda Crustacea",1238
"inhaerens","Acervulina","Protozoa",4023
"goliath","Acesta","Mollusca",3128
"patagonica","Acesta","Mollusca",3128
"curvirostris","Achaeus","Arthropoda Crustacea",875
"fissifrons","Achaeus","Arthropoda Crustacea",875
"minutissima","Achanthostomella","Ciliophora",2314
"communis","Achelia","Arthropoda Chilicerata",465
"dohrni","Achelia","Arthropoda Chilicerata",465
"spicata","Achelia","Arthropoda Chilicerata",465
"pegasus","Acheronia","Arthropoda Crustacea",684
…
```

This file was imported into the MS-EXCEL© spreadsheet program and exported as a DBF table, named as SAM.TDI. The field names for the genus and the specific epithet were set up to comply with the A.T.F standard.

The Species 2000 Database was offered in a MS-ACCESS© database format. The table SCINAMES was exported to a *comma delimited* file and imported into MS-EXCEL©. This spreadsheet was exported to a single table, DBF compatible, named as SAM.TDI and following the A.T.F. standard for the *genus* and *species* field names.

At the end of the conversion the five DBF compatible tables, named as SAM.TDI, contained at least two columns: GENUS and SPECIES. These tables, called *working tables*, had the following number of records, excluding the scientific name redundancy:

Working Table	No. of unique combinations of *genus + species*
Brazilian Atlantic Rain Forest	1,802
Brazilian Northeast checklist	7,691
ILDIS	15,616
Marine Invertebrates	6,745
Species 2000	26,850

Table 9 – Number of unique scientific names in databases used in the spelling error detection experiments

7.3.3. The invalid characters problem

In the first round of tests the presence of invalid characters was detected in all databases, as a result of the structural validation process, considered in section 6.3.1. In the second round of tests, these characters were removed in order to avoid compromising the execution of the spelling error algorithms.

Invalid characters, characters not in the allowed set [A..Z], [a..z] and "-", are often found associated with scientific names as a result, in most of cases, of a combination of poor data modelling and doubtful or imprecise taxon determination. Although the use of abbreviated status flags, such as *cf.*, *aff.*, *sp.nr.*, etc. is commonplace in association with species names, better data structures will provide a specific field for these status flags.

Genus	Species	Database
Panicum	aff.nervosum	PMA
Persea	cf.pyrifolia	PMA
Ophiactis	abyssicola var.cuspidata	Marine
Gymnangium	gracilicaule/armatum	Marine
Pterocarpus	sp.1	ILDIS
Dalbergia	sp.nr.macrosperma	ILDIS
Acetobacter (subgen.Acetobacter)	aceti	SP2K
Zornia	?gemella	CNIP
Trachypogon	Ness.	CNIP

Table 10 – Examples of use of invalid characters found in databases tested

Scientific names with invalid characters in the five databases tested (excluding the control database) were dynamically filtered from the working tables at the execution of the second round of tests using the following function:

```
function charerror(nome:string):boolean;
const
 caracters:string = ' ?,.<>:;"[]{}|\+=_)(*&^%$#@!';
var
 i:integer;
Begin
 charerror := false;
 For i:= 0 to 26 do
 Begin
 if pos(caracters[i],nome)>0 then charerror:= true;
 end;
end;
```

This algorithm assumes that any characters not in the specified list is valid, and thus it might need re-implementing if a different or larger character set, such as Unicode, is used. A specific application was developed in order to quantify the structural errors, specifically the presence of invalid characters in the *genus* or *species* fields, for the five working tables. The application generates a list of 313 names with invalid characters found in the five databases (Appendix I).

Database	No. of names with invalid characters
Atlantic Rain Forest	64
North-east checklist	43
ILDIS	134
Marine Invertebrates	67
Species 2000	5

Table 11 – Number of scientific names with invalid characters found in the databases

The most frequent invalid character found was the '.' (full stop), which was present in 254 of 313 returned names from all the five databases. The '.' (full stop) character is frequently associated with notations as *cf.*, *aff.*, and *sp.*, denoting uncertain or unfinished classification of certain taxa.

The presence of numbers was detected in 58 of 313 names, associated with the *sp.* notation, usually denoting undescribed or undetermined species.

The presence of invalid characters in the *genus* name was almost restricted to the characters '(' and ')', as result of the *domain schizophrenia* error pattern (see section 6.2.5), caused mostly by a poor data model and the consequent necessity to accommodate more than one name element in a single field (for example: "*Balanus (Megabalanus)*" and "*Holothuria (Lessonothuria)*").

7.4. The algorithms

A number of different types of algorithms were chosen in order to evaluate their performance in handling error detection in scientific names. These algorithms can be separated into two main categories: phonetic similarity algorithms and plain string similarity algorithms.

7.4.1. Phonetic similarity algorithms

Phonetic matching is used to identify strings that may be of similar pronunciation, regardless of their actual spelling. A typical application is a "white pages" enquiry line, where a telephone operator is verbally given a name, guesses at the spelling (or is provided with a spelling, which may be incorrect), and uses the guess to query a database of names. The phonetic matching system must then find in the database those strings most likely to be of the same or similar pronunciation to that of the query. Since there is no reliable way of automatically determining the pronunciation of a string in the English language, such matching must be inexact (Zobel and Dart, 1996).

Phonetic similarity algorithms, also known as phonetic matching techniques, encode strings of letters (names) according to their phonetic sounds, where the word “phonetic” refers to spoken sounds and not to the spelling of words. Names that share the same phonetic codes are assumed to be phonetically similar.

The Soundex algorithm

The *Soundex* algorithm has been used since the start of the twentieth century to retrieve names based on pronunciation rather than spelling. It was originally developed to search for names in the 1890 USA census files. *Soundex* indexes to the U.S. Census were compiled in the 1930s because wide-spread misspellings caused difficulties for the Social Security Administration in matching names of persons applying for old age benefits who had no birth or other proof of age records (Roughton and Tyckoson, 1985).

The original Soundex algorithm was patented by Margaret O'Dell and Robert C. Russell in 1918. The method is based on the six phonetic classifications of human speech sounds (bilabial, labiodental, dental, alveolar, velar, and glottal), which in turn are based on where you put your lips and tongue to make the sounds.

Soundex uses codes based on the sound of each letter to translate a string into a canonical form of at most four characters, preserving the first letter.

The Soundex algorithm can be described by the following outline:

1) Retain the first letter of the name, and drop all occurrences of a, e, h, i, o, u, w, and y in other positions.

2) Assign the following numbers to the remaining letters after the first:
    1. b, f, p, v
    2. c, g, j, k, q, s, x, z
    3. d, t
    4. l
    5. m, n
    6. r
3) If two or more letters with the same code were adjacent in the original name, omit all but the first.
4) Convert to the form letter, digit, digit, digit by adding trailing zeroes if there are less than 3 digits already, or by truncation if there are more than 3.

For example, *reynold* and *renauld* are both reduced to *r543*, but, more commonly, Soundex makes the error of transforming dissimilar-sounding strings such as *catherine* and *cotroneo* to the same code, and of transforming similar-sounding strings to different codes. There is no ranking of matches: strings are either similar or not similar.

The Phonix algorithm

The *Phonix* algorithm was developed to be used as a retrieval phonetic technique with the URICA library package, on bibliographical databases (Gadd, 1988; 1990).
The Phonix algorithm is a *Soundex* variant. While Soundex's phonetic property is restricted to the collection of similar sounding consonants into different classes, the algorithm for computing the Phonix codes uses elaborate substitution rules (Pfeifer *et al*., 1996).

Letters are mapped to a set of codes using the same algorithm, but a slightly different set of codes is used, and prior to mapping about 160 letter-group transformations are used to standardise the string. For example, the sequence *tjV* (where V is any vowel) is mapped to *chV* if it occurs at the start of a string, and *x* is transformed to *ecs*. These transformations provide context for the phonetic coding and allow, for example, *c* and *s* to be distinguished. The Phonix codes are shown below:

The categories 1 – 6 of *Phonix* are basically the same as those used in the Soundex algorithm, except that some letters from categories 1 and 2 are put into the new Phonix categories 7 and 8.

Phonetic Substitutions (The string SUB replaces the specified letters found at the START, MIDDLE or END of the word)			
SUB	START	MIDDLE	END
G	DG	DG	DG
KO	CO	CO	CO
KA	CA	CA	CA
KU	CU	CU	CU
SI	CY	CY	CY
SE	CE	CE	CE
KL	CL if CL*v*		
K	CK	CK	CK
K			GC
K			JC
KR	CHR if CHR*v*		
KR	CR if CR*v*		
R	WR		
NK	NC	NC	NC
KT	CT	CT	CT
F	PH	PH	PH
AR	AA	AA	AA
SH	SCH	SCH	SCH
TL	BTL	BTL	BTL
T	GHT	GHT	GHT
ARF	AUGH	AUGH	AUGH
LD		LJ if *v*LJ*v*	
LOW	LOUGH	LOUGH	LOUGH
KW	Q		
N	KN		

N			GN
N	GHN	GHN	GHN
N			GNE
NE	GHNE	GHNE	GHNE
NS			GNES
N	GN		
N		GN if GN*c*	GN if GN*c*
S	PS		
T	PT		
C	CZ		
Z		WZ if *v*WZ	
CH		CZ	
LSH	LZ	LZ	LZ
RSH	RZ	RZ	RZ
S		Z if Z*v*	
TS	ZZ	ZZ	ZZ
TS		Z if *c*Z	
REW	HROUG	HROUG	HROUG
OF	OUGH	OUGH	OUGH
KW		Q if *v*Q*v*	
Y		J if *v*J*v*	
Y	YJ if YJ*v*		
G	GH		
E			GH if *v*GH
S	CY		
NKS	NX	NX	NX
F	PF		
T			DT
TIL			TL
DIL			DL
ITH	YTH	YTH	YTH
CH	TJ if TJ*v*		
CH	TSJ IF TSJ*v*		
T	TS if TS*v*		
CH	TCH	TCH	TCH
VSKIE		WSK if *v*WSK	WSK if *v*WSK
N	MN if MN*v*		
N	PN if PN*v*		
SL		STL if *v*STL	STL if *v*STL
ENT			TNT
OH			EAUX
ECS	EXCI	EXCI	EXCI
ECS	X	X	X
ND			NED
DR	JR	JR	JR
EA			EE
S	ZS	ZS	ZS
AH		R if *v*R*c*	R if *v*R*c*

AH		HR if *v*HR*c*	HR if *v*HR*c*
AH			HR if *v*HR
AR			RE
AH			R if *v*R
LE	LLE	LLE	LLE
ILE			LE if *c*LE
ILES			LES if *c*LES
null			E
S			ES
AS			SS if *v*SS
M			MB if *v*MB
MPS	MPTS	MPTS	MPTS
MS	MPS	MPS	MPS
MT	MPT	MPT	MPT
Where v = vowel and c= consonant			
▪ Vowels a, e, i, o and u are ignored ▪ Consonants y h and w are ignored ▪ The second of any successive identical characters is ignored ▪ Non alphabetical characters are ignored ▪ Key length: 7 significant consonants ▪ Key format: Annnnnnn, where *n* represents a numerical value from 1 to 8 as defined by Table 13 below and *A* is the first character of the name AFTER the phonetic substitutions have been completed			

Table 12 – Phonix Phonetic Substitutions

The Phonix algorithm can be described by the following outline:

a) Perform phonetic substitutions (see Table 12);

   1. Only the specified characters are dropped, eg. The *v* or vowel is not dropped in the substitution of 'N' for 'PN' when 'PN*v*' is true;
   2. The substitutions are applied in the specified order;
   3. Process all occurrences of one substitution before proceeding to the next substitution parameter;
   4. The result of a substitution may create a new target string for substitution by subsequent parameters.

b) Retain the first character for the retrieval code.

c) Replace by '*v*' if A, E, I, O, U or Y.

d) Where names end in ES, drop the E.

e) Append an E where names end in A, I, O, U or Y.

f) Drop the last character regardless.

g) Drop the new last character if not A, E, I, O, U or Y.

h) Repeat g) until a vowel (including Y) is found.

i) Repeat the next three operations for each remaining character

   1. Strip any occurrence of A, E, I, O, U, Y, H and W.

   2. Retain one of any successive duplicate consonants.

   3. Replace the consonant by its numeric value (see Table 13).

j) Prefix the retrieval code with the retained first character (may be a '*v*').

Numeric Character Values							
1	2	3	4	5	6	7	8
B P	C G J K Q	D T	L	M N	R	F V	S X Z

Table 13 – Phonix Numeric Character Values

Examples of the codes produced by the Soundex and Phonix algorithms are shown in Table 14.

	Soundex code	**Phonix code**
macrocefalo	M262	M526
macrocephalum	M262	M526
malacophylla	M421	M542

Table 14 – Examples of Soundex and Phonix algorithms

7.4.2. String similarity algorithms

n-Gram

String similarity algorithms compute a numerical estimate of the similarity between two strings. *N*-grams are *n*-letter sub-sequences of words or strings, where *n* is usually one, two or three letters. If *n* is greater than 1, *n*-grams may overlap, in other words a letter may appear in more than one n-gram. If two strings are compared with respect to their *n*-grams, the set of all possible *n*-grams will be calculated separately for each string. Next, these sets will be compared, and the more *n*-grams occur in both of the sets, the more similar the two strings are. The *n*-gram method counts the number of *n*-grams which the two strings have in common, and calculates a *similarity coefficient*.

The general approach to this calculation of a similarity coefficient is performed by the following equation, where N_1 and N_2 are the *n*-gram sets of the two words to be compared (Pfeifer *et al*., 1996):

$$similarity\,coefficient := \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|}$$

One parameter that has to be chosen is the length of the *n*-gram. For the experiments, two lengths of the *n*-gram were tested: *n = 2* (bigram), see Table 15, and *n = 3* (trigram).

	li	ic	ca	an	ni	ia	ly	yc	Similarity coefficient
Licania	1	1	1	1	1	1			$\frac{4}{8} = 0.5$
Lycania			1	1	1	1	1	1	

Table 15 – Example of a bi-gram similarity coefficient calculation

Levenshtein distance

The Levenshtein distance (LD) is a measure of the similarity between two strings, which we will refer to as the source string (s) and the target string (t). The distance is the number of single-character deletions, insertions, or substitutions required to transform s into t. The greater the Levenshtein distance, the more different the strings are. For example,

- If s is “test” and t is “test”, then LD(s,t) = 0, because no transformations are needed. The strings are already identical.
- If s is “test” and t is “tent”, then LD(s,t) = 1, because one substitution (change “s” to “n”) is sufficient to transform s into t.

The Levenshtein distance is named after the Russian scientist Vladimir Levenshtein, who devised the algorithm in 1965. If you can't spell or pronounce Levenshtein, the metric is also sometimes called edit distance. The Levenshtein distance algorithm has been used in:

- Spell checking
- Speech recognition
- DNA analysis
- Plagiarism detection

The Levenshtein distance algorithm can be described by the following outline:

a) Set n to be the length of s.

Set m to be the length of t.

If n = 0, return m and exit.

If m = 0, return n and exit.

Construct a matrix containing 0..m rows and 0..n columns.

b) Initialize the first row to 0..n.

Initialize the first column to 0..m.

c) Examine each character of s (i from 1 to n).

d) Examine each character of t (j from 1 to m).

e) If s[i] equals t[j], the cost is 0.

If s[i] doesn't equal t[j], the cost is 1.

f) Set cell d[i,j] of the matrix equal to the minimum of:

The value of the cell immediately above plus 1: d[i-1,j] + 1.

The value of the cell immediately to the left plus 1: d[i,j-1] + 1.

The value of the cell diagonally above and to the left plus the cost: d[i-1,j-1] + cost.

g) After the iteration steps (c, d, e, f) are complete, the distance is found in cell d[n,m].

Example of Levenshtein Distance Algorithm

This section shows how the Levenshtein distance is computed when the source string is “GUMBO” and the target string is “GAMBOL”.

Steps *a* and *b*

		G	U	M	B	O
	0	1	2	3	4	5
G	1					
A	2					
M	3					
B	4					
O	5					
L	6					

Steps *c* to *f* when i = 1

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0				
A	2	1				
M	3	2				
B	4	3				
O	5	4				
L	6	5				

Steps *c* to *f* when i = 2

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1			
A	2	1	1			
M	3	2	2			
B	4	3	3			
O	5	4	4			
L	6	5	5			

Steps *c* to *f* when i = 3

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2		
A	2	1	1	2		
M	3	2	2	1		
B	4	3	3	2		
O	5	4	4	3		
L	6	5	5	4		

Steps *c* to *f* when i = 4

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2	3	
A	2	1	1	2	3	
M	3	2	2	1	2	
B	4	3	3	2	1	
O	5	4	4	3	2	
L	6	5	5	4	3	

Steps *c* to *f* when i = 5

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2	3	4
A	2	1	1	2	3	4
M	3	2	2	1	2	3
B	4	3	3	2	1	2
O	5	4	4	3	2	1
L	6	5	5	4	3	2

Step *g*

The distance is in the lower right hand corner of the matrix, i.e. 2. This corresponds to our intuitive realization that “GUMBO” can be transformed into “GAMBOL” by substituting “A” for “U” and adding “L” (one substitution and 1 insertion = 2 changes).

Skeleton-Key

Pollock and Zamora (Pollock and Zamora, 1984) presented a technique that is based on the idea that the consonants carry more information than the vowels. In contrast to phonetic techniques, the Skeleton-key codes the string based on the occurrence of the

consonants, followed by the occurrence of the vowels, rather than complex substitutions based on phonetic characteristics.

The Skeleton-key of a word consists of its first letter, followed by the remaining consonants and finally the remaining vowels, both in order of appearance. This key contains every letter at most once by eliminating duplicates (Pfeifer *et al*., 1996).

	Skeleton-key
Mobiliensi	MBLNSOIE
Mobilinensis	MBLNSOIE
Molibidensis	MLBDNSOIE

Table 16 – Examples of Skeleton-key codes

7.5. Running the tests

7.5.1. Hardware and development environment

The following set of tools was developed using the Borland® Delphi® 5.0 compiler, Enterprise Suite version, build 5.62. The source code for these tools is available from the author on request. A third party string manipulation library was used. This library, called TStringManager (http://www.geocities.com/ericdelphi/StrMan.html) was developed by Eric Grobler and is freely available to use.

The Microsoft Visual FoxPro 7.0 © Database Management System was used for general manipulation of DBF compatible tables and command line operations.

The hardware used to develop the tools and run the tests was an IBM-PC compatible, under the MS-Windows© 2000 Professional operating system.

7.5.2. Error classification

As pointed out by Klein (2001), there are four possible outcomes in error detection tasks: *Hit*, *Miss*, *False alarm* and *No action* (see Table 17).

	Behaviour of algorithm	
Input data	*Error detected*	*Error not detected*
Error exists	Hit	Miss
Error does not exist	False alarm	No action

Table 17 – Outcomes in error detection tasks (Klein, 2001)

Thus, an "appropriate algorithm" should be that one which has the highest rates of *hits* and *no action* and, lower or nil rates of *false alarms* and *misses*.

The missed errors problem

Since the errors present in the selected databases were not previously known, the category of existing errors which were "missed" could not be detected by the detection tool. Aiming to correct this deficiency at the experiment design, a tool was developed to generate a control database to be added to the experiments with the previous five selected databases, with the following arbitrary characteristics:

- 500 scientific names
- names selected randomly from the 5 selected databases
- similarity between the names (genus + species) ≤ 0.2 using the bigram algorithm

After the database generation, 25 additional scientific names were manually entered as a copy of an existing correct scientific name with one or more spelling errors, as follow:

Original name (correct)	Manually entered name (error)	DB source	Error type
Adenocarpus foliolosus	Adenocarpus folioloso	ILDIS	MU
Aeschynomene katangensis	Aeschynomene katangienses	ILDIS	MU

Alloiopodus pinguis	Aloiopodus pinquis	MARINE	MU
Artediellus gomojunovi	Artediellus gomoiunovi	SP2K	W
Cnemidocarpa bicornuata	Cnemidocarpa bicornata	MARINE	EX
Mayaca aubletii	Maiaca aubleti	CNIP	MU
Mayaca aubletii	Maiaca aubletti	CNIP	EX
Nannoperca oxleyana	Nanoperca oxleyana	SP2K	EX
Orthopappus angustifolius	Orthopappus angustifolia	PMA	MU
Pallenopsis obliqua	Pallenopsis oblicua	MARINE	W
Phractura longicauda	Phractura longicaudata	SP2K	MU
Rhytachne subgibbosa	Rhytachne subgibosa	CNIP	EX
Sparattosperma leucanthum	Sparattosperma leucanticum	CNIP	MU
Spilanthes acmella	Spilanthes aquimella	CNIP	MU
Strychnos brasiliensis	Strychnos brasiliense	PMA	MU
Swainsona cyclocarpa	Swainsona ciclocarpus	ILDIS	MU
Synodontis budgetti	Synodontis budgeti	SP2K	EX
Synodontis budgetti	Synodontis budgetii	SP2K	W
Syzygium jambolanum	Syzyguim jambolanum	CNIP	T
Tovomitopsis saldanhae	Tovomitopisis saldanhai	PMA	MU
Vicia murbeckii	Vicia murbecki	ILDIS	EX
Vicia sylvatica	Vicia silvatica	ILDIS	W
Wulffia stenoglossa	Wulfia stenoglosa	CNIP	MU
Zemysina globus	Zerrysina globus	MARINE	MU
Zornia gemella	Zonia gomela	CNIP	MU

Legend:

DB Source:
ILDIS: ILDIS Database
MARINE: Marine Invertebrates of New Zealand Database
SP2K: Species 2000
PMA: Atlantic Rain Forest Database
CNIP: North-east checklist Database

Error Type:
W: one wrong letter
T: transposition of two adjacent letters
EX: one extra letter or missing letter
MU: multi error

Table 18 – Spelling errors in the control database

The control database, with a total of 525 scientific names, including the 25 arbitrarily selected misspelled scientific names listed in Table 18, was added to the experimental procedures, in the same manner as the other five selected databases.

7.5.3. Algorithm Implementation Testing Tools

A first set of tools was developed with the objective of checking the correct implementation of the algorithms, using the Object Pascal Language, compatible with

the Delphi compiler. Each algorithm was implemented as a function and a simple interface was developed in order to permit the input of one or two strings, to execute the function and to show the result.

The first tool developed (*Tool #1*) was the one which tested the implementation of the phonetic similarity algorithms and the Skeleton Key algorithm. The Soundex and Phonix algorithms were implemented in Object Pascal, based on a implementation using the *C* language, available on the Internet at http://search.cpan.org/src/ULPFR/WAIT-1.800/soundex.c This implementation of the Phonix code was built based on the original papers by Gadd (1988; 1990). The Skeleton Key implementation was built from scratch, based on the definition of the algorithm published by Pfeifer, Poersch, and Fuhr (1995; 1996).

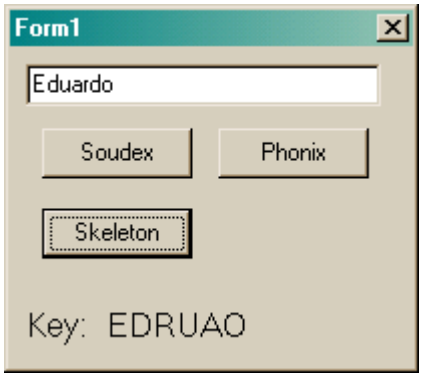


Figure 13 – Interface of the tool #1

A new version of the *Tool #1*, called here *Tool #2*, was developed for the validation of the implementation of the *n*-gram algorithms – *2*-gram ("Bigram") and *3*-gram ("Trigram"). The implementation of the *n*-gram algorithm was developed from scratch, based on the definitions published by Pfeifer *et al*. (1995; 1996) and Zamora *et al*. (1981).

Figure 14 – Interface of the Tool #2

A different tool, called *Tool #3*, was developed in order to validate the implementation of the Levenshtein distance algorithm. The implementation of the Levenshtein distance algorithm was based on the Delphi implementation by Alvaro Jeria Madariaga, freely available at http://www.merriampark.com/lddelphi.htm.

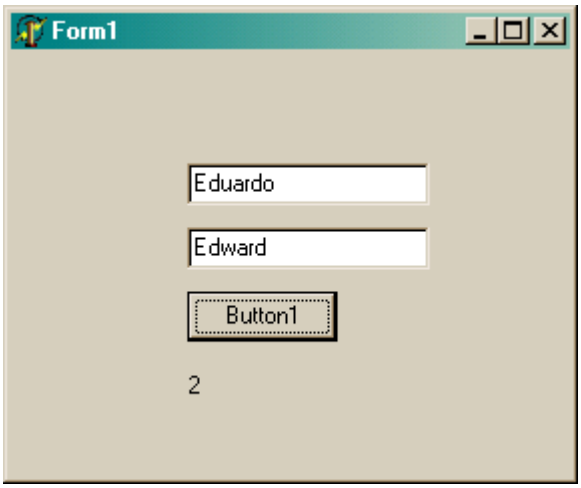


Figure 15 – Interface of the Tool #3

7.5.4. First round of tests

The first round of tests consisted of developing and running specific tools over the selected datasets in order to establish a first evaluation of the performance of the tools,

the methodology adopted and the experiment design as a whole. The following Figure 16 summarizes the schema for the first round of tests.

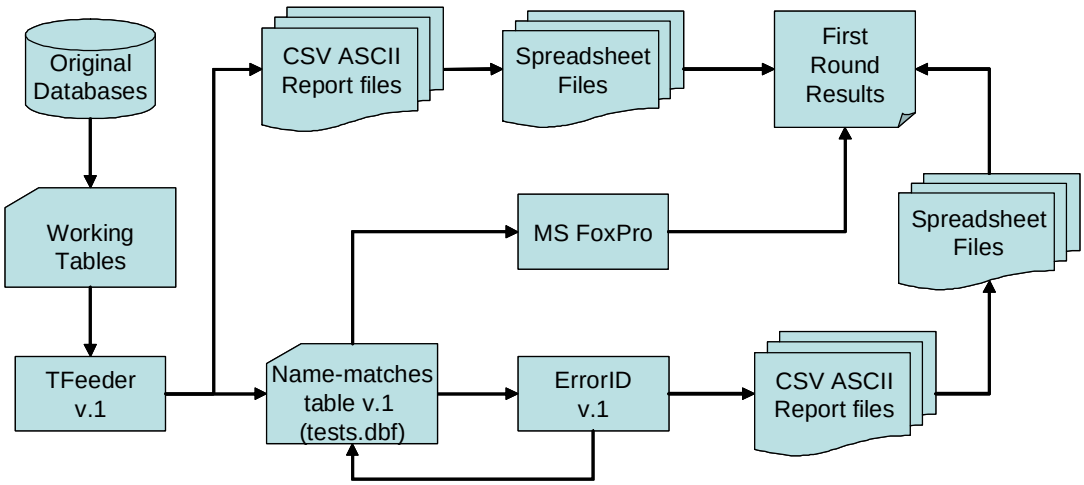


Figure 16 – Schema for the first round of tests

Spelling Error Detection Tool

Considering that the implementations of the algorithms were returning the expected results, a first version of a tool called *TFeeder* was developed. This version of the *TFeeder* was set up to receive the directory path to the database and the selection of the tests.

In the case of the *n*-gram test, two values for the parameter *n* were chosen (*n*=2 for bigram and *n*=3 for trigram) because, according to Zamora *et al*. (1981), bigrams and trigrams achieve the best results in retrieving words similar to a given word. Therefore, the detection tool was set up to detect plain string similarity based on the *n*-gram algorithm using *n*=2 (bigram) and *n*=3 (trigram). After considering various approaches by developing tools and running preliminary tests, three different approaches for the use of these *n*-gram tests were chosen for further investigation:

1. The "*-2*" approach

   The *-2* approach can be represented by the following steps:

   a. Merge the genus and species strings of the scientific name #1
   b. Merge the genus and species strings of the scientific name #2
   c. Calculate the number of exclusive *n*-grams
   d. List the pair if the number of exclusive *n*-grams ≤ 2

2. The "*-1*" approach

   The *-1* approach can be represented by the following steps:

   a. Merge the genus and species strings of the scientific name #1
   b. Merge the genus and species strings of the scientific name #2
   c. Calculate the number of exclusive *n*-grams
   d. List the pair if the number of exclusive *n*-grams ≤ 1

3. The "1.5" approach (15)

   The 1.5 approach can be represented by the following steps:

   a. Calculate the similarity index between the genus strings of the scientific names #1 and #2
   b. Calculate the similarity index between the species strings of the scientific names #1 and #2
   c. List the pair of names if the sum of the similarity indexes ≥ 1.5

These approaches were used with both the bigram and trigram algorithms, to generate six variant algorithms, as follows: 2-gram –1, 2-gram –2, 2-gram 1.5, 3-gram –1, 3-gram –2 and 3-gram 1.5.

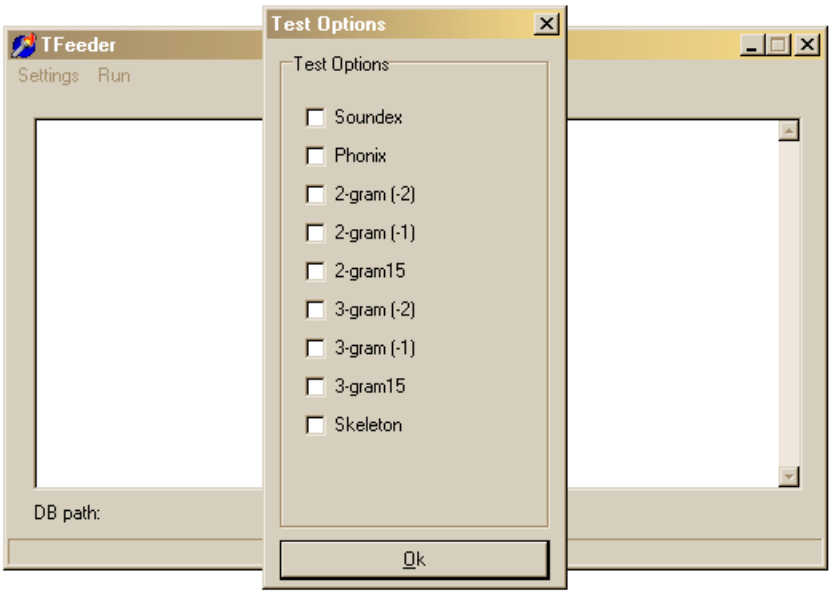


Figure 17 – Interface of the TFeeder tool, version 1

Version 1 of the *TFeeder* tool was set up to generate a different report for each working database table, in a CSV ASCII File, and populate a single flat file table (*tests.dbf*), one for each working table (taxonomic database) analysed, with the names detected by the algorithms.

Name	Type	Width	Description
GENUS1	Character	35	Genus of 1st scientific name
SPECIES1	Character	35	Species of 1st scientific name
GENUS2	Character	35	Genus of 2nd scientific name
SPECIES2	Character	35	Species of 2nd scientific name
CLASS	Character	5	The classification of the Error
2GRAM1	Character	5	Flag for the 2gram -1 test
2GRAM2	Character	5	Flag for the 2gram -2 test
2GRAM15	Character	5	Flag for the 2gram 1.5 test
3GRAM1	Character	5	Flag for the 3gram -1 test
3GRAM2	Character	5	Flag for the 3gram -2 test
3GRAM15	Character	5	Flag for the 3gram 1.5 test
SOUNDEX	Character	5	Flag fort Soundex test
PHONIX	Character	5	Flag for Phonix test
SKELETON	Character	5	Flag for Skeleton test

Table 19 – The structure of the table *tests.dbf*

```
genus1, species1, genus2, species2, error, soundex, skeleton, 2gram1, 2gram2, 3gram1, 3gram2
Beilschmedia,emarginata,Beilschmiedia,emarginata,,Y,Y,,Y,,
Beilschmedia,taubertiana,Beilschmiedia,taubertiana,,Y,Y,,Y,,
Brosimum,glaziovi,Brosimum,glaziovii,,Y,Y,Y,Y,Y,Y
Calymperes,lanchophyllum,Calymperes,lonchiphyllum,,Y,,,,,
Cecropia,glaziovi,Cecropia,glaziovii,,Y,Y,Y,Y,Y,Y
Chomelia,estrelana,Chomelia,estrellana,,Y,Y,Y,Y,,Y
Eugenia,tinguiensis,Eugenia,tinguyensis,,Y,,,Y,,
Geonoma,blanchettiana,Gouania,blanchetiana,,Y,,,,,
Inga,lentiscellata,Inga,lentiscifolia,,Y,,,,,
Jacaranda,macrantha,Jacaranda,micrantha,,Y,,,Y,,
Leandra,cf.gracilis,Leandra,cf.sylvestris,,Y,,,,,
....
```

Example of the CSV ASCII file generate by the TFeeder

The CSV ASCII file was imported to a spreadsheet where the features of indexing, counting and filtering were used to produce the first evaluations of the tests.

Spelling Error Classification Tool

A first error identification tool was developed in order to classify the errors. The *ErrorId* v.1 was set up to work together with the *tests.dbf* table, generated by the *TFeeder* v.1 tool and set up to classify the errors into the following categories:

1) Single errors (*Hit*)
    a) One extra or one missing letter (*Vicia faba* : *Vicia fabba*)
    b) One wrong letter (*Vicia faba* : *Vicia fada*)
    c) Transposition of two adjacent letters (*Vicia faba* : *Viica faba*)
2) Multiple errors (*Hit*)
    a) More than one “single error” (*Vicia faba* : *Viica fabba*)
3) Unrelated (*false alarm)*
    a) Different species epithet (*Vicia faba* : *Vicia allata*)
    b) Different genus (*Vicia glauca* : *Acacia glauca*)

4) Uncertainty (*doubt*)

   a) Uncertainty between unrelated (*false alarm*) and "Multiple error" (*Lysiloma latisiliqua* : *Lysiloma latisiliquum*)[21]

5) Structural error[22]

   a) Presence of "invalid character" (*Dalbergia* sp.1 x *Dalbergia sp.nr.macrosperma*)

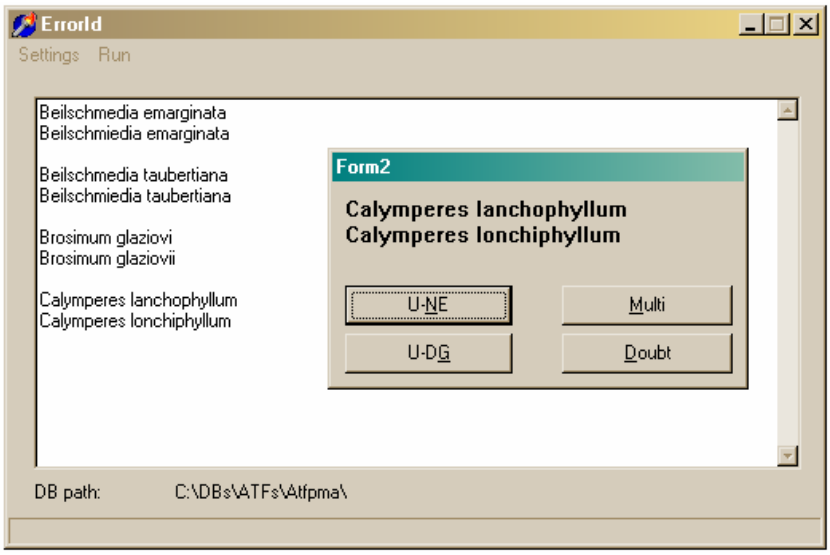


Figure 18 – The ErrorId V.1 interface

The *ErrorId* tool was designed to classify automatically all the single errors and structural errors, but to interact with the user in case of *multiple errors*, *false alarms* or *doubt*. The following set of rules was chosen to classify the errors following Damerau (1964), as far as possible:

1) If the difference in the lengths of the genus strings is greater than 3: *false alarm* involving different genus strings (U-DG)

[21] This specific example can be also a specialized kind of error in the taxonomic domain, called "gender error"
[22] See section 7.3.3

2) If the genus strings are identical but the difference in the lengths of the species strings is greater than 3: *false alarm* involving different species strings (U-NE)
3) If the difference in the lengths of the genus or species strings is equal to 1, check for the *extra/missing letter* error (E)
4) If the genus and species strings have the same length, check for the *transposition error* (T) and the *wrong letter* error (W)
5) If none of the preceding rules applies to the pair of scientific names, then present a interactive interface for the user to classify the type of error, with the addition of the options *multi-error* – more than one single error, and *doubt* – a error that cannot be classified as one of the errors mentioned above.

The *ErrorId* tool also updates the table *tests.dbf* with the error classification and generates a CSV ASCII file report, as shown below:

```
genus1, species1, genus2, species2, error, soundex, skeleton, 2gram1, 2gram2, 3gram1, 3gram2, phonix, 2gram15, 3gram15
acanthochitona,zelandica,acanthochitona,zelandicus,D,,,,,,,,
Acitellina,urinatoni,Acitellina,urinatoria,D,D,D,,D,,,D,D,D
Acitellina,urinatoni,Ascitellina,urinatoria,D,D,,,,,,D,,
Acitellina,urinatoria,Ascitellina,urinatoria,E,E,,,E,,E,E,E,E
Actaecia,euchroa,Actaecia,eughroa,W,W,,,W,,,W,W,
Actaecia,euchroa,Actoecia,euchiroa,M,M,,,,,,M,,
Actaecia,eughroa,Actoecia,euchiroa,M,M,,,,,,M,,
Adelasca,joqalqa,Adelascopora,jegolqa,U-DG,U-DG,,,,,,U-DG,,
Adelasca,joqalqa,Adelascopora,jeqolqa,U-DG,U-DG,,,,,,U-DG,,
Adelascopora,jegolqa,Adelascopora,jeqolqa,W,W,,,W,,,W,W,
Aeneator,valedicta,Aeneator,valedictus,D,D,,,D,,D,D,D,D
Aglaophamus,macoura,Aglaophamus,macroura,E,E,E,,E,,,E,E,
Aglaophamus,macoura,Aglaophamus,macrura,W,W,,,W,,,W,W,
Aglaophamus,macroura,Aglaophamus,macrura,E,E,,E,E,,E,E,E,
Aglaophamus,verrilli,Aglaophamus,verrillii,E,E,E,E,E,E,E,E,E
...
```

Example of the CSV ASCII file generate by the *ErrorId* tool

The assessment of the first round of tests

At the end of the first round of the tests some conclusions were made:

1. The process of importing the reports on CSV files to spreadsheets, one spreadsheet to each different working table, and using the spreadsheet as a analysis tool, could be improved with the use of a single table to store all the names detected in all tests and for all databases tested;
2. Some implementations of the algorithms should be optimized in order to reduce the time taken by the test with the large databases. Since each name in a database is compared to all other names present in the database, except with itself, the number of comparisons performed by the detection tool is represented by (*number of names* × *number of names-1*) / 2 (Table 20). Because of the large number of comparisons, the Species 2000 database test could not be completed in the first round, with the test interrupted after 2 days running.

Database	**No. of unique combinations of *genus* + *species***	**No. of comparisons performed by detection tool**
Atlantic Rain Forest	1,802	1,622,701
Northeast checklist	7,691	29,571,895
ILDIS	15,616	121,921,920
Marine Invertebrates	6,745	22,744,140
Species 2000	26,850	360,447,825
Control Database	525	137,550

Table 20 – Number of comparisons performed by the detection tool

3. The “Doubt” error classification was not represented in the four possible outcomes in error detection tasks (see Table 17, pg. 113), pointed out by Klein (2001) and adopted in this thesis as a standard for the tests.

7.5.5. The second round of tests

A second round of tests was defined to overcome the problems detected during the first round.

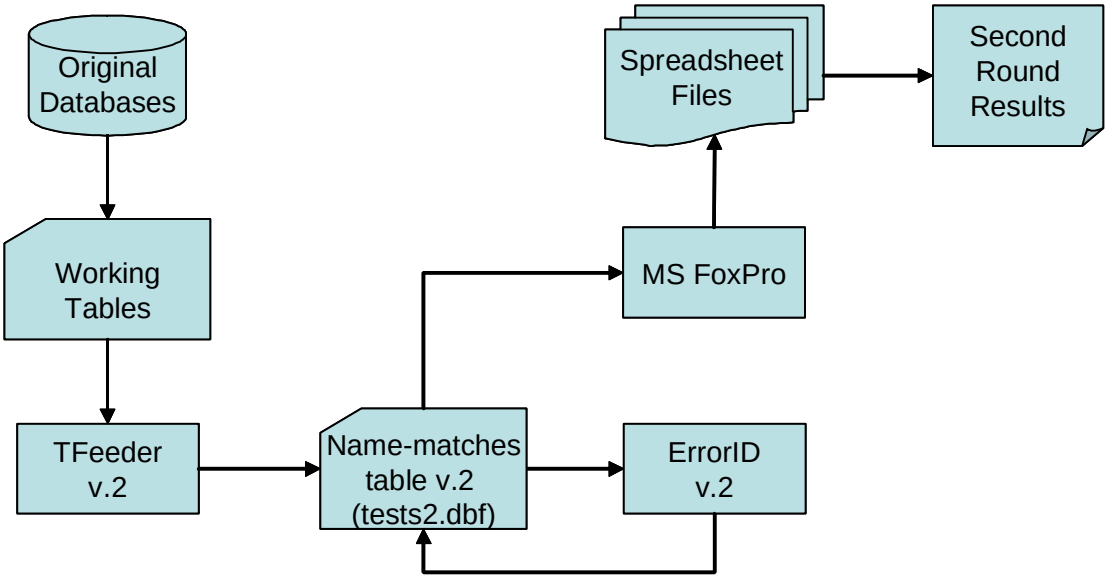


Figure 19 – Schema for the second round of tests

The second round of tests (Figure 19) was improved with a better methodology to handle and analyze the results, as a consequence of the improvement of the previously developed tools with the following characteristics:

Spelling error detection tool

The new version of the *TFeeder* tool was developed with two main improvements:

1. Improved implementation of the algorithms for fast processing;
2. A new structure for the name-matches table (*tests2.dbf)* in order to store the pairs of scientific names and all the other attributes which permit the later performance analysis of the tests from a single data source. The new table was created as a DBF compatible file in order to make possible the execution of quick analyses, based on the command line environment of the Microsoft Visual FoxPro 7.0 © Database Management System.

Name	Type	Width	Dec	Description
GENUS1	Character	30		Genus of 1st scientific name
SPECIES1	Character	30		Species of 1st scientific name
GENUS2	Character	30		Genus of 2nd scientific name
SPECIES2	Character	30		Species of 2nd scientific name
SIM_G_2G	Numeric	12	10	Similarity index between genus for bigram
SIM_S_2G	Numeric	12	10	Similarity index between species for bigram
SIM_GS_2G	Numeric	12	10	Σ of genus similarity index for bigram
SIM_G_3G	Numeric	12	10	Similarity index between genus for trigram
SIM_S_3G	Numeric	12	10	Similarity index between species for trigram
SIM_GS_3G	Numeric	12	10	Σ of genus similarity index for trigram
SNDX_G1	Character	6		Soundex code for genus of 1st scientific name
SNDX_S1	Character	6		Soundex code for species of 1st scientific name
SNDX_G2	Character	6		Soundex code for genus of 2nd scientific name
SNDX_S2	Character	6		Soundex code for species of 2nd scientific name
PHX_G1	Character	15		Phonix code for genus of 1st scientific name
PHX_S1	Character	15		Phonix code for species of 1st scientific name
PHX_G2	Character	15		Phonix code for genus of 2nd scientific name
PHX_S2	Character	15		Phonix code for species of 2nd scientific name
SKL_G1	Character	20		Skeleton code for genus of 1st scientific name
SKL_S1	Character	20		Skeleton code for species of 1st scientific name
SKL_G2	Character	20		Skeleton code for genus of 2nd scientific name
SKL_S2	Character	20		Skeleton code for species of 2nd scientific name
LV_DIST	Numeric	2		Levenshtein distance between 1st scientific name and 2nd scientific name
G2	Logical	1		Bigram test flag
G3	Logical	1		Trigram test flag
SNDX	Logical	1		Soundex test flag
PHX	Logical	1		Phonix test flag
SKL	Logical	1		Skeleton-key test flag
LV	Logical	1		Levenshtein distance test flag
CNIP	Logical	1		North-east checklist database flag
PMA	Logical	1		Atlantic Rain Forest database flag
ILDIS	Logical	1		ILDIS database flag
SP2K	Logical	1		Species 2000 database flag
MARINE	Logical	1		Marine Invertebrates database flag
CTRL	Logical	1		Control database flag
ER	Character	5		Error type
POS	Character	1		Error position (*genus or species*)
LT	Character	10		The letters affected in the error

Table 21 – The TFeeder v.2 name-matches table structure

At this stage, the Levenshtein algorithm was added to the set of algorithms used on the tests. Since the results from all algorithms and all databases (working tables) were now stored in a single table, a new interface was built to accommodate the exigencies of this new version.

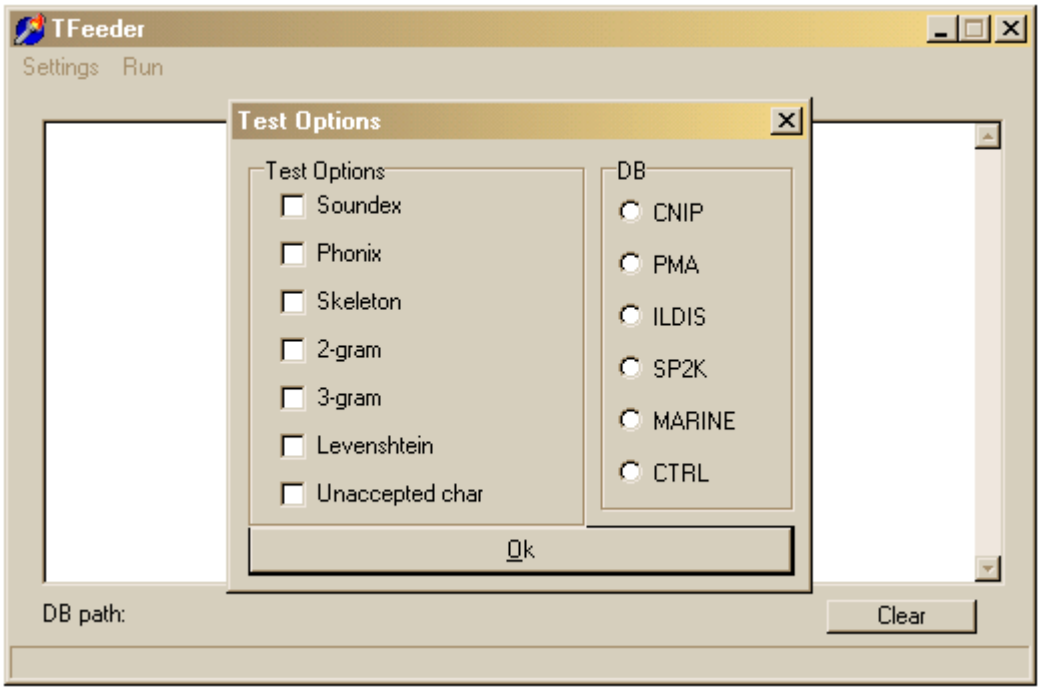


Figure 20 – *Tfeeder* v.2 interface

For algorithms based on character encoding (*Soundex*, *Phonix* and *Skeleton Key*), the *TFeeder* v.2 tool observes the following outline:

a) Select a scientific name (*genus* + *species*) from the given database and store the position

b) Calculate the code (*Soundex*, *Phonix* or *Skeleton*) for *genus* (*g1*) and *species* (*s1*)

c) Select the next scientific name

d) Calculate the code (*Soundex*, *Phonix* or *Skeleton*) for *genus* (*g2*) and *species* (*s2*)

e) If *g1* = *g2* and *s1* = *s2* then store the pair of scientific names in the name-matches table

f) Repeat steps *c*, *d* and *e* until the end of the given database

g) Return to the position previously stored

h) Select the next scientific name (*genus* + *species*) from the given database

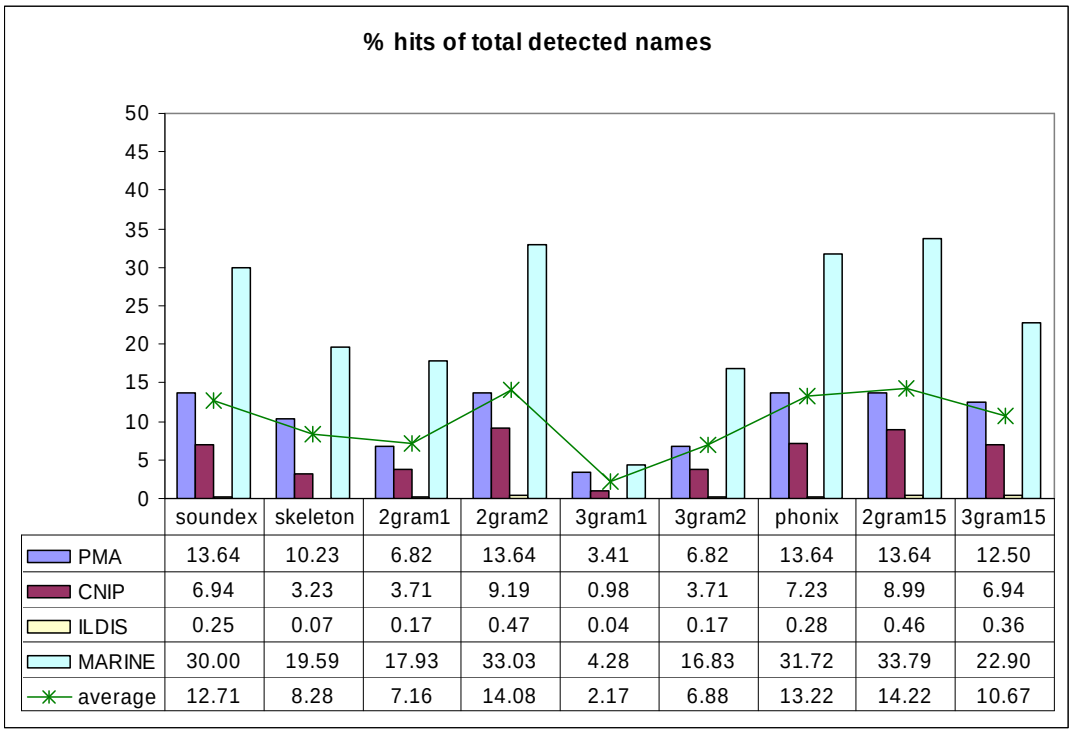
i) Repeat all the steps until the end of the given database

For plain string similarity distance (*n*-gram), the detection tool observes the following outline:

a) Select a scientific name (*genus* + *species*) from the given database and store the position
b) Select the next scientific name
c) Calculate the index of similarity between *genus* (*gis*) and *species* (*sis*)
d) If *gis* +*sis* ≥ 1.5 then store the pair of scientific names in the name-matches table
e) Repeat steps *b*, *c* and *d* until the end of the given database
f) Return to the position previously stored
g) Select the next scientific name (*genus* + *species*) from the given database
h) Repeat all the steps until the end of the given database

The limit of 1.5 as the minimum similarity (step *d* and corresponding to the *approach 1.5* cited on pg. 118) was established because it corresponds to the case where a minimum of 50% of the *n*-grams present in one of the name elements (*genus* or *species*) are shared by both names.

The choice of the approach *1.5* of the *n*-gram tests was reinforced by the preliminary analysis of the first round of tests, which showed, for example, that this approach had the best average percentage of *hits* of the total detected names (Figure 21).



	soundex	skeleton	2gram1	2gram2	3gram1	3gram2	phonix	2gram15	3gram15
PMA	13.64	10.23	6.82	13.64	3.41	6.82	13.64	13.64	12.50
CNIP	6.94	3.23	3.71	9.19	0.98	3.71	7.23	8.99	6.94
ILDIS	0.25	0.07	0.17	0.47	0.04	0.17	0.28	0.46	0.36
MARINE	30.00	19.59	17.93	33.03	4.28	16.83	31.72	33.79	22.90
average	12.71	8.28	7.16	14.08	2.17	6.88	13.22	14.22	10.67

Figure 21 – First round test analysis: number of real errors (*hits)* as a percentage of the total numbers of returned name pairs, by algorithm

For the *Levenshtein distance*, the detection tool observes the following outline:

a) Select a scientific name (*genus* + *species*) from the given database and store the position

b) Select the next scientific name

c) Calculate the distance between the corresponding *genus* (*gd*) and *species* (*sd*) parts of the two names

d) If $gd + sd \leq 4$ then store the pair of scientific names and the values of *gd* + *gs* in the name-matches table

e) Repeat steps *b*, *c* and *d* until the end of the given database

f) Return to the position previously stored

g) Select the next scientific name (*genus* + *species*) from the given database

h) Repeat all the steps until the end of the given database

The limit of 4 as the maximum distance between the scientific names was defined considering that Damerau (1964) found that approximately 80% of all misspelled words contained a single instance of one of the following four error types: insertion, deletion, substitution, and transposition. A Levenshtein distance of 4 would occur, for example, if there was a transposition in both the genus name and specific epithet.

Spelling error classification tool

The new version of the *ErrorId* tool was developed to scan the new name-matches table, fed by the TFeeder v.2, and to classify each pair of names using the same combination of automatic and manual methods into the same categories of errors as the first version and following the same set of rules (see pg. 120), except for the absence of the "*Doubt*" category as a possible option at the interactive classification interface, excluded as a classification option for the test because it was not represented in the four possible outcomes in error detection tasks (see Table 17, pg. 113), pointed out by Klein (2001) and adopted in this thesis as a standard for the tests

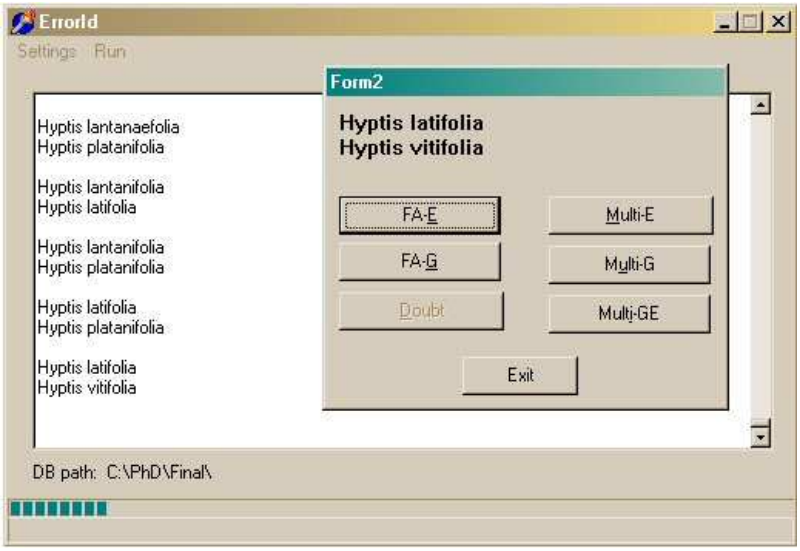


Figure 22 – Interactive interface of the error classification tool

The type of error for each pair of names was stored in the appropriate field in the name-matches table for later analysis.

The error classification tool was also designed to identify in which part of the name the error occurred and the position of the error (genus, species or both), and these were stored in the proper field in the name-matches table.

The name-matches table *test2.dbf* that was populated by the spelling error detection and classification tools described above contains the results which are analyzed below. This table is available from the author on request.

A set of programs was developed in the xBASE language using Microsoft Visual FoxPro 7.0 © in order to analyze the *tests2.dbf* name-matches table and produce a set of reports. These ASCII File reports (Appendix III) were used as input to MS-EXCEL© which was used to compile the final analysis of the tests and produce figures and graphics.

The analysis of the results of the experiments will be presented in a graphical form, followed by appropriate comments.

7.6. Results

7.6.1. Preliminary analysis of name-matches table

Spelling errors

After the tests, the name-matches table contained 51,645 pairs of names where 998 pairs of names were classified as “hits” and 50,647 pairs of names was classified as “false alarms”, which represents 98.1% of all pairs of names detected by the tests (Figure 23).

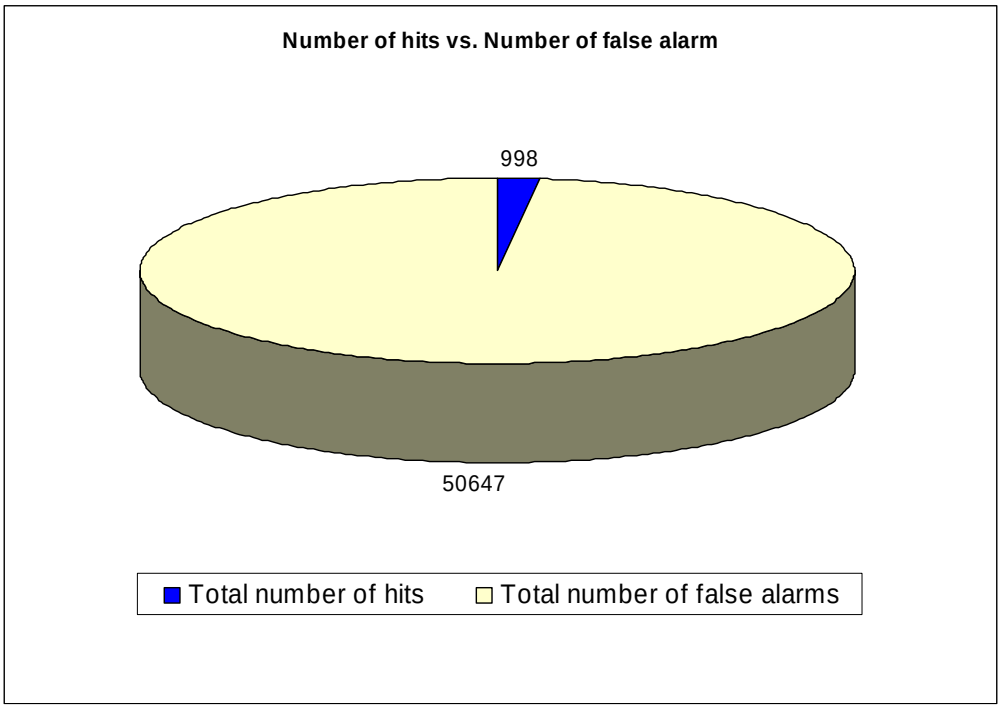


Figure 23 – Total number of “hits” vs. total number of “false alarms” in the name-matches table after the tests

Figure 24 shows the distribution of the types of errors in the total number of “hits”.

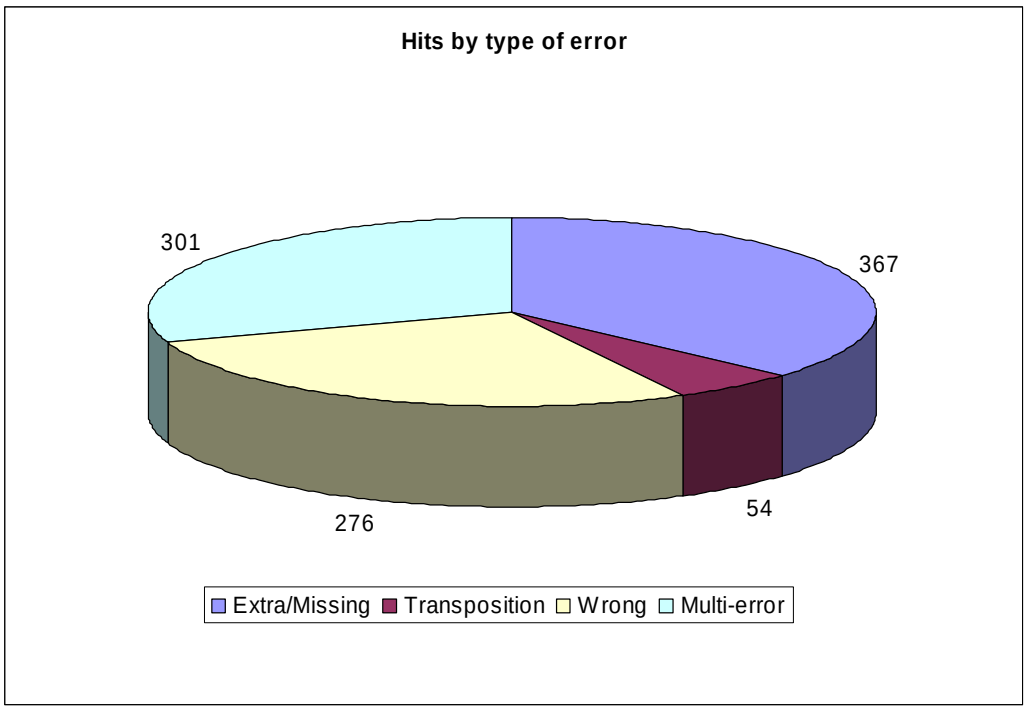


Figure 24 – Number of “hits” classified by type of error

7.6.2. Error rates in databases

All the five tested databases showed possible spelling errors in scientific names, as shown in Figure 25 and Figure 26, which include all the “hits” detected by one or more of the algorithms. The Marine Invertebrates of New Zealand Database showed a particularly large percentage of wrong names, explained by the process of construction of the database, from free text with no data quality control at data entry.

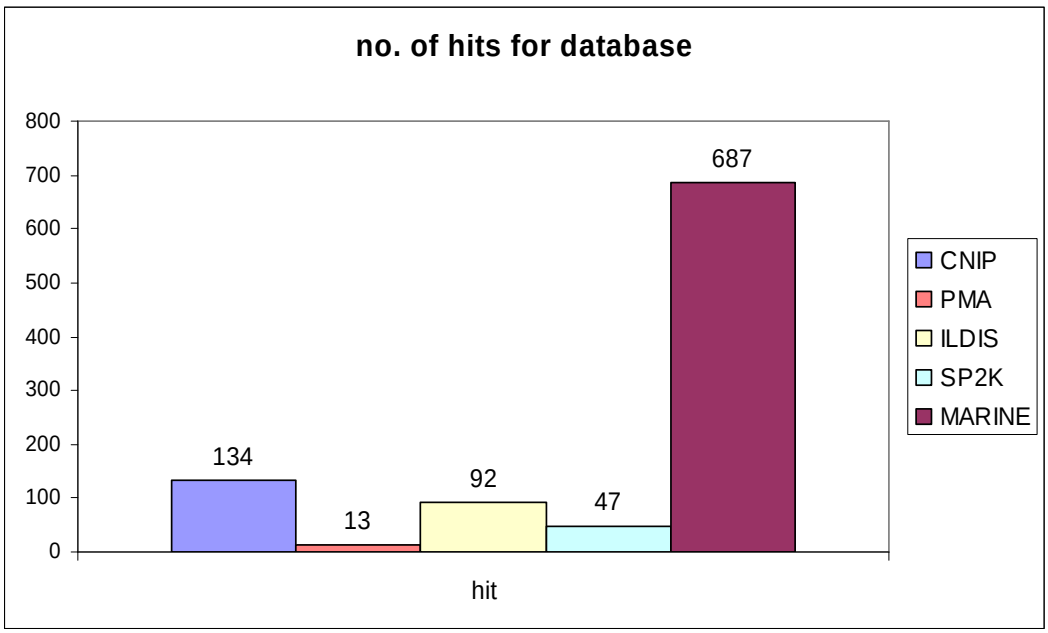


Figure 25 – No. of "hits" for the test databases

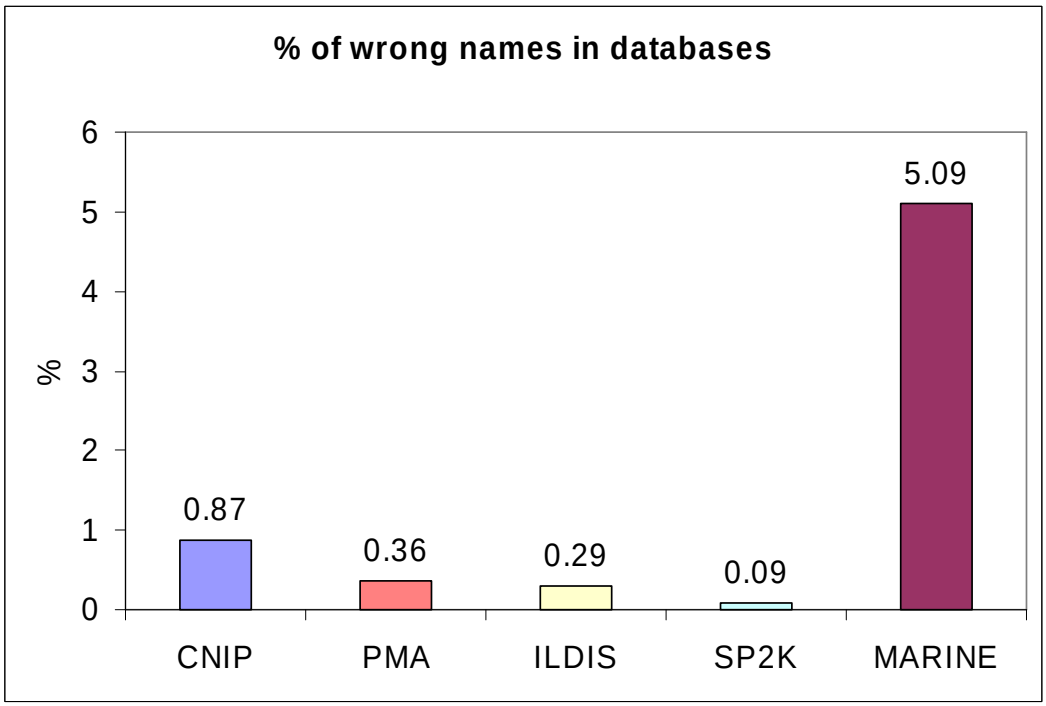


Figure 26 - % of wrong names in the test databases [23]

An unexpectedly large number of false alarms was detected in the ILDIS database (Figure 27) which can be explained by the large number of scientific names which

[23] Assuming one name of the detected pair is correct and the other name is a misspelling error

belong to each genus, in other words, it is a database with a restricted taxonomic domain (Figure 28), increasing the average similarity of names.

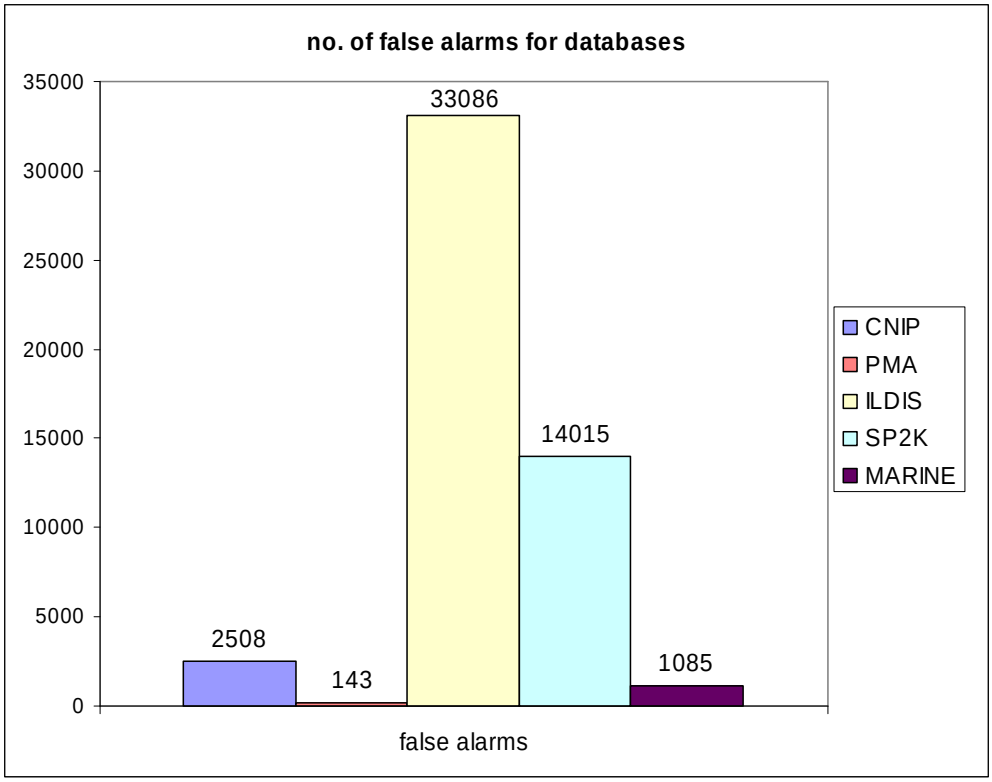


Figure 27 – Number of "false alarms" for the test databases

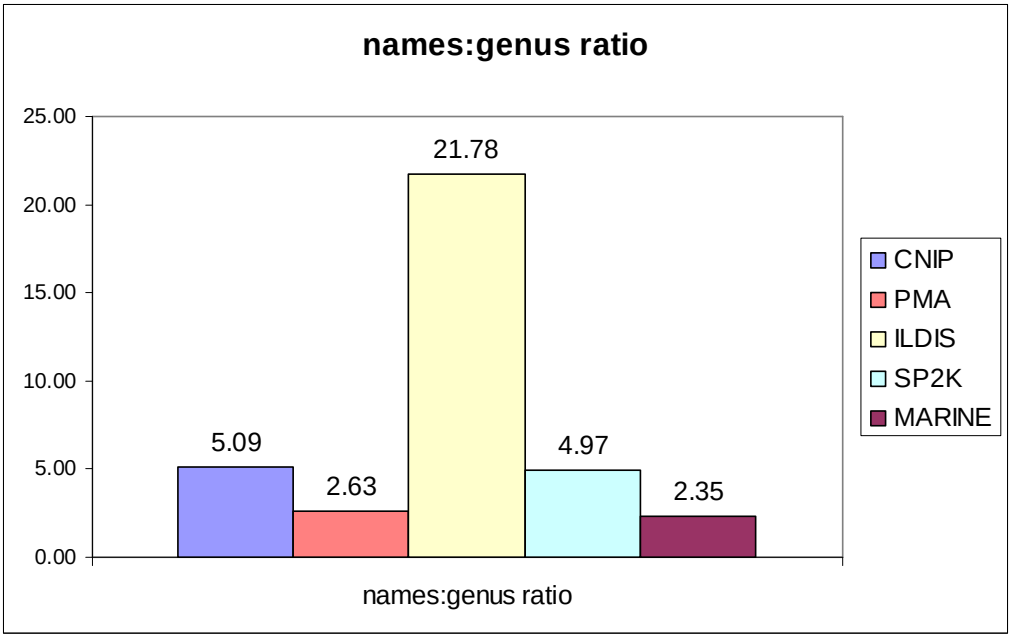


Figure 28 – Ratio of the number of scientific names per genus for the test databases

7.6.3. Performance of the algorithms

Figure 29 shows the number of “hits” detected by each algorithm tested. The name-matches table contained a total of 998 pairs of scientific names that were classified as “hits”.

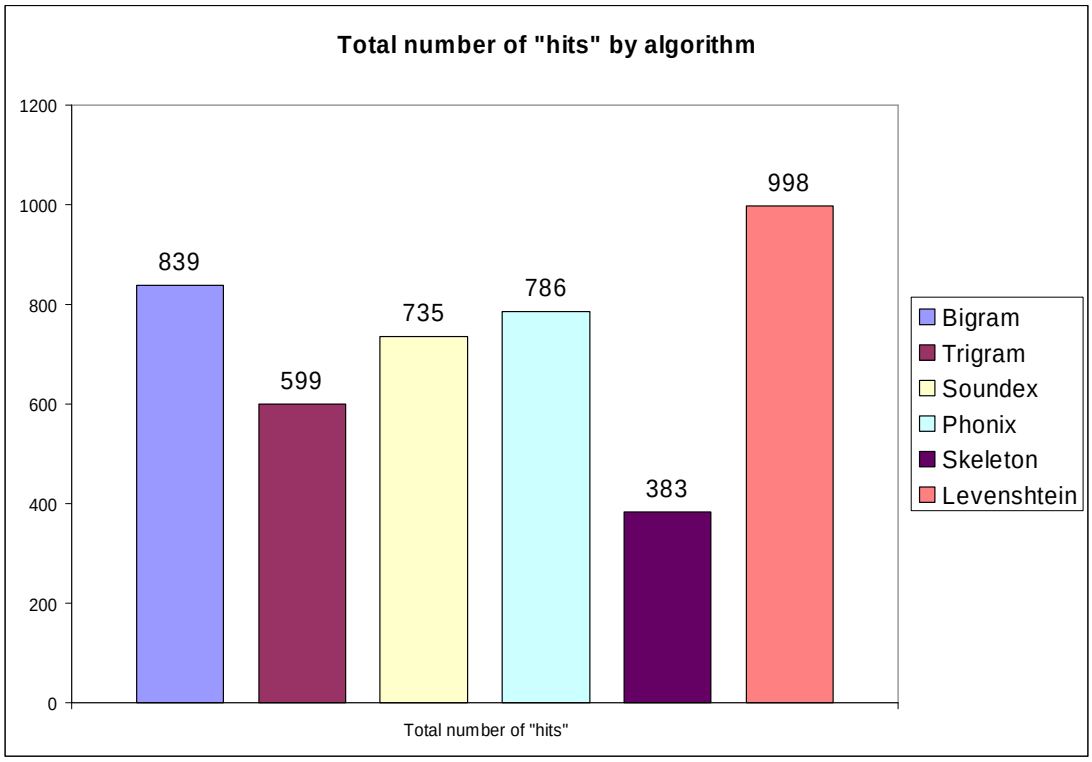


Figure 29 – Total number of “hits” detected by each algorithm

Figure 30 shows the number of “false alarms” detected by each algorithm tested. The name-matches table contained a total of 50,647 pairs of scientific names that were classified as “false alarms”.

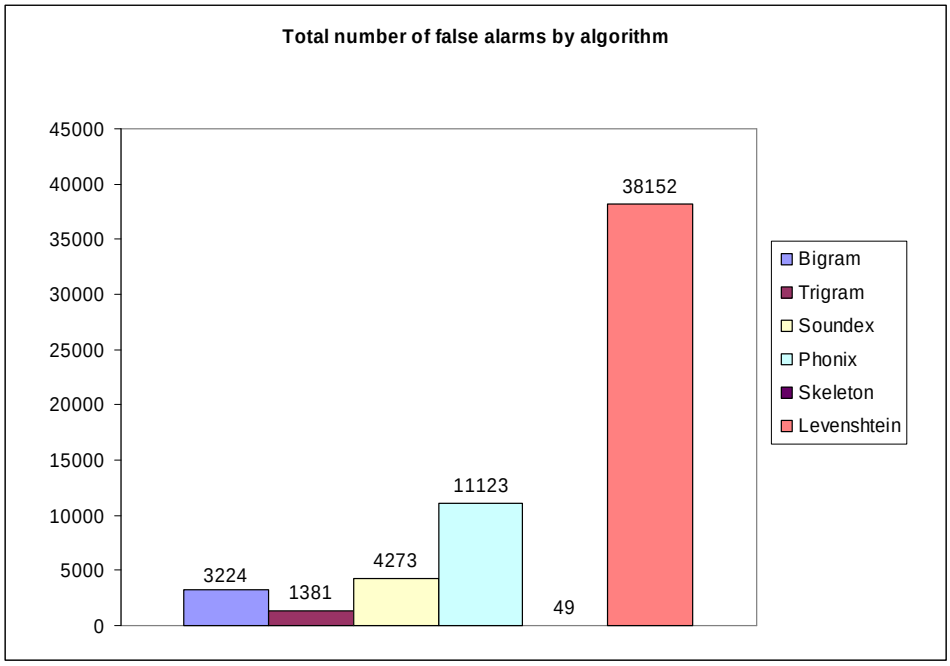


Figure 30 – Total number of false alarms detected by each algorithm

Figure 31 shows the relationship between the percentages of "hits" and "false alarms" and indicates that the *Levenshtein* algorithm was the only one to detect all pairs of names considered as a "hit".

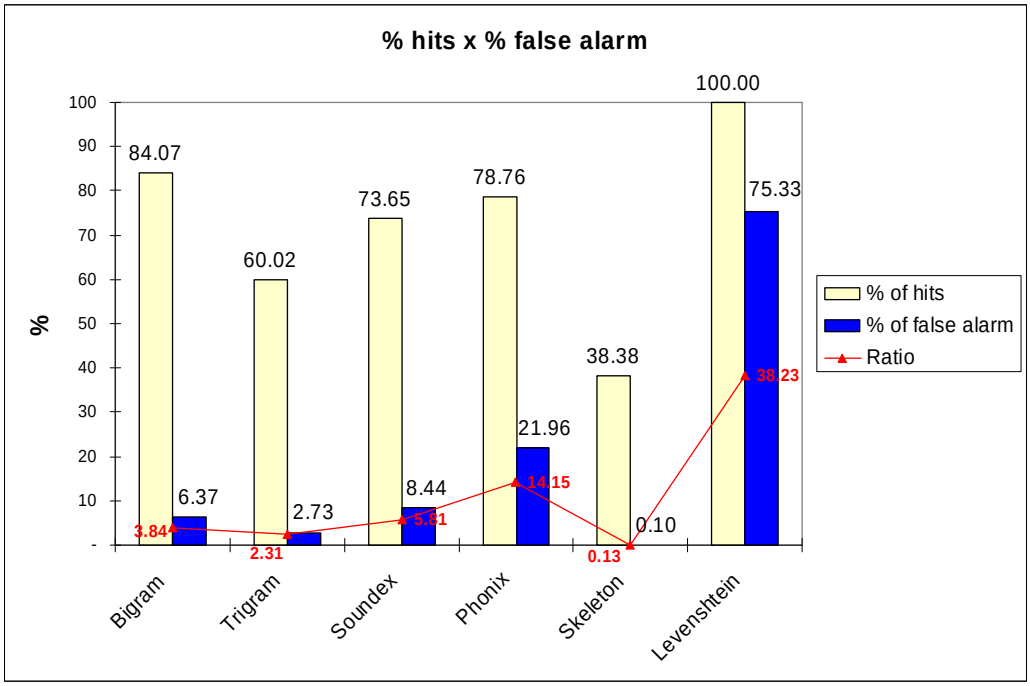


Figure 31 – Relationship between % of hits and % of false alarms by algorithm

However, the *Levenshtein* algorithm presents a high rate of "false alarms", and was responsible for 68.35% of the "exclusive false alarms" (which were detected only by one of the algorithms, in this case the *Levenshtein* algorithm), as showed in Figure 32.

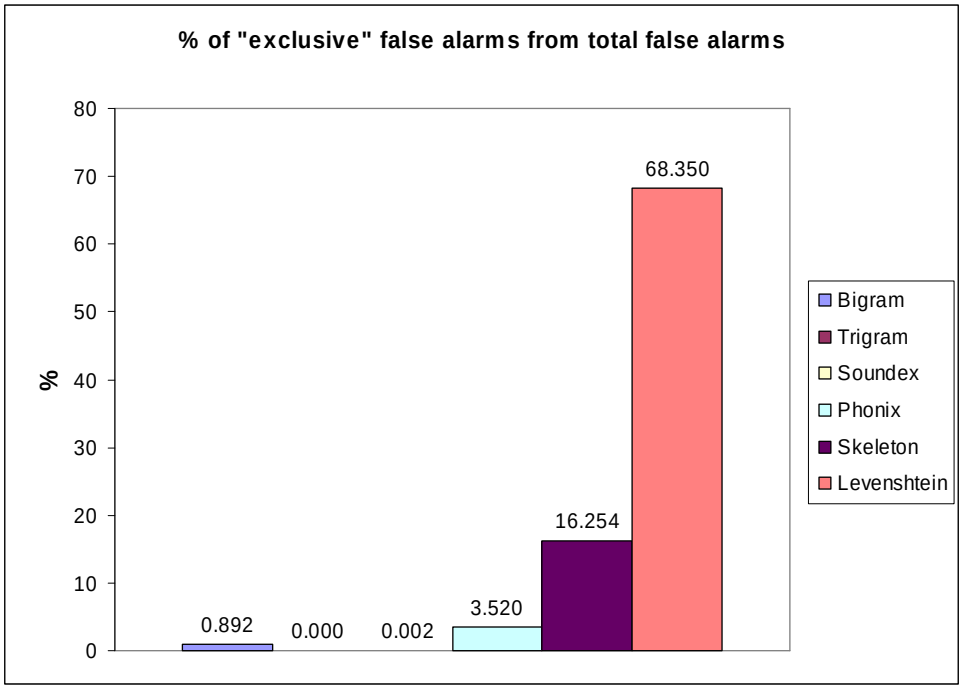


Figure 32 – Percentage of "exclusive" false alarms by algorithm

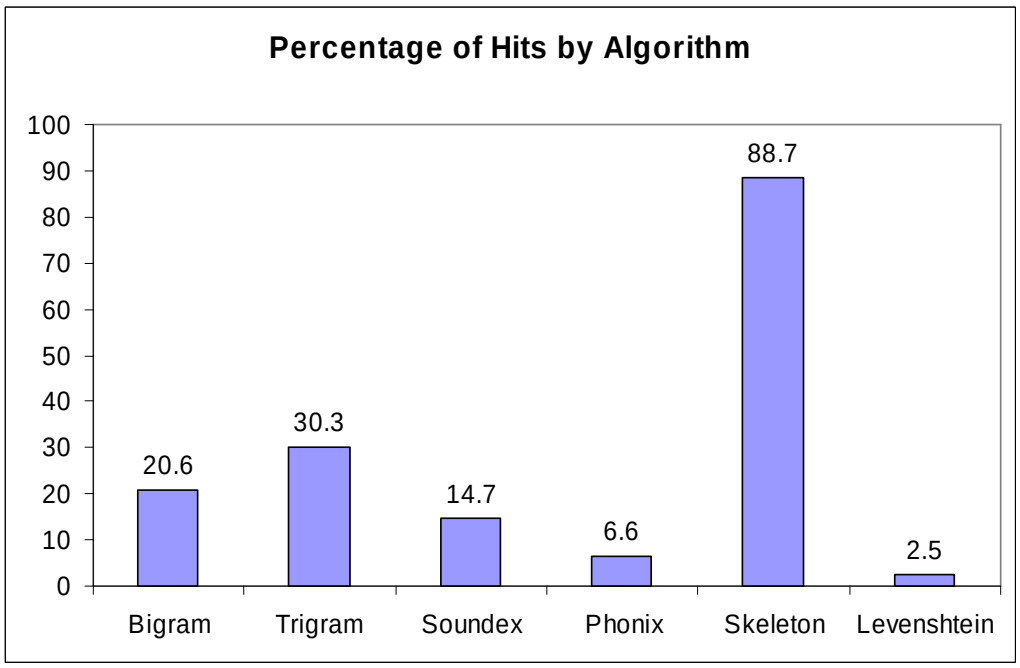


Figure 33 – Percentage of detected pairs which are "hits", by algorithm

Despite the high probability of detecting a hit shown by the Skeleton Key algorithm (Figure 33), its efficiency is the worst, detecting only 38.4% of the hits (Figure 31).

The Levenshtein algorithm test shows that the efficiency of the algorithm, in terms of the number of errors detected, has a strong relationship with the value of distance.

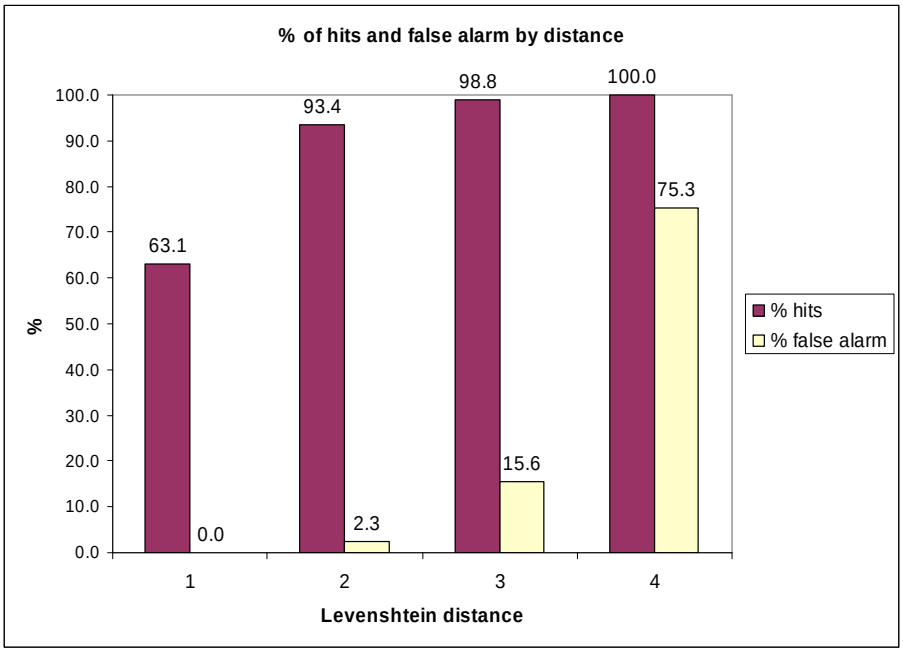


Figure 34 – Relationship between the Levenshtein distance and the efficiency of the algorithm

Figure 34 shows that when the value of the Levenshtein distance equals 1, there are no false alarms but only 63.1% of the total of hits were caught in this configuration. Configured to a Levenshtein distance equal to 2, the Levenshtein algorithm was able to catch 93.4% of all hits, with the rate of 2.3% of returned false alarms. With the distance set to 3, the algorithm caught 98.8% of the total hits but returns 15.6% of the false alarms. At distance 4, the number of false alarms jumps to 75.3% while the number of hits rises by a mere 1.2%. The figures suggest that the Levenshtein algorithm, configured to the distance value of 4, associated with a efficient algorithm to automatically detect and filter the false alarm behaviour, can be used with 100%

efficiency to detect spelling errors of scientific names, at least under the conditions tested.

7.6.4. Precision and Recall

The discovery of spelling errors in a database is analogous to the retrieval of desired records or documents from a database by means of specified queries or search terms. Retrieval system evaluation plays an important role in judging the efficiency and effectiveness of the retrieval process. Several different evaluation criteria, deemed most critical to the user population, have been identified in retrieval research; namely, recall, precision, effort, time, form of presentation and coverage. Among them, recall and precision have received the most attention in the literature (Raghavan *et al.*, 1989).

Recall is the proportion of the relevant documents which are retrieved, defined as the ratio of the number of relevant documents that are retrieved to the total number of relevant documents. Precision is the proportion of the documents retrieved which are relevant, defined as the number of relevant documents retrieved divided by the number of retrieved documents. In particular, a recall-precision graph is often used as a combined evaluation measure of retrieval systems, showing recall and precision on the *x* and *y* axes. Such a graph, given an arbitrary recall point, can tell us the corresponding precision value (Raghavan *et al.*, 1989).

For the purpose of evaluating the performance of the algorithms, we adapt the original definitions of *Precision* and *Recall* to be used in these tests as follows:

Considering *Precision* as:

$$\frac{\text{Number of pairs of "hits" retrieved by the algorithm}}{\text{Total number of pairs retrieved by the algorithm}}$$

And, considering *Recall* as:

$$\frac{\text{Number of pairs of "hits" retrieved by the algorithm}}{\text{Total number of pairs of "hits" in the database}}$$

The performance of the algorithms can be summarized in Figure 35.

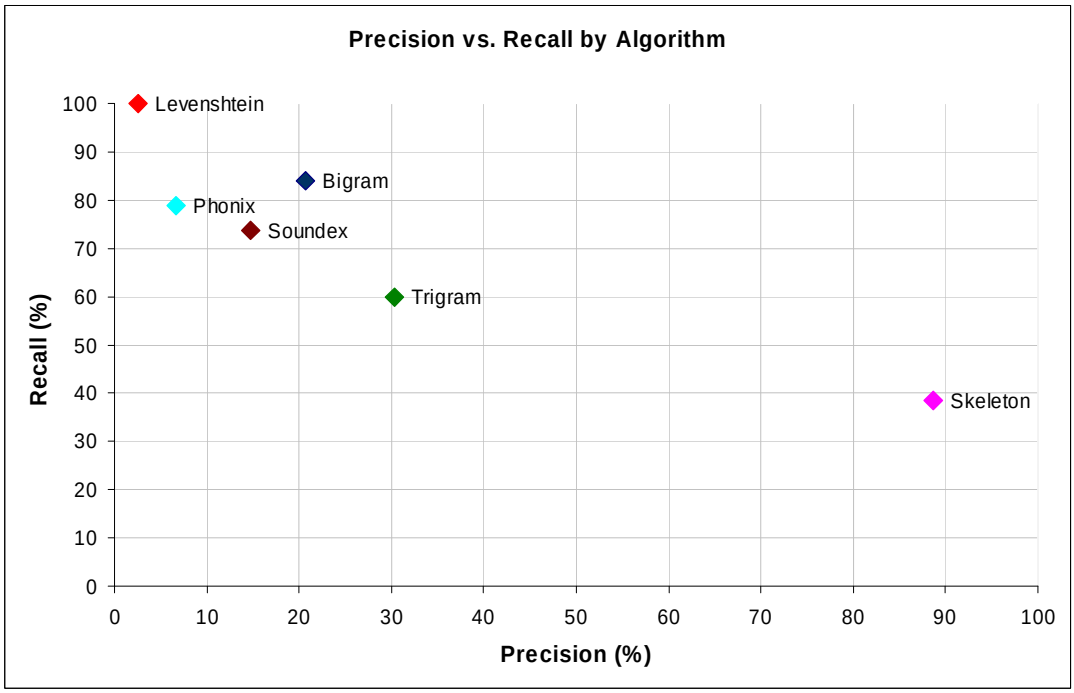


Figure 35 – Summary of the algorithms' performance using *Precision vs. Recall*

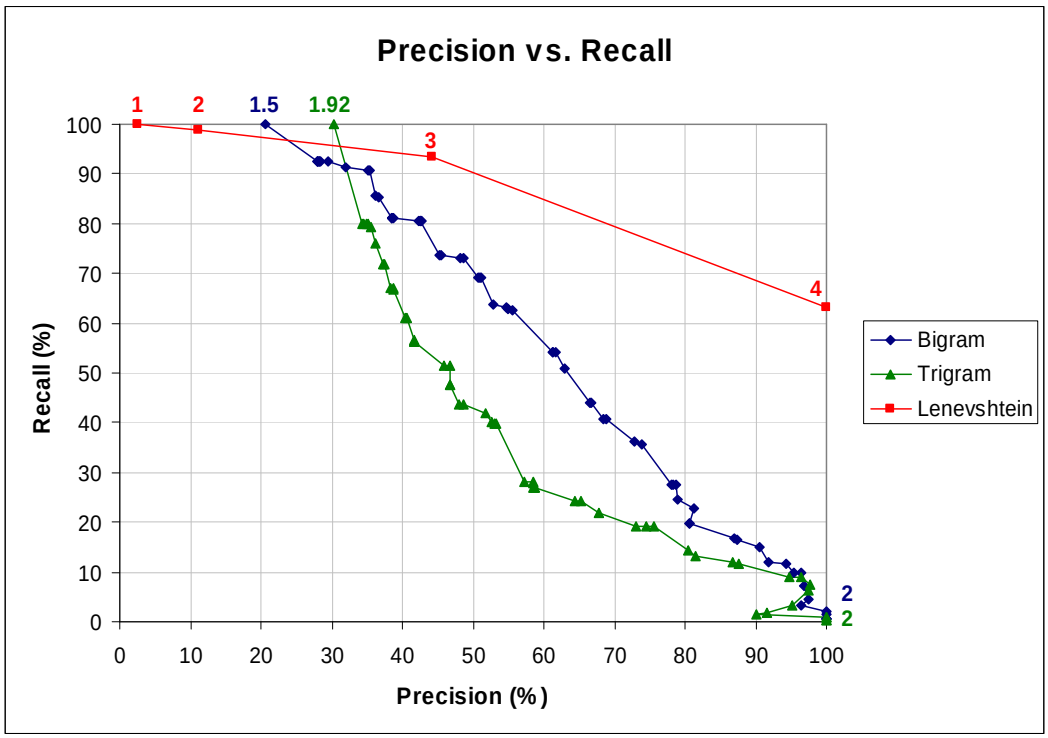


Figure 36 – Variation of *Precision* and *Recall* by similarity

In Figure 36, we can see the variation of performance as the variation of similarity index of *bigram* and *trigram* algorithm, and the variation of the distance of the *Levenshtein* algorithm.

7.6.5. Discussion of the results

As discussed at the beginning of this chapter, general-purpose spelling error detection techniques (as used in spelling checker tools) are usually based on checking for the existence of a given word in a “reliable source”, and efficient techniques have been devised for detecting strings that do not appear in a given word list, dictionary or lexicon (Kukich, 1992). Spelling correction techniques and tools go further: they both detect misspelled words and try to find the most likely correct word. This problem contains elements of pattern recognition and coding theory (Peterson, 1980).

The approach used in the tests in this thesis represents a mix of detection techniques, using the database or names list itself as a data dictionary; and correction techniques, using similarity algorithms to detect scientific names with a degree of similarity which suggest a spelling error.

The lack of dependence on a “reliable source” confers to the approach the advantage of an independent process, which can be executed by stand-alone tools without a dependence on external or third party data. On the other hand, this reduces the information available to help correct the errors detected.

The significant numbers of “false alarms” returned by the tests, associated with the algorithms that show the higher number of “hits” (Figure 31), suggest the necessity of auxiliary processes or tools to minimize the need for “human interaction” with the returned names lists, in order to distinguish the genuine “hits”.

Since 13.9% of false alarm behaviour was generated by differences in the generic epithet, the association of these techniques with a reliable lexicon, such as a reference *Genus* dictionary for example, could minimize the "false alarm" behaviour, and thus enhance the efficiency of these tools in practical use.

In the year 2000 the International Plant Name Index (www.ipni.org) launched its website with public access to its database, with "live access" to the database implemented in 2004. Using the URL encoding method (http://www.ipni.org/link_to_ipni.html) a list of compete scientific names can be checked for spelling errors against this reliable reference data dictionary. In order to evaluate the access to the IPNI database using a URL encoding system in a DBMS environment, an application was developed. The *ATF -» IPNI Check* tool was developed for the same environment described in section 7.5.1 in this thesis. The function of this tool was to check the scientific names present in a *sam.tdi*[24] table against the IPNI database, using the URL encoding technique.

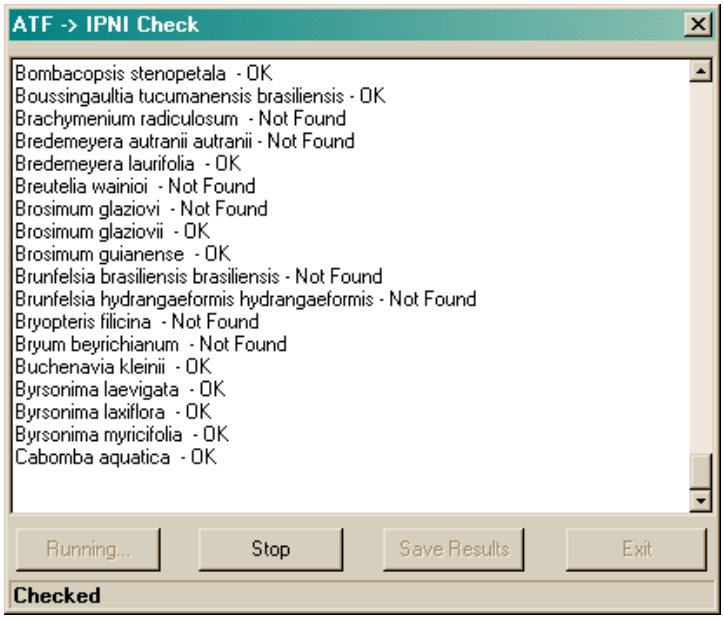


Figure 37 - ATF -» IPNI Check Interface

[24] *sam.tdi* table is compatible with the Alice Transfer Format, cited in section 7.3.

As can be seen in Figure 37, the Atlantic Rain Forest database was checked and, for example, the tool was able to define the correct name between a pair of scientific names detected by the spelling error test (*Brosimum glaziovi* and *Brosimum glaziovii*). The application of techniques such as this one to automatically correct spelling errors will be discussed in chapter 9 in this thesis.

The significant numbers of "false alarms" returned by the tests also suggest the necessity of development of a specific string similarity algorithm which can handle with better efficiency the particular characteristics of the Latin language, specifically for scientific nomenclature. This specific algorithm can be based on the improvement of phonetic similarity algorithms which already exist, adapted to handle Latin language. However, the association of different and improved algorithms, as a mix of phonetic and string similarity algorithms, as the experiments suggest, seems to be the best approach.

The *error classification tool* was capable of classifying automatically 92.36% of the returned pairs of names ("hits" or "false alarms"), whereas 7.64% of the returned pairs of names were classified interactively. However, this tool was not developed with the objective of determining automatically which one of the pair was correct. The definition of the "*doubt*" classification for a pair of names was set up at the interactive interface, in the first round of tests, to deal with a "real world" problem: a database administrator facing two very similarly spelled scientific names and, without appropriate skill and any auxiliary reference, needing to judge whether they may represent different taxa. Therefore, although this cannot be considered an integral part of a spelling error detection algorithm, it should be considered in any "real world" application.

Despite the building techniques and database administration efforts and tools (Table 22), all five databases tested show spelling errors. However, databases built with no data entry control, such as the Marine Invertebrates database, show the highest number of spelling errors (Figure 26).

Database	Database building and management technique
Atlantic Rain Forest	Alice Software
North-east checklist	Alice Software
ILDIS	Alice Software
Marine Invertebrates	Free Text
Species 2000	Compilation of different datasets

Table 22 – Database build/management technique

Algorithms based on character encoding, such as *Soundex*, *Phonix*, and *Skeleton Key*, have substantial advantages in execution performance since the codes can be generated in advance for each name, an "order (n)" process whose execution time is proportional to the number of names. *Bigram*, *Trigram* and *Levenshtein Distance* algorithms depend on comparison of the original strings, an "order (n^2)" process. Hence, the execution of these algorithms presented the worst performance.

Taking the results as a whole, and excluding the control database, there was a significant number of "hits" at species epithet (785), followed by "hits" on *Genus* (160) and "hits" in both name components (28), this last case found only in the Marine database.

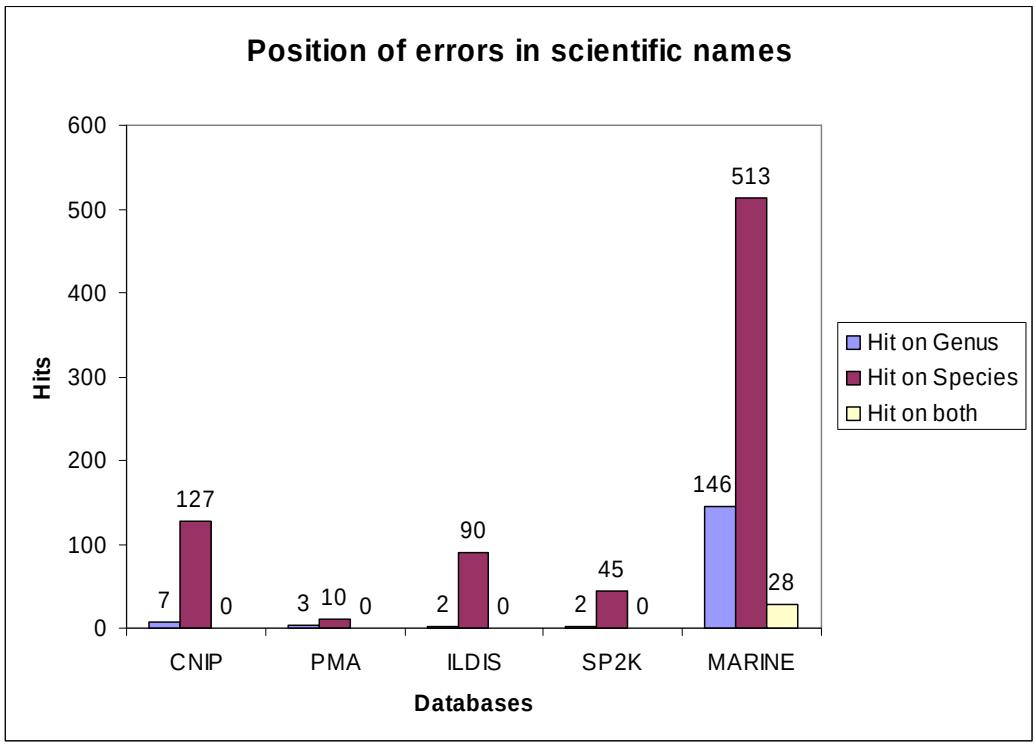


Figure 38 – Position of errors in the scientific names

The low rate for errors in the genus name for the first three databases, shown in Figure 38, can be explained by the fact that those databases were managed by a specialized relational database management system (Alice Software ®) which offers a genus dictionary from which the users may choose to select existing genus names. The low rate for genus errors in the Species 2000 database suggests the adoption of data entry control policies and data control during the compilation process. The high rate of errors in the genus position in the Marine database confirms the lack of a control policy at data entry and the basic nature of the Marine database management system.

The analysis of the frequency of the “wrong letter” error type produced the following Table 23:

Wrong letter	frequency
a «-» i	29
a «-» o	18
i «-» y	16
a «-» e	15
i «-» u	12
n «-» r	10
a «-» u	9
e «-» i	9
e «-» o	8
i «-» o	7
m «-» s	7
m «-» n	6
r «-» t	6
a «-» s	5
c «-» e	5
g «-» q	5
h «-» n	5
n «-» u	5
l «-» t	4
c «-» n	3
c «-» s	3
i «-» l	3
o «-» u	3
u «-» y	3
a «-» c	2
b «-» g	2
b «-» t	2
c «-» g	2
c «-» p	2
d «-» t	2
e «-» l	2
e «-» u	2
f «-» t	2
i «-» j	2
i «-» t	2
r «-» v	2
s «-» z	2
a «-» h	1
A «-» L	1
a «-» m	1
a «-» q	1
a «-» y	1
a «-» z	1
b «-» h	1
b «-» k	1
b «-» l	1
b «-» n	1
b «-» p	1
b «-» r	1
b «-» s	1
b «-» v	1
C «-» D	1
C «-» G	1
c «-» m	1
c «-» q	1
c «-» r	1
c «-» v	1
d «-» g	1
d «-» h	1
d «-» j	1
d «-» q	1
d «-» r	1
d «-» s	1
e «-» f	1
e «-» s	1
e «-» t	1
f «-» v	1
h «-» i	1
H «-» K	1
h «-» l	1
l «-» J	1
i «-» n	1
i «-» r	1
i «-» s	1
j «-» k	1
j «-» t	1
k «-» l	1
k «-» s	1
l «-» m	1
l «-»r	1
l «-» v	1
m «-» u	1
n «-» p	1
o «-» r	1
p «-» r	1
p «-» s	1
r «-» s	1
s «-» t	1
t «-» v	1
t «-» y	1
v «-» y	1

Table 23 – The frequency of the "wrong letter" error type analysis

The high occurrence of the replacement of *a* for *i* was, in the majority of cases (24 out of 29 pairs of names), because of the difference between "micr-" and "macr-" at the beginning of the specific epithet (Table 24).

The names in Table 24 were considered as "hits" in the analysis. However, they could be considered as "false hits" since the names have characteristics of a "hit" (similarity and classification as a wrong letter) but a strong likelihood to be two valid names and not a spelling error at all. Again, as cited at the section 7.6.3, a reliable reference, such as IPNI, can be use to minimize the effect of the "false hit".

Jacaranda macrantha Jacaranda micrantha
Bauhinia macrostachya Bauhinia microstachya
Calliandra macrocalyx Calliandra microcalyx
Lantana macrophylla Lantana microphylla
Argyrolobium macrophyllum Argyrolobium microphyllum
Crotalaria macrocarpa Crotalaria microcarpa
Detarium macrocarpum Detarium microcarpum
Diphysa macrophylla Diphysa microphylla
Indigofera macrantha Indigofera micrantha
Indigofera macrocalyx Indigofera microcalyx
Kennedia macrophylla Kennedia microphylla
Machaerium macrophyllum Machaerium microphyllum
Mimosa macrocephala Mimosa microcephala
Rhynchosia macrantha Rhynchosia micrantha
Sclerolobium macropetalum Sclerolobium micropetalum
Swartzia macrocarpa Swartzia microcarpa

Tephrosia macrantha Tephrosia micrantha
Trifolium macrocephalum Trifolium microcephalum
Astronesthes macropogon Astronesthes micropogon
Bathytroctes macrolepis Bathytroctes microlepis
Curimatopsis macrolepis Curimatopsis microlepis
Imparfinis macrocephala Imparfinis microcephala
Moringua macrochir Moringua microchir
Rhynchactis macrothrix Rhynchactis microthrix
Sertella fragida Sertella frigida
Resania lanceolata Resinia lanceolata
Paramoera rangatira Paramoera rangitira
Pagurus spanulimanus Pagurus spinulimanus
Chlamys kiwaensis Chlamys kiwiensis

Table 24 – Replacement of “a” for “i” in wrong letter error type

8. Taxonomic database quality assessment and metrics

8.1. Current concepts in data quality assessment

For English (1999), there are two basic categories of information quality characteristics and their related measures. The first is the inherent quality of the data. Inherent information quality characteristics are those that are independent of the way data are used. In a database, data have certain *static* quality characteristics. The second category consists of pragmatic quality characteristics. These include how intuitive the information is in its present format and how well it enables knowledge workers to accomplish their objectives. Pipino, Lee, and Wang (2002) present the same classification as English, calling these two basic characteristics *task-independent* and *task-dependent*, respectively.

For Kahn *et al*. (2002), information quality is an inexact science in terms of assessment and benchmarks. Although various aspects of quality and information have been investigated, there is still a critical need for methodologies that assess how well organizations develop information products and deliver information services to consumers.

However, for the inherent information quality, the focus of this thesis, assessments can be associated with the data quality dimensions related to inherent aspects of data, such as accuracy, completeness, consistency and timeliness.

The task-independent (or inherent) metrics reflect states of the data without the contextual knowledge of the application, and can be applied to any data set, regardless of the tasks at hand (Pipino *et al.*, 2002).

Pipino *et al.* (2002) presents a *simple ratio* measure, as a traditional data quality metric for quality dimensions such as *accuracy*, *consistency* and *completeness*.

The *simple ratio* measures the ratio of the number of desired outcomes to the total number of outcomes. Since most people measure exceptions, however, a preferred form is the number of undesirable outcomes divided by the total number of outcomes subtracted from 1. This simple ratio adheres to the convention that 1 represents the most desirable and 0 the least desirable score.

For accuracy, if one is counting the data units which are in error, the metric is defined as the number of data units in error divided by the total number of data units, subtracted from 1. In practice, determining what constitutes a data unit and what is an error requires a set of clearly defined criteria. For example, the degree of precision must be specified. It is possible for an incorrect character in a text string to be tolerable in one circumstance but not in another. For example, characters such as "." might be acceptable in the context of phytosociological data sets (e.g. *Vicia sp.*) and not in nomenclatural databases.

The *completeness* dimension can be viewed from many perspectives, leading to different metrics. At the most abstract level, one can define the concept of schema completeness, which is the degree to which entities and attributes are not missing from

the schema. At the data level, one can define column completeness as a function of the missing values in a column of a table. A third type is called population completeness. If a column should contain at least one occurrence of all 50 US states, for example, but only contains 43 states, then we have population incompleteness. Each of these three types (schema, column and population completeness) can be measured by taking the ratio of the number of missing items to the total number of items required (i.e. not just counting the items which are present) and subtracting from 1.

Motro and Rakov (1996; 1998) approached data quality assessment considering *integrity* as a combination of *accuracy* and *completeness*. In their approach, a database view is *accurate* if it includes ***only*** information that occurs in the real world; a database view is *complete* if it includes ***all*** the information that occurs in the real world. Hence, a database view has integrity, if it includes the whole truth (*completeness*) and nothing but the truth (*accuracy*).

8.2. Applying the concept of data quality assessment to the taxonomic database domain

In order to demonstrate the application of these data quality assessment concepts, we will investigate their application to the taxonomic database domain. We denote the *actual* (stored) database instance *D*, and we denote the *ideal* (real world) database instance *W*. Of course, *W* is a hypothetical instance which is unavailable. The stored instance *D* is an approximation of the ideal instance *W*.

Therefore, a measure of the quality of a given database is the answer to the question "*how good is the approximation?*"

To determine the goodness of the approximation we must measure the similarity of the two database instances. Since each instance is a set (of tuples), we can use measures that compare two sets of elements. One such measure is based on two components as follows:

- Accuracy of the database relative to the real world: $\frac{|D \cap W|}{|D|}$
- Completeness of the database relative to the real world: $\frac{|D \cap W|}{|W|}$

As both *D* and *W* may be very large, and *W* may be unknown or unavailable, the estimation of *accuracy* and *completeness* should be based on samples of *D* and *W*.

To determine the portion of the stored information that is true (*accuracy*), the following procedure can be used:

1) Sample *D* to produce *Dsample*.
2) For each $x \in Dsample$, determine whether $x \in W$.
3) Calculate the *accuracy* estimate as the number of "hits" $|x \in W|$ divided by the sample size $|Dsample|$.

Step 1, sampling a database, is simple. But step 2 is difficult: we need to determine the presence of database elements in *W* when *W* is unavailable and without actually constructing it. This is accomplished either by human verification (*knowledge workers*)

of the sample elements or by, comparing with a reputable data source (*database bashing*).

To determine the portion of the true (real-world) information that is stored (*completeness*), the following procedure can be used:

1) Sample *W* to produce *Dsample*.
2) For each $x \in Wsample$, determine whether $x \in D$.
3) Calculate the *completeness* estimate as the number of "hits" $|x \in D|$ divided by the sample size $|Wsample|$.

Step 2, verifying the presence of elements in the database, is simple. But step 1 is difficult: we need to sample *W* when *W* is unavailable and without actually constructing it. However, a relation between $|Wsample|$ and the *taxonomic database domains* (see Section 6.3.2 - Database domain exception) can be established. Considering the example cited in the Section 6.3.2 - Database domain exception: a hypothetical database about "Woody Leguminosae of Brazil":

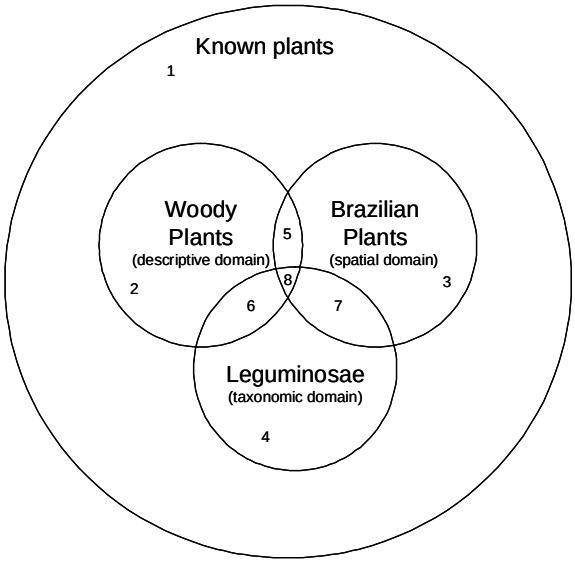


Figure 39 - Hypothetical database representation

As demonstrated, W_8 (the *ideal instance* at the region 8 of the figure above) is the intersection of the three sets: woody plants, Brazilian plants and Leguminosae plants. Thus, taxonomic database domains can be a reasonable way to determine $|Wsample|$ since, for some specific situations such as taxonomic revisions or regional floras, some stored databases *D* can play a reliable $|Wsample|$ role in a *database bashing* routine.

This approach is particularly useful for high hierarchical ranks of taxonomic nomenclature. Considering the example above, it should be straightforward for *knowledge workers* to produce a list of all known Leguminosae genera which belong to Brazil and, separately, are woody[25]; their intersection will determine the population completeness of the genus attribute of W_8.

For the same reason, databases which adopt as standard other hierarchically organized sets of values to describe specific attributes, such as the "Economic Botany Data Collection Standard" (Cook, 1995) and "World Geographical Scheme for Recording

[25] "Woody" means here "trees or shrubs", i.e. not herbs.

Plant Distributions" (Hollis and Brummitt, 1992), can use the same approach in order to assess the data quality level of these attributes.

These techniques of data quality assessment can be applied in the context of the *taxonomic data domains*, bearing in mind the peculiar characteristics of the *classification data domain* which may have several "different instances of the real world", according to the different opinions of taxonomists.

8.3. Taxonomic data quality assessment in practice

8.3.1. The ITIS Project

Institutional and cooperative taxonomic database projects have, in recent years, begun to take account of data quality issues and the importance of the establishment of data quality indicators. One example of the definition and utilization of taxonomic data quality indicators came from the ITIS Project (ITIS - Integrated Taxonomic Information System, 2004).

The Integrated Taxonomic Information System (ITIS) was established in the mid-1990's as a cooperative project among several US federal agencies to improve and expand upon taxonomic data (known as the "NODC Taxonomic Code") maintained by the National Oceanographic Data Center (NODC), National Oceanic and Atmospheric Administration (NOAA). It later expanded with partners in Canada and Mexico. ITIS inherited approximately 210,000 scientific names with varying levels of data quality from the NODC data set. While many important taxonomic groups were not well represented (e.g., terrestrial insects), the rate of errors and omissions within represented taxonomic groups ranged from relatively low (e.g., few misspellings or occasional

typographical errors) to rather high (e.g., many species names without authors or dates, or species assigned to wrong groups).

The ITIS mission is to create a scientifically credible database of taxonomic information, placing primary focus on taxa of interest to North America, with world treatments included as available. Within this framework, the initial data content development and quality assurance strategy was to begin with the NODC data and proceed on two tracks: (1) adding new names or checklists with a high level of taxonomic credibility, and (2) reviewing and verifying the legacy NODC data, thereby bringing it to a minimal, or higher, standard of data quality. Pending review and improvement, the unverified legacy data have been retained in the ITIS database to meet the needs of ITIS partners and collaborators who use the names and their associated unique identifiers (Taxonomic Serial Numbers - TSNs) in specific applications. Since the 1996 import of the legacy dataset, ITIS has grown to more than 334,000 scientific names, more than 55% of which have been verified in the literature, leaving about 140,000 names as unverified legacy data.

Although the ITIS database initially was populated with names derived from the NODC Taxonomic Code, it is being expanded to link individual names to one or more credible references (e.g., print publications, recognized experts, and databases). Those references may or may not also be linked to other names contained within the group, i.e. a family name may be linked to a publication that was used to verify its status and position, but that publication might not reference the subordinate genera or species within the family. This process of verification based on credible references is at the core of ITIS activities.

Depending on the rank of the scientific name (e.g. kingdom, family, subspecies, etc.), each ITIS name record has one to three data quality indicators associated with it:

- Record Credibility Rating (an example of the *Believability Taxonomic Data Quality Dimension*)
- Latest Record Review (an example of the *Timeliness Taxonomic Data Quality Dimension*)
- Global Species Completeness (an example of the *Completeness Data Quality Dimension*)

Every scientific name record in ITIS, regardless of the name's rank, has the data quality indicator "**Record Credibility Rating**" denoting whether it has undergone internal review. Because the NODC records originally imported into the ITIS database were of unknown quality, each was assigned a Record Credibility Rating of "unverified." As these records are reviewed, credibility ratings are changed to either the highest value, "verified – standards met", or "verified – minimum standards met." A rating of "verified – standards met" indicates to the user that all elements in the record and the position of the scientific name in the hierarchy are perceived to be accurate and supported by one or more credible references. If data in the record have been reviewed but are incomplete and/or contain accuracy, placement, or nomenclatural issues, or are from a non-peer reviewed source, a rating of "verified – minimum standards met" is assigned. During the process of adding new names to ITIS, some of the unverified legacy records in the same taxonomic group are vetted (i.e., unverified records are verified and the Record Credibility Rating is adjusted). As a result, more than 30% of the original NODC records have now been verified, and efforts to improve the quality of ITIS legacy data continue.

"**Latest Record Review**", a second ITIS data quality indicator, is assigned to records with names at ranks above species (e.g., genera, families, orders), and represents the

year that the record was last reviewed by ITIS. For example, if a family name is listed as “verified – standards met,” with a Latest Record Review rating of “1997,” a user can assume that the record was reviewed in that year. Users should be aware however, that taxonomic changes might have been published since that review and not yet incorporated into the ITIS files. (Additionally, users can refer to dates of cited publications which provide another indication of the currency of the record.) For original NODC data, all records were initially rated as “unknown” for this data quality indicator, but are being adjusted as records are reviewed.

The third ITIS data quality indicator, “**Global Species Completeness**,” indicates whether or not, for a given valid/accepted name (i.e., current standing) for a taxon at the rank of genus or higher, all known valid/accepted species for that rank were incorporated into ITIS at the time of review. Completeness ratings of “unknown” (such as were given to the original NODC data records) are adjusted to the appropriate level, “partial” or “complete,” when adequate information supports a change. Both “Latest Record Review” and “Global Species Completeness” indicators also are used by ITIS in making decisions about the timeliness of peer review of a group (ITIS - Integrated Taxonomic Information System, 2004).

8.3.2. The Brazilian Northeast Plant Checklist

The Brazilian Northeast Plant Checklist database was set up initially from a published checklist (Barbosa *et al.*, 1996), which was built based on different publications about plant species occurring in the northeast region of Brazil. Some of the references used to set up the checklist have no link between the species name and a herbarium voucher. This characteristic, species data without an association with the recorded biological fact, the herbarium voucher, has a significant impact on the Believability Taxonomic Data Quality Dimension (see section 5.2) for a specific class of Knowledge Workers: The taxonomists.

In order to overcome this impact, the first step was to establish a name status variable and a revision flag at the database, similar to the “Record Credibility Rating” adopted by ITIS.

At phase I of the database, all the scientific names were set up with a status as “provisional”, indicating that the name had not been verified by a reliable reference: a herbarium voucher citation or a taxonomist.

The revision phase II consisted of the revision of the names based on selected reliable literature. During phase III, a list of names, organized by family, was sent to taxonomists to revise.

As the revision of the database advanced, by incorporating new reliable references (phase II) or by the revision by taxonomists (phase III), the names were promoted to the status “accepted” or “synonym” and the revision flag received values of either “revised by reliable bibliography” or “accepted as a corrected name by”, associated with the name of a taxonomist.

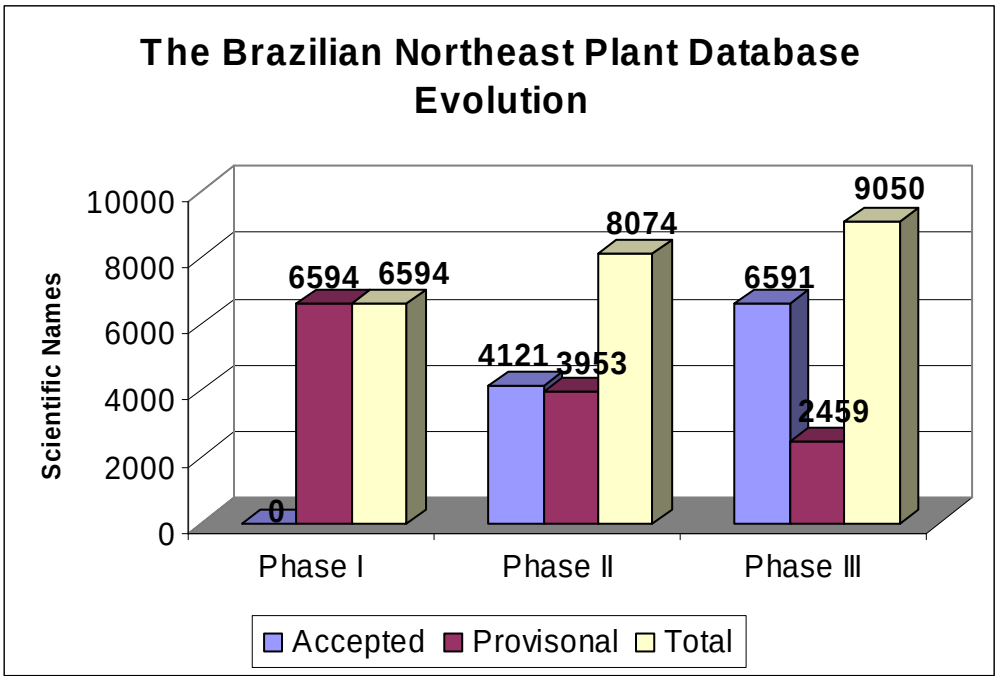


Figure 40 – Evolution of the Brazilian Northeast Plant Database

In both examples, the implemented data quality indicators will provide to knowledge workers and end users a measure of quality in some domains and for some dimensions. Yet in both examples, although these indicators seem to have been defined for operational and management reasons, they are also relevant to the users as means for assessing data quality.

9. Discussion

9.1. Can erroneous data in taxonomic databases be automatically corrected?

At this point it should be clear that automated error detection techniques can be used with a reasonable level of success to detect erroneous data in taxonomic databases, especially errors in scientific names. To select a reduced number of records from large databases, to be verified by knowledge workers, is a significant assistance to promote data quality, but could we go further? In the case of detecting spelling errors, rather than simply detecting errors, appropriate algorithms can suggest a correct name based on reliable data dictionaries or detecting similarity between names. However, fully automated correction has had very restricted practical application so far.

Froese (1997) presented an algorithm which automates and simplifies the process of verifying scientific names of fish. Two aims were defined for Froese's algorithm:

1. To assign a valid name automatically to as many as possible of the submitted names;
2. To provide all necessary information to assist the user in selecting a valid name for cases where a valid name cannot be assigned automatically.

Based on the rules of the International Code of Zoological Nomenclature, the algorithm relies on three auxiliary tables:

1. The FishBase FAMILIES table;
2. The FishBase GENERA table;

3. The FishBase SYNONYMS table, which contains valid names, synonyms and misspelled names.

Associated with those reference data dictionaries, Froese's algorithm follows four major steps:

1. If a submitted name exactly matches a valid, synonymous or misspelled name, the algorithm adds the valid name (if synonymous) or synonyms (if valid), author and year, family and additional references;
2. If no exact match is found, the algorithm tries to find unambiguous corresponding names by testing a set of different combinations of valid genus (when the submitted genus was a synonymous) and modifications of the suffix of the specific epithet, according to a set of common misspelling suffices. If a match is found, the algorithm adds the additional information, described at step one;
3. If a matching name hasn't been found in step 1 or 2, the algorithm checks the validity of the generic name. If a matching valid, synonymous or misspelled generic name is found, the algorithm adds the valid genus and family to the submitted name. If no matching generic name can be found, the algorithm tries to find similar names, following a series of misspelling tests. The resulting comments and possible generic name are printed. Based on these suggestions, obviously misspelled generic names are corrected manually and the algorithm is rerun starting with the step 1 above.
4. If steps 1-3 have not found a matching name, the algorithm makes suggestions for those names for which at least a valid generic name had been found, by checking for misspellings in the specific epithet.

Following those steps, if no exact matches are found, Froese's algorithm generates a list of suggestions, which need the intervention of a specialist (knowledge worker) in order to manually correct the name (Table 25).

Input	Output
Abantennarius neocaledoniensis	= Antennarius coccineus (Lesson 1830), Ref. Pietsch, T.W. and D.B.Grobecker, 1987
Ablavys binotata	= Ablabys binotatus (Peters 1855), Ref. Poss, S.1986, p. 479
Abramis pekinensis	= Parabramis pekinensis (Basilewsky 1855), Ref. Berg, L.S., 1964, p. 358

Table 25 – Examples of the results of Froese's algorithm

Musso (1988) presented an algorithm to clean up old legacy data which may, for example, have been entered in upper-case letters. It automatically transliterates a species name into a normalized scientific name, in accordance with the majority of the international recommendations. Musso's algorithm does not use any dictionary and focuses on identifying the name elements and transliterating then into the appropriate upper or lower case. The algorithm is capable of identifying and normalizing authority names as well (Table 26).

Original name	Automated corrected name
SAXIFRAGA FLORULENTA MORETTI	Saxifraga florulenta Moretti
HELIANTHEMUM LUNULATUM (ALL)%DC. IN LAM.ET. %DC.	Helianthemum lunulatum (All.)DC. In Lam. & DC.
ACHILLEA ERBA_ROTTA ALL.	Achillea erba-rotta All.

Table 26 – Examples of the results of Musso's algorithm

Musso's algorithm represent a simple script which follows pre-defined rules and recognizes special characters (such as "%") in order to standardize the format of the scientific name.

The LITCHI Project (http:// litchi.biol.soton.ac.uk) has been working on a model of the knowledge integrity rules in a taxonomic treatment held in taxonomic databases, in order to detect and correct taxonomic classification conflicts. The project produced, among several other useful rules and models, a *Conflict Reasoning Engine* which automatically detects conflicts in taxonomic checklists, and an interactive interface to promote semi-automated data correction (Jones *et al.*, 2001). In the specification of the rules and possible repairs (Brandt, 1999), LITCHI intends to achieve, in specific situations, an automated "repair process". The automated repair relies on an analysis of the "pieces of evidence", which can suggest an automated repair or indicate the need for a skilled *knowledge worker* to resolve the conflict.

Data other than scientific names may also be corrected in certain circumstances. For example, an occurrence of a *missing data values* error pattern may be automatically corrected in some specific cases, such as missing latitude and longitude for a known recorded locality (spatial data sub-domain). However, automated correction of species descriptive domain records, without a link to primary data sources (specimen descriptive data sub-domain), will usually be impracticable. If the link exists, morphological characteristics of specimen records can be used to complete species descriptor data domains, for example, the average size of the specimens to achieve the species "habit", or the specimen spatial occurrence to achieve the geographic distribution of habitat.

As pointed out by Maxted *et al*. (1993), descriptive data from superior levels of the taxonomic hierarchy can also be used to tackle missing data values at subordinate hierarchical levels. For example, a white flower characteristic ascribed to a whole genus

can be applied to all the species which belongs to this genus, in a process called *Data Propagation*.

In the same paper, Maxted *et al.* showed the automatic synthesis of descriptive data using the taxonomic hierarchy, in a process of *Data Aggregation* from subordinate hierarchical levels in order to define descriptive data for a taxon from specimen data.

Other automated data cleaning techniques appear to have useful application to automatically correct the *duplicate occurrences* error pattern of scientific names, caused by misspelling errors. Hernandez and Stolfo (1998) present an "Equational Theory" which is a framework to allow tests for duplicate records to be specified and executed. It could be applied to improve scientific name spelling error detection and, maybe, achieve fully automated correction.

For Hernandez and Stolfo (1998), the comparison of records to determine their equivalence (e.g. two scientific names with high degree of string similarity) is a complex inferential process that considers much more information in the compared record and in related tables. They use a declarative language to make it easy to specify criteria for detecting duplicate records. Working with personal data, Hernandez and Stolfo gives an example supposing two person names are spelled similarly but not identically, and have the exact same address. We might infer they are the same person. On the other hand, suppose two records have exactly the same social security numbers, but the names and addresses are completely different. We could either assume the records represent the same person who changed his name and moved, or the records represent different persons, and the social security number field is incorrect for one of

them. Without any further information, we may perhaps assume the latter. The more information there is in the records, the better inferences can be made (Hernandez and Stolfo, 1998).

This example can be used for two scientific names, with the specific epithet spelled nearly (but not completely) identically. Both names have the same genus, family and authority values. In this case we could perhaps infer that both names represent the same taxon and one of those names (or possibly both) is misspelled.

Spelling error algorithms have proved, in this thesis, to be an efficient tool to create clusters of scientific names that could, by the similarity of their spelling, represent the same entity. The key to carry out an automated correction is to choose the correct name. Additional databases and/or data can be used to check the occurrence of both names in order to increase the level of confidence of the choice. As pointed out by Hernandez and Stolfo, the more information there is associated with the records, the better inferences can be made.

Following this principle, automated correction of a duplicate occurrence of scientific names, caused by misspelling errors, can be demonstrated by the following example:

	Nomenclatural Data Domain			
	Family	**Genus**	**Specie**	**Author**
Record #1	Fabaceae	Vicia	faba	L.
Record #2	Fabaceae	Vycia	faba	L.

	Classification Data Domain	
	Status	**Synonyms**
Record #1	Accepted	Faba bona Medikus Faba faba (L.) House Faba major Desf Faba sativa Bernh. Faba vulgaris Moench Orobus faba Brot. Vicia esculenta Salisb. Vicia vulgaris Gray
Record #2	Accepted	Faba faba (L.) House Faba minor Roxb. Faba sativa Bernh. Faba vulgaris Moench Vicia esculenta Salisb. Vicia vulgaris Gray

	Species Descriptive Data Domain		
	Vernacular names	**Flower colour**	**Fruit**
Record #1	Ackerbohne, Bokla, Broad Bean, Fava Bean, Feve des Marais, Pois Blanc, Small Bean, Tick Bean, Windsor Bean	white	pubescent
Record #2	Ackerbohne, Bokla, Broad Bean, Faba Bean, Fava Bean, Feve des Marais, Field Bean, Habas, Horse Bean, Pois Blanc	white with the wings smeared in black	

Reference databases	**presence of *Vicia faba***	**presence of *vycia faba***
Database #1	Y	N
Database #2	Y	N
Database #3	N	N
Database #*n*	Y	N

Based on the data comparison of the records for *Vicia faba* and *Vycia faba*, we can assume with a high degree of confidence that both records are intended to represent the

same entity, in this case the same taxon and an automated correction (or merge) can be done. Of course, the merging may be non-trivial to accomplish.

Using this example, for the specific situation of the *duplicate occurrences* error pattern of scientific names caused by misspelling errors, we assume that automated correction can be successfully accomplished, with a different level of confidence, depending on the amount of information associated with each taxon record. However, the level of confidence, the amount of taxonomic or descriptive information needed, ways to specify this information concisely and methods for merging duplicate records should be explored by further research on this area.

9.2. Can incorrect data in taxonomic databases be prevented?

Error prevention is considered to be far superior to error detection, since error detection is often costly and cannot guarantee to be 100% successful at any stage (Embury, 2001).

Error prevention requires a better understanding and definition of the taxonomic information chain, since “bad data” is, in a sense, a result of a “bad process” (English, 1999). Hence, data defect prevention should be treated in the scope of the whole process of information quality improvement.

When an incorrect data item is detected in a database, there are two fundamental approaches to deal with it:

- Be *reactive* to the problem and fix the consequences (data cleansing);

- Be *proactive* by analyzing the causes of the problem and eliminating the cause (information process quality improvement).

Data cleansing fixes the symptoms of information quality problems after the problem has happened. Information process quality improvement analyses and eliminates the cause.

Based on the *data life cycle model* presented in chapter 4 (Figure 3) of this thesis, we can assume that the acquisition activities determine all the inherent data quality[26]. Consequently, an information process quality improvement should include activities that involve data modelling and its implementation, and data acquisition and data management, which will require a better understanding and definition of the *taxonomic information chain*.

Apart from the design and implementation of the database itself, where bad design and bad implementation will lead to overall poor data quality, error prevention can be related, separately, to the *data producers*, *data intermediaries* and, in some cases, *knowledge workers*.

Considering *data producers,* those who are responsible for acquiring specific values from the "real world" for the attributes of the individual instances of the defined entity classes, error prevention will involve adoption of standards and training in, for example, correct observation, measurement and description of the biological or environmental

[26] *Inherent data quality* – The degree to which data accurately reflects the real-world object that the data represents or, simply stated, the data accuracy.

fact, the handling equipment and specific taxonomic training (using dichotomous keys, correct observation and measurement of morphological structures etc.).

Data intermediaries are, usually, data handlers or transcribers without the capacity for evaluation or judgement of the data values themselves. They are responsible for data-entry activities, which produce the most frequent data quality problems (Eckerson, 2002). Hence, error prevention techniques for data intermediaries must involve validation routines and referential integrity checks as part of the database management system interface.

Not surprisingly, survey respondents cite data-entry errors by data intermediaries as the most common source of data defects. Examples of errors include misspellings, transposition of numerals, incorrect or missing codes, data placed in the wrong fields and unrecognizable scientific names, vernacular names, abbreviations or acronyms. These types of errors are increasing as taxonomic databases move their interfaces to the Web. Web interfaces may lack data entry controls and may allow *knowledge workers*, who may be less aware of data quality issues, to enter data directly into databases (Eckerson, 2002).

An example of how better interfaces and validation routines can improve data quality can be seen in a typical case of a data intermediary entering a misspelled name *Vycia faba*. Even database management systems specially developed to handle scientific names, such as Alice Software®, do not complain about the prior existence of a record for *Vicia faba* and will accept the name as a new valid record. In this case, the use of a

data entry validation routine, such as a “do you mean” feature, as discussed in section 10.1, will have significant impact on error prevention.

The adoption of recognized data standards and coherent data models together with clearly defined data process and management, responsible curatorial procedures on the primary data sources and the adoption of validation routines will therefore have a significant impact on error prevention in taxonomic databases.

9.3. Can taxonomic databases be certified?

This question appears to be a just the “tip of an iceberg” because it raises some related questions: “Why should Taxonomic Databases be certified?”, “How can Taxonomic Databases be certified?” and ”What level of certification is appropriate?”

With the advent of the Internet and the World Wide Web, abundant sources of “taxonomic information” became available to the general public. Some of those associated with reputable scientific institutions and projects confer on the information available a reasonable reputation. However, “reasonable reputation” is a subjective concept and a fragile attribute on which to base actions and decisions that involve, for example, rational and sustainable use of biodiversity resources.

Distinguishing reliable information from reputable information or even from ordinary unreliable information is not a straightforward task, even for taxonomists. They generally appeal to additional data like the author of the information, the institution and whether there is reference or citation to “verifiable material”, such as a herbarium

voucher. For databases, this additional data is usually called *metadata*[27]. However, the presence or accessibility of *metadata* for a given database still does not guarantee data quality and cannot be considered a data quality certification by itself. For *knowledge workers*, metadata can give important information in order to establish a certain degree of quality, in some taxonomic data quality dimensions. For example, the name of a specialist associated with a checklist, as a reviewer, can imply or even assure a good quality for the *classification data domain*. However, for *end users*, or the general public, *metadata* can be, in most cases, meaningless. Therefore, a taxonomic data quality certification can be used as a flag, indicating, in appropriate language, that a specific source of taxonomic information is *accurate, complete, consistent, timely* or whatever taxonomic data quality domain can be appropriately applied. With certification, taxonomic information consumers, from *knowledge workers* to *end users*, will be capable of recognizing, for example, from among the profusion of web pages with taxonomic information, those pages with reliable information.

There is also a secondary reason as to "*why*" Taxonomic Databases should be certified: to achieve certification, taxonomic database owners will be motivated to promote an improvement in the whole information quality process. Exposing taxonomic databases to criticism and public evaluation has proved to be an effective method to promote data quality (Dalcin *et al*., 1997). This may also be true for open-source program code, for example, in the case of security applications.

As shown in this thesis, the inherent data quality of taxonomic databases can be assessed in different dimensions and domains. However, to effectively establish a data

[27] *Metadata* is defined as "data about data" or "information about data or other information"

quality certification programme, a protocol should be developed and proposed, including the different aspects of the *taxonomic information chain*: primary data sources (the raw material), the information chain itself (the process) and the database (the product).

It will be appropriate for independent organizations, like the Taxonomic Database Working Group (http://www.tdwg.org), to conduct a taxonomic data quality certification programme, in order to set up a standardized certification protocol for public and private taxonomic databases. Once such a protocol exists, it will need to be applied, either by an appropriate agency, or perhaps by means of self-certification.

9.3.1. Suggestions for a Data Quality Certification Model for Taxonomic Databases

Taxonomic data quality certification procedures should include the following elements, among others:

- Public taxonomic databases should adopt data quality indicators, to different levels and for different taxonomic data quality dimensions and taxonomic data domains;
- Data quality indicators should be published in the public interface of the database, as Metadata;
- A timeliness indicator can be represented by revision dates associated with taxonomic data domains, data dictionaries and even individual data values (e.g. gazetteer updated on 2002, nomenclatural synonym added on 23/10/1989, ARACEAE reviewed by an expert in July 2002, etc.)

- A completeness indicator can be represented by a quantitative indicator, maybe a percentage, of the taxonomic data domains present at the database (see chapter 3). For example: 87% of genera have authority values; 98% of specimen records have latitude and longitude values; 65% of taxon records have a habit value, etc.
- A believability or credibility indicator can be represented by "verified" and "unverified" flags, associated with the records or even data values, similar to those adopted by ITIS Project (see section 8.3.1);
- Periodical reports about the data management procedures, updates, downtime etc should be publicly available;

9.4. What is the impact of taxonomic data quality on database merging and linking?

As discussed in the Chapter 2 of this thesis, several initiatives and significant efforts have been undertaken in order to build *global species databases* and to produce a compiled global checklist of living organisms (Bisby, 1993; 1997; 2000; Edwards *et al*., 2000; Embury *et al*., 2001; Embury *et al*., 1999). However, few of these initiatives have given special attention to data quality problems, mainly those which involve fundamental taxonomic data dimensions and basic error patterns.

As pointed out in chapter 6, a scientific name is a label - a textual representation - for a biological entity: the taxon. In taxonomic databases, this particular string of characters usually represents the main key to store, retrieve and access all the information related to a particular taxon or "entity". Several initiatives which are merging or linking taxonomic databases are using and offering the scientific name as a "unique identifier" without appropriate verification, for example, of spelling. As pointed in chapter 7 of this

thesis, where in every database used in the experiments, spelling errors were found that indicated possible entity duplication, we can assume that the same problems will occur when organisations or institutions decide to link or merge all databases together.

The assumption, by a database owner, of a high level of quality of their databases may be more dangerous than the lack of specific tools, personnel and procedures to assess the data quality in their databases. On the other hand, the recognition that the database is not “perfect”, or prone of errors, will encourage improvement of the quality.

Merging and linking taxonomic databases is not a trivial task. Complex concepts and rules are involved and often broken, in a semantically ambiguous world. Researchers in this area very often assume that their “raw material of research” – taxonomic databases – are “ideal databases”, as required by logical and mathematical studies. Unfortunately, this assumption is rarely (maybe never) true. Therefore, besides the semantic complexity of merging taxonomic databases, and in the light of the results of the tests applied in chapter 7 of this thesis, we can assume that the data quality level of the result of merging or linking one or more taxonomic databases will be equal to the data quality level of the worst database involved in the process.

In general, merging and linking use data from the database to:

a) Test which entities are to be merged (e.g. duplicate taxon detection);
b) Decide which data elements correspond with each other and can be merged;
c) Obtain the actual data and merge or link it.

Thus poor data quality can affect the linking or merging process in all these stages, and any deleterious effects of poor data may be multiplied.

Although it is beyond the scope of this thesis to design algorithms for linking or merging data records or data sets, clearly such algorithms must be designed with the effects of data quality in mind.

10. Conclusion and recommendations

10.1. Conclusions

Data quality studies need solid bases to generate practical benefits. Those bases include, for example, the understanding and description of the processes involved in data acquisition, treatment and management; information production, usage and delivery; and data modelling and implementation.

Data quality studies also bring to the scene concepts of data quality dimensions which depend on those bases to be appropriately applied, in order to establish a solid and practical relationship between the concept of data quality and taxonomic data.

Despite the advances in taxonomic databases in recent decades, data quality issues related to taxonomic databases have received very little attention, from both the disciplines of taxonomy and information science. Therefore, because of the lack of previous studies providing understanding and definitions of the taxonomic information chain in the context of data quality, the first part of this thesis was dedicated to proposing definitions of *taxonomic data domains* and *taxonomic data quality dimensions* in order to establish a framework to support the practical work, emphasizing the problem of spelling errors in scientific names. However, these definitions may need further refinement in the face of the complexity of taxonomic information, from a holistic point of view.

Consequently, a first conclusion of this work is the urgent need for the establishment of those solid bases in order to support forthcoming studies on data quality applied to taxonomic databases.

The most significant aspect of data quality issues in taxonomic database nowadays is related with the Internet which promotes a shift in the way taxonomic information is delivered.

A few years ago, taxonomic information was delivered exclusively by printed materials where the previous verification of the first pass[28] by a copy editor[29] was a standard procedure (Figure 41). Publications about taxa were revised by *knowledge workers* before formally being published as books or other printed materials and, consequently, becoming available to *end users*.

[28] An early printed edition of the manuscript, which is reviewed for accuracy by the author and copy editor before publication.
[29] Corrects grammar and spelling in a manuscript and checks facts for accuracy and conformity.

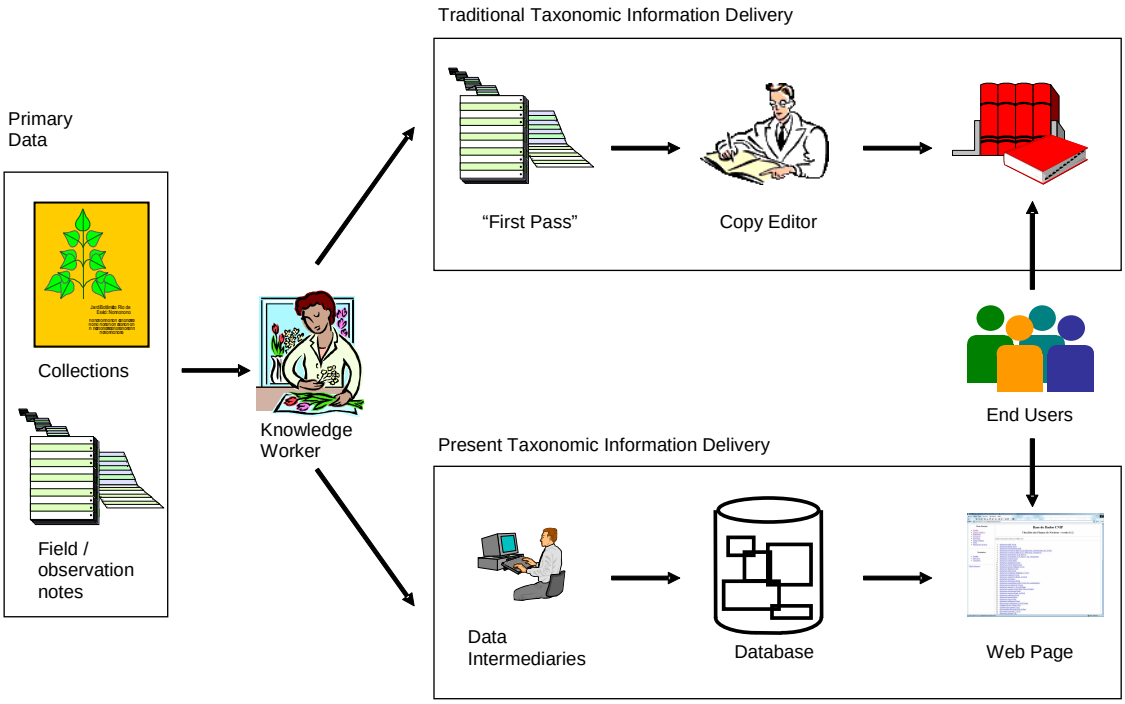


Figure 41 – Comparison of traditional and present taxonomic information delivery

Nowadays, taxonomic databases have become primary data sources for applications and languages capable of delivering taxonomic information on demand through the Internet, directly to consumers. This "quiet revolution" (Bisby, 2000), despite the advances and benefits, can lead to a scenario where, for example, there will be no need for a *knowledge worker* to be in touch with the biological material itself, since all the aspects and data of the specimen, including images, sounds and detailed measurements, are "digitized" and accessible by a few "clicks". In other words, the content of taxonomic databases could become interpreted, or even recognized, as a primary data source in order to generate new taxonomic information products, or used to make biodiversity assessments. Since databases are prone to errors, relying exclusively on databases to produce knowledge, without strong links to verifiable biological material could lead to incorrect assessment and decision making.

In the same way, the increasing pervasiveness of the Internet among the general public has been conferring to the World Wide Web a status of primary data source and, again, this is easily interpreted as it is a source of reliable data by the general public.

Even the “interoperability of biodiversity databases”, the “holy grail” of recent global efforts, meetings and organizations involved with taxonomic databases, made possible with the advent of the Internet, relies on linking related information using scientific names, which are themselves threatened by taxonomic data quality issues.

This new development of the *taxonomic information chain* comes to endorse the necessity of a better understanding, description and definition of the *taxonomic information chain* itself, in order to support a framework where the concepts of data quality can be appropriately applied.

The second conclusion from this thesis is that, as in any other database, taxonomic databases are prone to errors. The practical work in this thesis, working with five different taxonomic databases and focusing on the evaluation of the *nomenclatural data domain*, demonstrated the detection of data quality problems in all the five databases, over almost all the *taxonomic error patterns* proposed and described in section 6.26.

However, as pointed out in the discussion (section 9.2), incorrect data is just a consequence of a “bad process”. It is outside the scope of this thesis to answer questions which could delimit or identify which part of the process or information chain was responsible for the incorrect data. Were the data entered wrongly? Were the data entered correctly but it is incorrect at the source? Were the incorrect data a result of

unsuccessful data conversion or migration? Those are some questions which address data quality issues as a whole; questions that should arise and be treated in any taxonomic database assessment, as a part of an information quality process improvement.

Consequently, the third conclusion of this thesis, based also on the literature review, is that the majority of taxonomic database projects have neglected aspects of data quality control and, we can assume, have no information quality process improvement plans or policy.

Three of the five databases used on the experiments of this thesis were built and managed using a database management system, developed specifically to handle taxonomic nomenclature. However, even this specially designed tool was incapable of clearly understanding or detecting mistyped scientific names, which can result in, for example, a duplicate occurrence of a taxon.

There is a fourth conclusion of this thesis which pointed out that implementations of taxonomic database management systems have also neglected to promote an effective data entry validation routine or appropriate interface and algorithms that could be recognized effectively as contributing to a “taxonomic intelligent” database program. In other words, taxonomic databases need to incorporate effectively the “intelligence” needed not just to promote, for example, the referential integrity of their related tables, but also the “intelligence” necessary to promote overall data quality. Again, it necessarily involves an accurate knowledge of the *taxonomic information chain*. For the same reason, taxonomic databases should incorporate the same “intelligence” when

requested to deliver information to the general public. Let us examine the following examples:

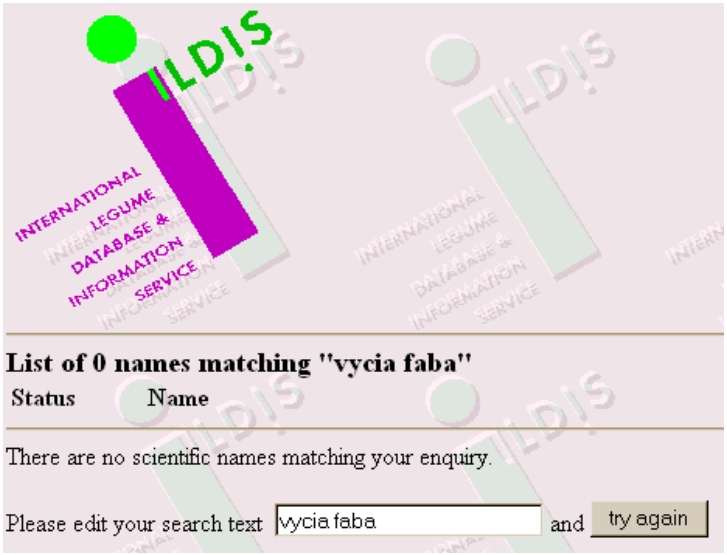


Figure 42 – Example of a query to the ILDIS database

Figure 42 represents a query to the ILDIS database web site (LegumeWeb), using a misspelling of the scientific name *Vicia faba*. Despite the existence of data on *Vicia faba* in the ILDIS database, a simple misspelled request will return no names.

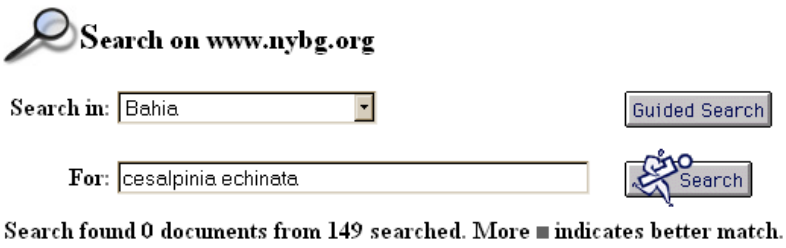


Figure 43 – Example of a query to the New York Botanical Garden Database

Figure 43 represents a query to the New York Botanical Garden Database Web Service and, despite the existence of data on *Caesalpinia echinata*, again the query using a simple misspelled name will return no data at all.

These two examples are used here to illustrate the absence of “intelligence” in these interfaces or, more precisely, the incapacity of the interface to recognize the intention behind the operation performed by the user.

Fortunately, recently implemented web search engines have adopted a better approach for misspelled queries. The Taxonomic Search Engine (Figure 44), implemented by Rod Page as a test bed for a web service approach to federating taxonomic name databases (http://darwin.zoology.gla.ac.uk/~rpage/portal/) has implemented recently what he calls a “did you mean” feature, using a Google API (http://www.google.com/apis/index.html) to recognize misspelled queries (Figure 45).

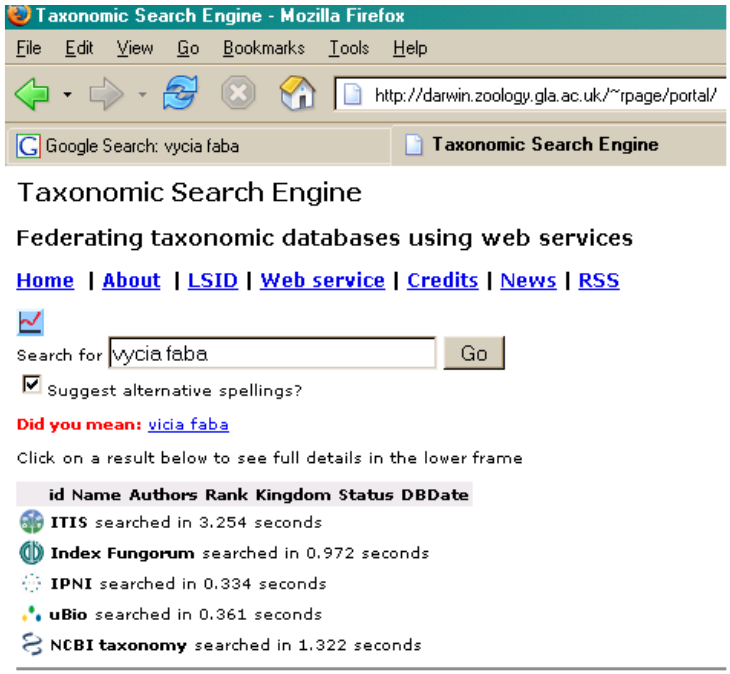


Figure 44 – Rod Page’s Taxonomic Search Engine interface and its “did you mean” feature

Figure 45 – Google API for misspelled queries in action

The results of the experiments carried out on chapter 7, lead to the fifth and sixth related conclusions as well. The fifth conclusion is that many spelling errors in taxonomic databases can be automatically detected; and the sixth, that the algorithms available to detect spelling errors in scientific names must be improved.

The results of the experiments performed on chapter 7 have clearly demonstrated that, using algorithms to detect string similarity and to perform string manipulation, it is possible to detect and pick out pairs of names that strongly suggest a duplicate representation of a single taxon. However, the significant number of occurrences of "false alarm" behaviour, and the relationship between the "hit" and the "false alarm" behaviour suggest that the algorithms used in the experiment, based on their originally published definitions, can be adapted in order to handle scientific names with better performance and efficiency. This is particularly true for phonetic similarity algorithms since they were originally developed to handle the names of people.

As a final conclusion, data quality as applied to taxonomic databases needs significantly more studies in order to cover all the aspects of the *taxonomic information chain*, and an

taxonomic data domains. Even those definitions, proposed here in the limited scope of this thesis must be explored by future studies, in order to support forthcoming developments in taxonomic databases which will effectively improve the data quality in taxonomy.

10.2. Recommendations

Based on the conclusions above, we can point out the following recommendations:

10.2.1. General recommendations for knowledge workers and database administrators

1. Setting up a proper taxonomic database is not a trivial task and its complexity should not be underestimated by knowledge workers, or even by the information technology workforce. Database modelling, attribute definitions, data flow process, interface design, etc. must be developed and implemented with data quality issues in mind;
2. In view of the fact that every database is prone to error, knowledge workers must bear in mind that they should not rely uncritically on taxonomic databases or their automatically generated products as reliable data sources for critical assessments or decisions without careful consideration of their reliability;
3. Taxonomic database projects should have data quality policies and must incorporate data quality assessment and improvement procedures in their operational routine. Examples:

a. Data quality assessment and improvement must involve primary data sources;
b. Data producers must be trained to gather information as precisely, accurately and completely as possible;
c. Data intermediaries should be trained in basic concepts of taxonomy in order to be more sensitive to issues about inherent data quality;
d. Data producers and knowledge workers should be involved in database design, especially in data modelling and interface design;

4. Taxonomic database management systems must incorporate data entry and data request validation routines and become more “taxonomically intelligent”;
5. Taxonomic databases with public access, especially those with access through the Internet, must adopt data quality indicators in order to make it possible for end users and knowledge workers to evaluate the reliability of the data;
6. Taxonomic databases should take advantage of the criticism of experts in order to contribute to further data quality improvement. Publication should not be delayed until the database is considered as a final product. In this case, the taxonomic database interface should provide a means to accept, manage and incorporate feedback from contributors.

10.2.2. Recommendations for further research

1. The processes involved in taxonomic data acquisition, manipulation, storage and usage, here called the *taxonomic information chain*, must be a focus of further studies in order to establish a solid framework where data quality concepts and techniques can be applied and evaluated, in order to address data quality issues in taxonomic databases;

2. The concepts of Taxonomic Data Quality Dimensions and Taxonomic Data Domains should be further validated by other practical applications in the “real world”;
3. The aspects of the “pragmatic data quality” of taxonomic databases, not deeply considered in this thesis, must be investigated in further research, in order to establish a complete understanding of data quality as it affects the acquisition, storage, delivery and usage of taxonomic information;
4. Spelling error algorithms should be improved in order to better handle the characteristics of scientific names and taxonomic nomenclature;
5. The applicability of spelling error algorithms as a tool for data quality assessment and data defect detection should be tested with other types of taxonomic data, such as authority names, bibliographic references, geographical data (place names), etc.;
6. Automated correction techniques, not considered on this thesis, should be explored, perhaps using the concepts of taxonomic data domains and taxonomic error patterns described in this thesis;
7. Data quality certification for taxonomic databases should be discussed, developed and proposed by an independent organization in order to promote the overall data quality of taxonomic databases and the recognition of reliable taxonomic information sources;

References

Adams, R. P. 1974. Computer graphics plotting and mapping of data in systematics. Taxon 23:53-70.

Adey, M. E., R. Allkin, F. A. Bisby, T. D. Macfarlane, and R. J. White. 1984. The Vicieae Database: An Experimental Taxonomic Monograph. Pages 175-188 *in* Databases in Systematics (R. Allkin, and F. A. Bisby, eds.). Academic Press, London.

Alberga, C. N. 1967. String similarity and misspellings. Communications of ACM 10:302-313.

Allkin, R. 1980. Some difficulties of computer-stored taxonomic descriptions.Instituto Nacional de Investigaciones sobre Recursos Bióticos (INIREB), Xalapa, Mexico.

Allkin, R. 1984a. Computer management of taxonomic data with examples for Sicilian Vicieae (Leguminosae). Webbia 38:577-583.

Allkin, R. 1984b. Handling taxonomic descriptions by computer. Pages 263-278 *in* Databases in systematics (R. Allkin, and F. A. Bisby, eds.). London.

Allkin, R. 1986. Nuevos métodos para la identificación de angiospermas en el estado de Veracruz, Mexico, Medellín, Colombia.

Allkin, R. 1987. ILDIS Exchange Data Format: Trial Data Definitions in XDF(C. C. ILDIS, Biology 44, ed.) The University, Southampton.

Allkin, R. 1988. Taxonomically intelligent database programs. Pages 315-331 *in* Prospects in Systematics (D. L. Hawksworth, ed.) Oxford University Press, Oxford.

Allkin, R. 1989. Data management models for taxonomy*in* Research Seminar Series IBM Scientific Research Centre, Winchester, HANTS UK.

Allkin, R., and F. A. Bisby. 1988. The Structure of Monographic Databases. Taxon 37:756-763.

Allkin, R., and R. J. White. 1989. XDF - a language for the definition and exchange of biological data sets. Description & Manual, Version 3.3.

Allkin, R., R. J. White, and P. J. Winfield. 1992. Handling the Taxonomic Structure of Biological Data. Math. Comput. Modelling 16:1-9.

Allkin, R., and P. J. Winfield. 1990. An introduction to ALICE: a database system for species diversity and checklist databases*in* DELTA Newsletter (R. Webster, ed.) USDA, Beltsville.

Allkin, R., and P. J. Winfield. 1993. ALICE Species checklist and database management system. Pages 473 *in* Advances in Computer Methods for Systematic Biology: Artificial Intelligence, Databases and Computer Vision (R. Fortuner, ed.) Johns Hopkins University Press, Baltimore and London.

Argus, G. W., and J. W. Sheard. 1972. Two simple labeling and data retrieval systems for herbaria. Can. J. Bot. 50:2197-2209.

Barbosa, M. R. V., S. J. Mayo, A. A. J. F. Castro, F. L. Freitas, M. S. Pereira, P. C. Gadelha Neto, and H. M. Moreira. 1996. Checklist preliminar das angiospermas. Pages 253-415 *in* Pesquisa Botânica Nordestina: Progressos e Perspectivas. (E. V. B. Sampaio, S. J. Mayo, and M. R. V. Barbosa, eds.). Soc. Bot. Brasil, Reg. Pernambuco, Recife.

Barron, D. W. 1984. Current Database Design - the User's View. Pages 35-41 *in* Databases in Systematics (R. Allkin, and F. A. Bisby, eds.). Academic Press, London.

Beach, J. H., S. J. Ozminski, and D. E. Boufford. 1993. An Internet Botanical Specimen Data Server. Taxon 42:627-629.

Beaman, J. H., and J. C. Regalado Jr. 1989. Development and Management of Microcomputer Specimen-Oriented Database for the Flora of Mount Kinabalu. Taxon 38:27-42.

Berendsohn, W. G. 1995. The concept of "potential taxa" in databases. Taxon 44:207-212.

Berendsohn, W. G. 1997. A taxonomic information model for botanical databases: the IOPI model. Taxon 46:283-309.

Berendsohn, W. G., A. Anagnostopoulos, G. Hagedorn, J. Jakupovic, P. L. Nimis, B. Valdés, A. Güntsch, R. J. Pankhurst, and R. J. White. 1999. A Comprehensive Reference Model for Biological Collections and Surveys. Taxon 48:511-562.

Beschel, R. E., and J. H. Soper. 1970. The automation and standardization of certain herbarium procedures. Can. J. Bot. 48:547-554.

Bisby, F. A. 1984a. Automated Taxonomic Information Systems. Pages 301-322 *in* Current Concepts in Plant Taxonomy (V. H. Heywood, and D. M. Moore, eds.). Academic Press.

Bisby, F. A. 1984b. Information Services in Taxonomy. Pages 17-33 *in* Databases in Systematics (R. Allkin, and F. A. Bisby, eds.). Academic Press, London.

Bisby, F. A. 1986. International Legume Database and Information Service (ILDIS). Taxon 35:217.

Bisby, F. A. 1988. Communications in taxonomy. Pages 277-291 *in* Prospects in Systematics (D. L. Hawksworth, ed.) Clarendon Press, Oxford.

Bisby, F. A. 1993. Botanical strategies for compiling a global plant checklist. Pages 145-157 *in* Designs for a Global Plant Species Information System (F. A. Bisby, G. F. Russell, and R. J. Pankhurst, eds.). Clarendon Press, Oxford.

Bisby, F. A. 1995. Plant names in botanical databases. Hunt Institute for Botanical Documentation, Pittsburgh.

Bisby, F. A. 1997. Putting names to things and keeping track: the Species 2000 programme for a coordinated catalog of life. Pages 59-68 *in* Information technology, plant pathology and biodiversity (P. Bridge, P. Jeffries, D. R. Morse, and P. R. Scott, eds.). Wallingford.

Bisby, F. A. 2000. The Quiet Revolution: Biodiversity Informatics and the Internet. Science 289:2309-2312.

Bisby, F. A., and S. M. Brandt. Year. Species 2000: A Global Architeture for the Catalog of Life *in* The International Joint Workshop for Studies on Biodiversity. National Institute for Environmental Studies, Japan:3-8.

Bosbach, K., A. Scott, and J. Steffen. 1990. ARS BOGOS - Automated registration system of the Botanical Garden of the University of Osnabrück. Botanic Gardens Conservation News 1:41-49.

Brandt, S. M. 1999. LITCHI Conflicts to the Rules and Possible Repairs - Version 3. Pages 31.

Brenan, J. H., and J. T. Williams. 1975. Computers in Botanical Collections. Plenum New York.

Brummitt, R. K. 1992. Vascular Plant Families and Genera. Royal Botanic Gardens, Kew.

Brummitt, R. K., and C. E. Powell. 1992. Authors of plant names. Royal Botanic Gardens, Kew.

Burley, J. R., P. R. Scott, and A. W. Speedy. 1997. Biodiversity: The role of information technology in distributing information. Pages 157-172 *in*

Biodiversity Information - Needs and Options (D. L. Hawksworth, P. M. Kirk, and S. D. Clarke, eds.). CAB International, Wallingford, Oxon.

Canhos, V. P., G. P. Manfio, and D. A. L. Canhos. 1997. Networks for distributing information. Pages 147-156 *in* Biodiversity Information - Needs and Options (D. L. Hawksworth, P. M. Kirk, and S. D. Clarke, eds.). CAB International, Wallingford, Oxon.

Carling, R. C. J., and J. Harrison. 1997. Biodiversity information on the Internet: Cornucopia or confusion? Biodiversity Letter 3:125-135.

Codd, E. F. 1970. A Relational Model of Data for Large Shared Data Bank. Communications of ACM 13:377-387.

Cook, E. M. 1995. Economic botany data collection standard. Royal Botanic Gardens, Kew.

Croft, J. R. (ed) 1989. HISPID - Herbarium Information Standards and Protocols for Interchange of Data - Version 1. Australian National Botanic Gardens, Canberra.

Crovello, T. J. 1967. Problems in the use of electronic data processing in biological collections. Taxon 16:481-494.

Crovello, T. J., L. A. Hauser, and C. A. Keller. 1984. BRASS BAND (The Brassicaceae data bank at Notre Dame): An example of database concepts in systematics. Pages 219-233 *in* Databases in Systematics (R. Allkin, and F. A. Bisby, eds.). Academic Press, London.

Dalcin, E. C., L. Solano, and R. Pizarro. 1997. De banco de dados a centro de informações e serviços: Uma experiência para a Reserva Ecológica de Macaé de Cima. Pages 307-313 *in* Serra de Macaé de Cima: Diversidade florística e conservação em Mata Atlântica (R. R. Guedes-Bruni, and H. C. de Lima, eds.). Jardim Botânico do Rio de Janeiro, Rio de Janeiro.

Dallwitz, M. J. 1980. A general system for coding taxonomic descriptions. Taxon 29:41-46.

Damerau, F. J. 1964. A technique for computer detection and correction of spelling errors. Communications of the ACM 7:171-176.

Duff, F. 1997. Overview of the UNEP/GEF Biodiversity Data Management Project (BDM). Pages 115-124 *in* Biodiversity Information - Needs and Options (D. L. Hawksworth, P. M. Kirk, and S. D. Clarke, eds.). CAB International, Wallingford, Oxon.

Eckerson, W. W. 2002. Data Warehousing Special Report: Data quality and the bottom line. Applications Development Trends May 2002.

Edwards, J. L., M. A. Lane, and E. S. Nielsen. 2000. Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. Science 289:2312-2314.

Embury, S. M. 2001. Data quality issues in information systems. Pages 41 *in* Database Systems Cardiff University, School of Computer Science, Cardiff.

Embury, S. M., S. M. Brandt, J. Robinson, I. Sutherland, F. A. Bisby, W. A. Gray, A. C. Jones, and R. J. White. 2001. Adapting integrity enforcement techniques for data reconciliation. Information Systems 26:657-689.

Embury, S. M., A. C. Jones, I. Sutherland, W. A. Gray, R. J. White, J. Robinson, F. A. Bisby, and S. M. Brandt. Year. Conflict Detection for Integration of Taxonomic Data Sources *in* 11th International Conference on Scientific and Statistical Database Magagement. IEEE Computer Society Press, Cleveland, Ohio, USA:204-213.

English, L. P. 1999. Improving data warehouse and business information quality: methods for reducing costs and increasing profits. John Wiley & Sons, Inc., New York.

Everard, M. 1993. A global plant taxonomy database: design considerations. Pages 198-218 *in* Design for a global plant species information system (F. A. Bisby, G. F. Russell, and R. J. Pankhurst, eds.). Claredon Press, Oxford.

Freeston, M. W. 1984. The Implementation of Databases on Small Computers. Pages 43-52 *in* Databases in Systematics (R. Allkin, and F. A. Bisby, eds.). Academic Press, London.

Froese, R. 1997. An algorithm for identifying misspellings and synonyms in lists of scientific names of fishes. Cybium 21:265-280.

Gadd, T. N. 1988. 'Fisching fore werds': phonetic retrieval of written text in information systems. Program 22:222-237.

Gadd, T. N. 1990. Phonix: The algorithm. Program 24:363-366.

Gibbs-Russell, G. E., and T. H. Arnold. 1989. Fifteen years with the computer: Assessment of the "PRECIS" taxonomic system. Taxon 38:178-195.

Gibbs-Russell, G. E., and P. Gonsalves. 1984. PRECIS - A curatorial and biogeographical system. Pages 137-153 *in* Databases in Systematics (R. Allkin, and F. A. Bisby, eds.). Academic Press, London.

Gómez-Pompa, A., N. P. Moreno, L. Gama, and V. Sosa. 1984. Flora of Veracruz: Progress and Prospects. Pages 165-173 *in* Databases in Systematics (R. Allkin, and F. A. Bisby, eds.). Academic Press, London.

Gómez-Pompa, A., N. P. Moreno, V. Sosa, L. Giddings, and R. M. Soto. 1985. Flora de Veracruz project: an update on database management of collections and related information. Taxon 34:645-649.

Gómez-Pompa, A., and L. I. Nevling Jr. 1988. Some Reflections on Floristic Databases. Taxon 37:764-775.

Graham, M., M. F. Watson, and J. B. Kennedy. 2002. Novel Visualisation Techniques for Working with Multiple, Overlaping Classification Hierarchies. Taxon 51:351-358.

Green, D. G. 1994. Databasing Diversity - a Distributed, Public-domain Approach. Taxon 43:51-62.

Greuter, W., J. McNeill, F. R. Barrie, H. M. Burdet, V. Demoulin, T. S. Filgueiras, D. H. Nicolson, P. C. Silva, J. E. Skog, P. Trehane, N. J. Turland, and D. L. Hawksworth. 2000. International Code of Botanical Nomenclature (Saint Louis Code). Koeltz Scientific Books, Königstein.

Hagedorn, G., and G. Rambold. 2000. A Method to establish and revise descriptive data sets over the Internet. Taxon 49:517-528.

Hall, A. V. 1972. Computer-based data banking for taxonomic collections. Taxon 21:13-25.

Hall, A. V. 1974. Museum specimen record data storage and retrieval. Taxon 23:23-28.

Hernandez, M. A., and S. J. Stolfo. 1998. Real-world data is dirty: data cleaning and the merge/purge problem. Journal of Data Mining and Knowledge Discovery 2:9-37.

Heywood, V. H. 1984. Eletronic Data Processing in Taxonomy and Systematics. Pages 1-15 *in* Databases in Systematics (R. Allkin, and F. A. Bisby, eds.). Academic Press, London.

Hollis, S., and R. K. Brummitt. 1992. World geographical scheme for recording plant distributions. Hunt Institute for Botanical Documentation, Pittsburgh.

ICZN., W. D. L. Ride, International Trust for Zoological Nomenclature, Natural History Museum (London England), and International Union of Biological Sciences. General Assembly. 1999. International code of zoological nomenclature, 4th

edition. International Trust for Zoological Nomenclature c/o Natural History Museum, London.

ITIS - Integrated Taxonomic Information System. 2004. Data Development History and Data QualityNational Museum of Natural History, Washington, D.C.

IUCN-BGCS. 1987. I.T.F. - The International Transfer Format for Botanic Garden Plant Records, 1st. edition. Hunt Institute for Botanical Documentation, Carnegie Mellon University, Pittsburgh.

Jardine, N., and R. Sibson. 1977. Mathematical Taxonomy. Wiley & Sons, London and New York.

Jeffrey, C. 1977. Biological Nomenclature, Second Edition edition. Edward Arnold, London.

Johnston, B. C. 1980. Computer Programs for Constructing Polyclave Keys from Data Matrices. Taxon 29:47-51.

Jones, A. C., I. Sutherland, S. M. Embury, W. A. Gray, R. J. White, J. S. Robinson, F. A. Bisby, and S. M. Brandt. Year. Techniques for effective integration, maintenance and evolution of species databases *in* 12th International Conference on Scientific and Statistical Database Management. IEEE Computer Society Press, Berlin.

Juma, C. 1997. The role of information in the operation of the convention on biological diversity. Pages 125-128 *in* Biodiversity Information - Needs and Options (D. L. Hawksworth, P. M. Kirk, and S. D. Clarke, eds.). CAB International, Wallingford, Oxon.

Kahn, B. K., D. M. Strong, and R. Y. Wang. 2002. Information Quality Benchmarks: Product and Service Performace. Communications of ACM 45:184-192.

Klein, B. D. 2001. Detecting errors in data: Clarification of the impact of base rate expectations and incentives. Omega 29:391-404.

Krauss, H. M. 1973. The use of generalized information processing systems in the biological sciences. Taxon 22:3-18.

Kukich, K. 1992. Techniques for automatically correcting words in text. ACM Computing Surveys 24:377-439.

Levitin, A., and T. C. Redman. 1995. Quality Dimensions of a Conceptual View. Information Processing and Management 31:81-88.

MacDonald, R. D. 1966. Eletronic data processing methods for botanical garden and arboretum records. Taxon 15:291-295.

Mackinder, D. C. 1984. The database of the IUCN Conservation Monitoring Center. Pages 91-101 *in* Databases in Systematics (R. Allkin, and F. A. Bisby, eds.). Academic Press, London.

Maletic, J. I., and A. Marcus. 1999. Progress report on automated data cleansing. Pages 13 The Department of Mathematical Sciences Division of Computer Science, The University of Memphis, Memphis, TN.

Maletic, J. I., and A. Marcus. 2000. Data Cleansing: Beyond integrity Analysis. Pages 10 Division of Computer Science, Department of Mathematical Sciences, The University of Memphis, Memphis.

Marcus, A., J. I. Maletic, and K. I. Lin. 2001. Ordinal Association Rules for Error Identification in Data Sets. Pages 15 Division of Computer Science, Department of Mathematical Sciences, The University of Memphis, Memphis, TN.

Maxted, N., R. J. White, and R. Allkin. 1993. The Automatic Synthesis of Descriptive Data Using the taxonomic Hierarchy. Taxon 42:51-62.

Moreno, N. P., and R. Allkin. 1981. Reporte preliminar sobre el conjunto de caracteres vegetativos para la Flora de VeracruzINIREB.

Moreno, N. P., and R. Allkin. 1988. Métodos computadorizados y algunas aplicaciones al estudio de la flora de Mexico. Bol. Soc. Bot. México 48:65-73.

Morin, N. R., and J. Gomon. 1993. Data banking and the role of the natural history collections. Ann. Missouri Bot. Gard. 80:317-322.

Morse, L. E. 1974. Computer-assisted storage and retrieval of the data of taxonomy and systematics. Taxon 23:29-43.

Morse, L. E., J. A. Peters, and P. B. Hamel. 1971. A general data format for summarizing taxonomic information. BioScience 21:174-180.

Motro, A., and I. Rakov. Year. Estimating the Quality of Data in Relational Databases *in* 1996 Conference on Information Quality:94-106.

Motro, A., and I. Rakov. 1998. Estimating the quality of databases. Lecture Notes in Computer Science 1495:298-308.

Musso, J. A. 1988. Carl_linne: an Automat to Transliterate Scientific Name of Species in Accordance with International Recommendations. J. Comp. Biol. 3:109-110.

Naumann, F. 2001. From Database to Information Systems - Information Quality Makes the Difference. Pages 17 IBM Almaden Research Center, San Jose, CA.

Pankhurst, R. J. 1974. Automated Identification in systematics. Taxon 23:45-51.

Pankhurst, R. J. 1975. Biological Identification with Computers. Academic Press, London and New York.

Pankhurst, R. J. 1986. A package of computer programs for handling taxonomic databases. Cabios 2:33-39.

Pankhurst, R. J. 1988. Database Design for Monographs and Floras. Taxon 37:733-746.

Pankhurst, R. J. 1993. Taxonomic Databases: The PANDORA System. Pages 229-240 *in* Advances in Computer Methods for Systematic Biology: Artificial Intelligence, Databases, Computer Vision (R. Fortuner, ed.) The Johns Hopkins University Press, Baltimore, Maryland.

Partridge, T. R., M. J. Dallwitz, and L. Watson. 1986. A primer for the DELTA system on VMS, MS-DOS and PRIMOSCSIRO - Div. of Entomology, Canberra.

Perring, F. H. 1963. Data-processing for the Atlas of the British Flora. Taxon 12:183-190.

Peterson, J. L. 1980. Computer programs for detecting and correcting spelling errors. Communications of the ACM 23:676-687.

Pfeifer, U., T. Poersch, and N. Fuhr. Year. Searching Proper Names in Databases *in* Hypertext - Information Retrieval - Multimedia, Synergieeffekte elektronischer Informationssysteme - HIM '95. Universitätsverlag Konstanz, Konstanz:259-276.

Pfeifer, U., T. Poersch, and N. Fuhr. 1996. Retrieval effectiveness of proper name search methods. Information Processing and Management 32:667-679.

Pipino, L. L., Y. W. Lee, and R. Y. Wang. 2002. Data Quality Assessment. Communications of ACM 45:211-218.

Pollock, J. J., and A. Zamora. 1984. Automatic spelling correction in scientific and scholarly text. Communications of ACM 27:358-368.

Pullan, M. R., M. F. Watson, J. B. Kennedy, C. Raguenaud, and R. Hyam. 2000. The Prometheus Taxonomic Model: a Pratical Approach to Representing Multiple Classifications. Taxon 49:55-75.

Raghavan, V. V., S. G. Jung, and P. Bollmann. 1989. A critical investigation of recall and precision as measures of retrieval system performance. ACM Transactions on Information Systems 7:205-229.

Raguenaud, C., J. Kennedy, and P. J. Barclay. 1999. Database Support for Taxonomy, Prometheus Napier University, School of Computing, Database and Objects Systems Group, Edinburgh.

Raguenaud, C., M. R. Pullan, M. F. Watson, J. B. Kennedy, M. F. Newman, and P. J. Barclay. 2002. Implementation of the Prometheus Taxonomic Model: a comparison of database models and query languages and an introduction to the Prometheus Object-Oriented Model. Taxon 51:131-142.

Rahm, E., and H. H. Do. 2000. Data Cleaning: Problems and Current Approaches. Bulletin of the Technical Committee on Data Engineering 23:3-13.

Redman, T. C. 1996. Data Quality for the Information Age. Artech House, Inc.

Redman, T. C. 1998. The Impact of Poor Data Quality on the Typical Enterprise. Communications of ACM 48:79-82.

Reznicek, A. A., and G. F. Estabrook. 1986. Herbarium specimens label and annotation printing program for IBM PC computers. Taxon 35:640-641.

Roughton, K. G., and D. A. Tyckoson. 1985. Browsing with sound: Sound-based codes and automated authority control. Information Technology and Libraries 4:130-136.

Shelter, S. G. 1974. Demythologising biological data banking. Taxon 23:71-100.

Silva, P. C. 1966. Machine data processing and plant taxonomy. Science 152:934-935.

Sneath, P. H. A., and R. R. Sokal. 1973. Numerical Taxonomy: the principles and practice of Numerical Classification. Freeman & Son, San Francisco.

Sokal, R. R. 1963. The principles and practice of numerical taxonomy. Taxon:190-199.

Sokal, R. R., and P. H. A. Sneath. 1966. Efficiency in taxonomy. Taxon 15:1-21.

Soper, J. H., and F. H. Perring. 1967. Data processing in the herbarium and museum. Taxon 16:13-19.

Squires, D. F. 1966. Data Processing and Museum Collections: A Problem for the Present. Curator IX:216-227.

Stafleu, F. A. 1966. Authors, taxa, names and computers. Taxon 15:109-114.

Stribling, J. B., S. R. Moulton II, and G. T. Lester. 2003. Determining the quality of taxonomic data. J. N. Am. Benthol. Soc. 22:621-631.

Strong, D. M., Y. W. Lee, and R. Y. Wang. 1997. Data quality in context. Communications of ACM 40:103-110.

Sugden, A., and E. Pennisi. 2000. Diversity digitized. Science 289:2305.

Sutherland, I., J. Robinson, S. M. Brandt, A. C. Jones, S. M. Embury, W. A. Gray, R. J. White, and F. A. Bisby. Year. Assisting the integration of taxonomic data: The LITCHI toolkit *in* 16th International Conference on Data Engineering (ICDE 2000). IEEE Computer Society Press.

Veregin, H. 1998. Data Quality Measurement and Assessment, NCGIA Core Curriculum in GISciene. University of Minnesota, Department of Geography, http://www.ncgia.ucsb.edu/giscc/units/u100/u100.html, posted March 23, 1998.

Vit, D. H., D. G. Horton, and P. Johnston. 1977. A computer program for printing herbarium labels. Taxon 26:425-428.

Vogellehner, D., and T. Speck. 1988. DIDEA-FR. A computer program system for Botanic Gardens. Taxon 37:876-884.

Wand, Y., and R. Y. Wang. 1996. Anchoring Data Quality Dimensions in Ontological Foundations. Communications of ACM 39:86-95.

Wang, R. Y., M. P. Reddy, and H. B. Kon. 1995a. Toward quality data: An attribute-based approach. Decision Suport System 13:349-372.

Wang, R. Y., V. C. Storey, and C. P. Firth. 1995b. A framework for analysis of data quality research. IEEE Transactions on Knowledge and Data Engineering 7:623-639.

Watson, L. 1971. Basic taxonomic data: The need for organisation over presentation and accumulation. Taxon 20:131-136.

Wetmore, C. M. 1979. Herbarium Computerization at the University of Minnesota. Systematic Botany 4:339-350.

Whalwn, A. (ed) 1993. HISPID - Herbarium Information Standards and Protocols for Interchange of Data - Version 2. National Herbarium of New South Wales, Sydney.

White, R. J., R. Allkin, and P. J. Winfield. 1993. Systematic Databases: The BAOBAB Design and the ALICE System. Pages 298-311 *in* Advances in Computer Methods for Systematic Biology: Artificial Intelligence, Databases, Computer Vision (R. Fortuner, ed.), Baltimore, Maryland.

Wilson, E. O. 2000. A Global Biodiversity Map. Science 289:2279-.

Wilson, H. D. 2001. Informatics: new media and paths of data flow. Taxon 50:381-337.

Wood Jr., C. E., R. S. Cowan, and G. Buchheim. 1963. Botanical nomenclature, punched cards, and machines. Taxon 12:2-12.

Zaïane, O. R. 1999. Introduction to Data Mining. Pages 15 *in* Principles of Knowledge Discovery in Databases University of Alberta, Department of Computing Science.

Zamora, E., J. J. Pollock, and A. Zamora. 1981. The use of trigram analysis for spelling error detection. Information and Processing Management 17:305-316.

Zarucchi, J. L. 1991. ILDIS: the world's legume species diversity project*in* Designs for a Global Plant Species Information System (F. A. Bisby, R. J. Pankhurst, and G. F. Russell, eds.). Oxford Univ. Press.

Zarucchi, J. L., P. J. Winfield, R. M. Polhill, S. Hollis, and F. A. Bisby. 1993. The ILDIS Project on the wold's legume species diversity. Pages 131-144 *in* Design for a global plant species information system (F. A. Bisby, G. F. Russel, and R. J. Pankhurst, eds.). The Systematics Association Special Volume No. 48, Oxford.

Zhong, Y., S. Jung, S. Pramanik, and J. H. Beaman. 1996. Data Model and Comparison and Query Methods for Interacting Classifications in a Taxonomic Database. Taxon 45.

Zobel, J., and P. Dart. Year. Phonetic String Matching: Lessons from Information Retrieval *in* 19th International Conference on Research and Development in Information Retrieval. ACM Press, Zurich, Switzerland:166-172.

Appendix I - List of species with invalid characters

Glossary:

POS – Position of error
S – Species
G – Genus
T – True for the Database
CNIP - Brazilian Northeast checklist Database
PMA – Brazilian Atlantic Rain Forest Database
ILDIS – International Legume Database & Information Service Database
MARINE - Marine Invertebrates of New Zealand Database
SP2K – Species 2000 Database

GENUS	SPECIES	POS	CNIP	PMA	ILDIS	MARINE	SP2K
Abrus	sp.A	S			T		
Abrus	sp.	S			T		
Acacia	sp.133	S			T		
Acacia	sp.2	S			T		
Acacia	sp.C	S			T		
Acacia	sp.D	S			T		
Acacia	sp.F	S			T		
Acacia	sp.132	S			T		
Acacia	sp.131	S			T		
Acacia	cf. langsdorfii	S	T				
Acacia	sp.E	S			T		
Acacia	sp. aff. riparia	S	T				
Acetobacter (subgen.Acetobacte	aceti	G					T
Acetobacter (subgen.Gluconoace	liquefaciens	G					T
Adenocalymma	aff.reticulatum	S		T			
Aeschynomene	sp.B	S			T		
Aeschynomene	sp.F	S			T		
Aeschynomene	sp.G	S			T		
Aeschynomene	sp.nr.bella	S			T		
Aeschynomene	sp.E	S			T		
Aeschynomene	sp.D	S			T		
Aeschynomene	sp.aff.mimosifolia	S			T		
Aeschynomene	sp.C	S			T		
Aeschynomene	sp.A	S			T		
Aeschynomene	sp.H	S			T		
Alysicarpus	sp.A	S			T		
Amorpha	sp.1	S			T		
Ampharetid	n.d.	S				T	
Amphimas	sp.	S			T		
Angostura	cf.odoratissima	S		T			
Angylocalyx	sp.	S			T		
Annona	aff. montana	S	T				
Anthonotha	sp.A	S			T		
Argyrolobium	spp.1	S			T		
Aricidea	Allia-Sp.	S				T	
Aricidea	Acesta-Sp.2	S				T	
Aspidosperma	cf. illustre	S	T				

Aspilia	cf. cupulata	S	T				
Asplundia	aff.maximiliani	S		T			
Astragalus	sp.1	S			T		
Athenaea	sp. 1	S	T				
Axiopsis	n.sp	S				T	
Baccharis	cf. rivularis	S	T				
Balanus (Megabalanus)	linzei	G				T	
Balanus (Solidobalanus)	auricoma	G				T	
Baphia	sp.B	S			T		
Bauhinia	sp. 1.	S	T				
Beania	inermis cryptophragma	S				T	
Berlinia	sp.1	S			T		
Berlinia	sp.	S			T		
Blechnum	cf.regelianum	S		T			
Bolusia	sp.	S			T		
Brachystegia	sp.cf.bakeriana	S			T		
Brachystegia	sp.nr.russelliae	S			T		
Calliotropis	pelseneeri rossiana	S				T	
Calophaca	sp.1	S			T		
Camptosema	aff.coccineum and pedicellatum	S	T				
Camptosema	sp. aff. coriaceum	S	T				
Cassia	sp.A	S			T		
Cassia	sp.B	S			T		
Cayaponia	cf.pilosa	S		T			
Cayaponia	cf.tayuya	S		T			
Cecropia	cf.lyratiloba	S		T			
Ceradocus	rubromaculatus haumuri	S				T	
Ceradocus	rubromamulatus haumuri	S				T	
Cestrum	cf. viminale	S	T				
Cestrum	aff.sessiliflorum	S		T			
Chamaecrista	cf. nictitans	S	T				
Chaperiopsis (Clipeochaper	chathamensis	G				T	
Chesneya	sp.1	S			T		
Chlamys	gemmulata radiata	S				T	
Chryso-Hypnum	diminutivum	G		T			
Chusquea	aff.tenella	S		T			
Chusquea	aff.oxylepis	S		T			
Coelorhynchus (Oxymacrurus	bollonsi	G				T	
Coelorhynchus (Oxymacrurus	cookianus	G				T	
Colossendeis	megalonyx megalonyx	S				T	
Cominella	adspersa melo	S				T	
Conyza	cf. bonariensis	S	T				
Copaifera	sp.1	S			T		
Cordyla	sp.A	S			T		
Cordyla	sp.C	S			T		
Cordyla	sp.B	S			T		
Cordyla	sp.	S			T		
Costus	aff.spiralis	S		T			
Crotalaria	cf. incana	S	T				
Crotalaria	sp. aff. breviflora	S	T				
Crotalaria	sp.	S			T		
Crudia	sp.A	S			T		

Cryptocarya	aff.hamosa	S		T			
Cupania	sp. near paniculata	S	T				
Cupania	cf. revoluta	S	T				
Cupania	sp. near oblongifolia	S	T				
Cymatona	kampyla delli	S				T	
Cynometra	sp.A	S			T		
Cynometra	sp.B	S			T		
Cyperus	aff.meyenianus	S		T			
Dalbergia	sp.1	S			T		
Dalbergia	sp.A	S			T		
Dalbergia	aff. catingicola	S	T				
Dalbergia	sp.B	S			T		
Dalbergia	sp.nr.macrosperma	S			T		
Dalbergia	sp.nr.pachycarpa	S			T		
Dallina	new.sp.	S				T	
Daniellia	sp.	S			T		
Derris	sp.2	S			T		
Derris	sp.1	S			T		
Diacranthera	cf. crenata	S	T				
Dialium	sp.	S			T		
Dichorisandra	cf.pubescens	S		T			
Dichrostachys	sp.A	S			T		
Didelotia	sp.	S			T		
Dioclea	sp.11722	S			T		
Dioclea	sp.9068	S			T		
Ditassa	cf. oxyphylla	S	T				
Dolichos	sp.B	S			T		
Dolichos	sp.C	S			T		
Dolichos	sp.A	S			T		
Dorstenia	aff.turneraefolia	S		T			
Droogmansia	cf.whytei	S			T		
Elephantorrhiza	sp.1	S			T		
Emarginula	n.sp	S				T	
Erythrina	sp.2	S			T		
Erythrina	sp.cf.E.buesgenii	S			T		
Erythrina	sp.A	S			T		
Erythrina	sp.1	S			T		
Erythrophleum	sp.1	S			T		
Estea	subfusca aupouria	S				T	
Eunice	n.d.	S				T	
Fenestrulina	malusii pulchra	S				T	
Fenestrulina	malusii incompta	S				T	
Galactia	aff. remansoana	S	T				
Galactia	sp.1	S			T		
Galega	sp.1	S			T		
Gaylussacia	aff.fasciculata	S		T			
Geonoma	aff.luetzelburghii	S		T			
Geonoma	aff.fiscellaria	S		T			
Gilbertiodendron	sp.aff.G.dewevrei	S			T		
Gilbertiodendron	sp.A	S			T		
Globocassidulina	crassa rossensis rossens	S				T	
Glycyrrhiza	sp.1	S			T		

Goniada	aff. brunnea	S				T	
Guatteria	cf. candolleana	S	T				
Gymnangium	gracilicaule/armatum	S				T	
Gyrothyris	forma.chathamensis	S				T	
Haematoxylum	sp.1.	S			T		
Heliconia	aff.lacletteana	S		T			
Heliconia	aff.spatho-circinada	S		T			
Hippothoa	divaricata pacifica	S				T	
Holothuria (Halodeima)	atra	G				T	
Holothuria (Halodeima)	edulis	G				T	
Holothuria (Lessonothuria)	paradalis	G				T	
Holothuria (Mertensiothuri	dofleini	G				T	
Holothuria (Mertensiothuri	leucospilota	G				T	
Holothuria (Mertensiothuri	pervicax	G				T	
Holothuria (Microthele)	nobilis	G				T	
Holothuria (Microthele)	fuscogilva	G				T	
Holothuria (Platyperona)	difficilis	G				T	
Holothuria (Stauropora)	dofleini	G				T	
Holothuria (Stauropora)	pervicax	G				T	
Holothuria (Thymiosycia)	impatiens	G				T	
Holothuria (Thymiosycia)	arenicola	G				T	
Holothuria (Thymiosycia)	hilla	G				T	
Holothuria (Vaneyothuria)	uncia	G				T	
Holothuria (Vaneyothuria)	neozelanica	G				T	
Humularia	sp.nov.aff.kassneri	S			T		
Huperzia	cf.dichotoma	S		T			
Hyalinoecia	aff. tubicola	S				T	
Hyalinoecia	tubicola longibranchiata	S				T	
Hymenocarpos	sp.1	S			T		
Hymenophyllum	cf.polyanthos	S		T			
Hymenostegia	sp.	S			T		
Hypolytrum	aff.schraderianum	S		T			
Hypopterygium	aff.tamariscinum	S		T			
Ichnanthus	cf. riedelii	S	T				
Ilex	cf. brevicuspis	S	T				
Indigofera	sp.aff.amitinae	S			T		
Indigofera	sp.aff.volkensii	S			T		
Indigofera	sp.C	S			T		
Indigofera	sp.3234	S			T		
Indigofera	sp.A	S			T		
Indigofera	sp.94	S			T		
Indigofera	sp.93	S			T		
Indigofera	sp.92	S			T		
Isocladus	Isocladus sp.	S				T	
Justicia	cf. thunbergioides	S	T				
Kotschya	sp.A	S			T		
Lathyrus	sp.1	S			T		
Lathyrus	spp.1	S			T		
Leandra	cf.laxa	S		T			
Leandra	cf.gracilis	S		T			
Leandra	cf.sylvestris	S		T			
Lenibiana [sic]	tubicila	G				T	

Lepidophyllum	cf.radicale	S		T			
Leucaena	sp.1	S			T		
Licaria	aff.reitzkleiniana	S		T			
Limatula	japonica delli	S				T	
Limatula	powelli powelli	S				T	
Lonchocarpus	sp.2	S			T		
Lonchocarpus	sp.1	S			T		
Lotononis	spec.	S			T		
Lotus	sp.1	S			T		
Machaerium	sp.1	S			T		
Machaerium	cf. angustifolium	S	T				
Maldane	sarsi antarctica	S				T	
Maoricolpus	roseus manaukauensis	S				T	
Maoricolpus	roseus manukauensis	S				T	
Marlierea	aff.teuscheriana	S		T			
Medicago	sp.1	S			T		
Merostachys	aff.ternata	S		T			
Meteuthria	multituberculata rossian	S				T	
Miconia	aff.ibaguensis	S		T			
Miconia	cf.jucunda	S		T			
Microgramma	cf.crispata	S		T			
Mikania	aff.myriantha	S		T			
Mikania	cf. hemisphaerica	S	T				
Millettia	sp.A	S			T		
Millettia	sp.3	S			T		
Millettia	sp(leucantha)	S			T		
Millettia	sp.1	S			T		
Millettia	sp.2	S			T		
Mimosa	aff. adenophylla	S	T				
Monopetalanthus	sp.C	S			T		
Monopetalanthus	sp.B	S			T		
Monopetalanthus	sp.	S			T		
Moroteuthis	lönnbergii	S					T
Mucuna	sp.3	S			T		
Mucuna	sp.2	S			T		
Mundulea	sp.1	S			T		
Nectandra	aff.leucantha	S		T			
Neilo	wairoana delli	S				T	
Nidularium	aff.rosulatum	S		T			
Ocotea	cf.velloziana	S		T			
Oncidium	aff.lietzei	S		T			
Ononis	sp.1	S			T		
Onykia	appellöfi	S					T
Ophiactis	abyssicola var.cuspidata	S				T	
Ormocarpum	sp.5	S			T		
Ormocarpum	sp.B	S			T		
Ormocarpum	sp.6	S			T		
Oxystigma	sp.	S			T		
Palicourea	aff.mansoana	S		T			
Panicum	aff.nervosum	S		T			
Perknaster	sladeni sladeni	S				T	
Persea	aff.venosa	S		T			

Persea	cf.pyrifolia	S		T			
Persea	aff.americana	S		T			
Phonicosia	circinata.	S				T	
Phylo	felix Kinberg	S				T	
Piptadenia	cf. viridiflora	S	T				
Pleurothallis	aff.hamosa	S		T			
Polhillia	sp.A	S			T		
Pouteria	aff.microstrigosa	S		T			
Prockia	crucis.	S	T				
Prosopis	sp.	S			T		
Pteraster	stellifer hunteri	S				T	
Pterocarpus	sp.1	S			T		
Pueraria	sp.4	S			T		
Pueraria	sp.3	S			T		
Pueraria	sp.2	S			T		
Pueraria	spp.1	S			T		
Pueraria	sp.1	S			T		
Quesnelia	cf.lateralis	S		T			
Rhynchosia	sp.A	S			T		
Rhynchosia	sp.nov.	S			T		
Rhynchosia	sp.	S			T		
Rhynchosia	sp.B	S			T		
Ruellia	cf. geminiflora	S	T				
Sabicea	aff.cinerea	S		T			
Salvinia	aff.auriculata	S		T			
Schotia	sp.	S			T		
Scoloplos	marginatus mcleani	S				T	
Scorpiurus	sp.1	S			T		
Scutochaperia	serrata biporosa	S				T	
Sebastiania	aff.klotzschiana	S		T			
Securigera	sp.1	S			T		
Seila (Hebeselia)	regia	G				T	
Sepia	appellöfi	S					T
Serjania	cf.sylvestris	S		T			
Serjania	sp. cf. salzmanniana	S	T				
Serjania	cf. paradoxa	S	T				
Sicydium	cf.deltoides	S		T			
Solanum	aff. distichophyllum	S	T				
Solanum	aff.schizandrum	S		T			
Solanum	aff.glomuliflorum	S		T			
Solanum	aff.convolvulus	S		T			
Solaster	aff.japonicus	S				T	
Sophronitis	aff.grandiflora	S		T			
Sophronitis	aff.mantiqueirae	S		T			
Sorocea	aff.hilarii	S		T			
Sorocea	aff.guilleminiana	S		T			
Sphagnum	aff.erythrocalyx	S		T			
Sutherlandia	sp.1	S			T		
Talisia	cf. guianensis	S	T				
Tephrosia	sp.A	S			T		
Tephrosia	sp.1	S			T		
Terebratulina	n.sp	S				T	

Tontelea	aff.riedeliana	S		T			
Tosarhombus	n.sp	S				T	
Trachypogon	Ness.	S	T				
Travisia	olens novaezealandiae	S				T	
Trichilia	aff.venosa	S		T			
Trifolium	sp.A	S			T		
Unonopsis	aff. stipitata	S	T				
Vernonia	aff.puberula	S		T			
Vernonia	cf. scorpioides	S	T				
Vicia	sp.A	S			T		
Vigna	sp.C	S			T		
Vigna	sp.B	S			T		
Wedelia	cf. villosa	S	T				
Xylopia	cf. frutescens	S	T				
Xylopia	aff. lanceolata	S	T				
Zornia	?gemella	S	T				

Appendix II - List of “wrong letter” and “transposition” errors detected

```
----------------------------------
Wrong Letter
----------------------------------
A <-> L
Aotus mollis
Lotus mollis

Total: 1
----------------------------------
C <-> D
Ciphysa robinioides
Diphysa robinioides

Total: 1
----------------------------------
C <-> G
Codonophilus imbricatus
Godonophilus imbricatus

Total: 1
----------------------------------
H <-> K
Holostoneura novaezelandiae
Kolostoneura novaezelandiae

Total: 1
----------------------------------
I <-> J
Ianthina exigua
Janthina exigua

Total: 1
----------------------------------
a <-> c
Daviesia striata
Daviesia stricta

Trifolium striatum
Trifolium strictum

Total: 2
----------------------------------
a <-> e
Centrolobium microchaeta
Centrolobium microchaete

Sida galhairensis
Sida galheirensis

Escharoides praestita
Escharoides praestite

Eunice tridentata
Eunice tridentate

Globigerinella aequilateralis
Globigerinella aequilaterelis
```

Glycera tessalata
Glycera tesselata

Hyala rubra
Hyale rubra

Malluvium calcareus
Malluvium calcereus

Melarhapha cincta
Melarhaphe cincta

Melarhapha oliveri
Melarhaphe oliveri

Metavargula mazari
Metavargula mazeri

Parawaldeckia parata
Paraweldeckia parata

Parawaldeckia thomsoni
Paraweldeckia thomsoni

Paridotea ungulata
Paridotea ungulate

Soletellina siliqua
Soletellina silique

Total: 15
---
a <-> h
Brachioteuthis beani
Brachioteuthis behni

Total: 1
---
a <-> i
Jacaranda macrantha
Jacaranda micrantha

Bauhinia macrostachya
Bauhinia microstachya

Calliandra macrocalyx
Calliandra microcalyx

Lantana macrophylla
Lantana microphylla

Chlamys kiwaensis
Chlamys kiwiensis

Pagurus spanulimanus
Pagurus spinulimanus

Paramoera rangatira
Paramoera rangitira

Resania lanceolata
Resinia lanceolata

Sertella fragida
Sertella frigida

Argyrolobium macrophyllum
Argyrolobium microphyllum

Crotalaria macrocarpa
Crotalaria microcarpa

Detarium macrocarpum
Detarium microcarpum

Diphysa macrophylla
Diphysa microphylla

Indigofera macrantha
Indigofera micrantha

Indigofera macrocalyx
Indigofera microcalyx

Kennedia macrophylla
Kennedia microphylla

Machaerium macrophyllum
Machaerium microphyllum

Mimosa macrocephala
Mimosa microcephala

Rhynchosia macrantha
Rhynchosia micrantha

Sclerolobium macropetalum
Sclerolobium micropetalum

Swartzia macrocarpa
Swartzia microcarpa

Tephrosia macrantha
Tephrosia micrantha

Trifolium macrocephalum
Trifolium microcephalum

Astronesthes macropogon
Astronesthes micropogon

Bathytroctes macrolepis
Bathytroctes microlepis

Curimatopsis macrolepis
Curimatopsis microlepis

Imparfinis macrocephala
Imparfinis microcephala

Moringua macrochir

Moringua microchir

Rhynchactis macrothrix
Rhynchactis microthrix

Total: 29
---
a <-> m
Mycoplasma auris
Mycoplasma muris

Total: 1
---
a <-> o
Calea villosa
Colea villosa

Calopogonium caeruleum
Calopogonium coeruleum

Lisianthius caerulescens
Lisianthius coerulescens

Manihot caerulescens
Manihot coerulescens

Simaba maiana
Simaba moiana

Solanum caavurana
Solanum coavurana

Sparattanthelium botocudarum
Sparattanthelium botocudorum

Astrononian novozealandicum
Astrononion novozealandicum

Aulacomya maoriana
Aulocomya maoriana

Lumbrineris aotearoae
Lumbrineris aoteoroae

Mallacoota subcarinata
Mallocoota subcarinata

Modiolus arealatus
Modiolus areolatus

Modiolus areolatus
Modiolus areolotus

Navicula trampii
Navicula trompii

Notostamus westergreni
Notostomus westergreni

Pseudobaletia indiana
Pseudoboletia indiana

Pycnogonum rhinoceras
Pycnogonum rhinoceros

Sigmadocia glacialis
Sigmodocia glacialis

Total: 18
---
a <-> q
Leptochiton inquinatus
Leptochiton inquinqtus

Total: 1
---
a <-> s
Tawera spisa
Tawera spiss

Inga brachystachya
Inga brachystachys

Borreria acabiosoides
Borreria scabiosoides

Cereus aquamosus
Cereus squamosus

Paraphoxus waipiro
Parsphoxus waipiro

Total: 5
---
a <-> u
Notopandalus magnoculus
Notopandulus magnoculus

Parapeneus bifasciatus
Parupeneus bifasciatus

Phoxocephalas regium
Phoxocephalus regium

Zegalerus tenuis
Zegulerus tenuis

Astragalus laristanicus
Astragalus luristanicus

Crotalaria lanata
Crotalaria lunata

Lupinus barkeri
Lupinus burkeri

Machaerium lanatum
Machaerium lunatum

Haplochromis bartoni
Haplochromis burtoni

```
Total: 9
---------------------------------
a <-> y
Pyrgo murrhana
Pyrgo murrhyna

Total: 1
---------------------------------
a <-> z
Pecten novaezelandiae
Pecten novzezelandiae

Total: 1
---------------------------------
b <-> g
Genypterus blacodes
Genypterus glacodes

Scutus breviculus
Scutus greviculus

Total: 2
---------------------------------
b <-> h
Sphaeroidinella debiscens
Sphaeroidinella dehiscens

Total: 1
---------------------------------
b <-> k
Adamussium colbecki
Adamussium colkecki

Total: 1
---------------------------------
b <-> l
Clidemia bulbosa
Clidemia bullosa

Total: 1
---------------------------------
b <-> n
Fissidentalium zelabdicum
Fissidentalium zelandicum

Total: 1
---------------------------------
b <-> p
Astragalus beckii
Astragalus peckii

Total: 1
---------------------------------
b <-> r
Astragalus harbisonii
Astragalus harrisonii

Total: 1
---------------------------------
b <-> s
Lanicides bilobata
```

Lanicides silobata

Total: 1
---
b <-> t
Mimosa brachycarpa
Mimosa trachycarpa

Lycenchelys alba
Lycenchelys alta

Total: 2
---
b <-> v
Romancheina barica
Romancheina varica

Total: 1
---
c <-> e
Archacopsis ramusculus
Archaeopsis ramusculus

Neobuccinum eatoni
Neobuceinum eatoni

Notoplax violacca
Notoplax violacea

Littorina coccinca
Littorina coccinea

Maurca blacki
Maurea blacki

Total: 5
---
c <-> g
Actaecia euchroa
Actaecia eughroa

Lafoensia clyptocarpa
Lafoensia glyptocarpa

Total: 2
---
c <-> m
Metrodorea caracasana
Metrodorea maracasana

Total: 1
---
c <-> n
Seguenzia cocopia
Seguenzia conopia

Senna cana
Senna nana

Tephrosia cana
Tephrosia nana

```
Total: 3
---------------------------------
c <-> p
Hybanthus ipecacuanha
Hybanthus ipepacuanha

Atherinella callida
Atherinella pallida

Total: 2
---------------------------------
c <-> q
Pallenopsis oblicua
Pallenopsis obliqua

Total: 1
---------------------------------
c <-> r
Leporinus octomaculatus
Leporinus ortomaculatus

Total: 1
---------------------------------
c <-> s
Calloria inconspicua
Calloria insonspicua

Penion aducta
Penion adusta

Paramyxine cheni
Paramyxine sheni

Total: 3
---------------------------------
c <-> v
Astacilla levis
Astavilla levis

Total: 1
---------------------------------
d <-> g
Licania ridida
Licania rigida

Total: 1
---------------------------------
d <-> h
Paradexamine pacifica
Parahexamine pacifica

Total: 1
---------------------------------
d <-> j
Melamphaes danae
Melamphaes janae

Total: 1
---------------------------------
d <-> q
```

```
Piriqueta duarteana
Piriqueta quarteana

Total: 1
---------------------------------
d <-> r
Schistura prashadi
Schistura prashari

Total: 1
---------------------------------
d <-> s
Limatula hodgdoni
Limatula hodgsoni

Total: 1
---------------------------------
d <-> t
Dyckia dissidiflora
Dyckia dissitiflora

Chondrostoma duriense
Chondrostoma turiense

Total: 2
---------------------------------
e <-> f
Oxalis neuwiedii
Oxalis nfuwiedii

Total: 1
---------------------------------
e <-> i
Maytenus brasilienses
Maytenus brasiliensis

Lagenocarpus guianenses
Lagenocarpus guianensis

Piptadenia moniliformes
Piptadenia moniliformis

Stachytarpheta lychnites
Stachytarpheta lychnitis

Cylichna thetides
Cylichna thetidis

Heteromolpadia marenzelleri
Heteromolpadia marenzelliri

Micropora coreacea
Micropora coriacea

Plocamium cartilagineum
Plocamium cartilaginium

Trichophoxus capellatus
Trichophoxus capillatus

Total: 9
```

```
---------------------------------
e <-> l
Cuspidaria trailei
Cuspidaria trailli

Urothoe elizae
Urothoe elizal

Total: 2
---------------------------------
e <-> o
Kolestoneura novaezelandiae
Kolostoneura novaezelandiae

Mysidetes pesthon
Mysidetes posthon

Osthimosia milleporoides
Osthimosia milloporoides

Sertella hippocrepis
Sertella hippocropis

Telepora digitata
Telopora digitata

Thyasira peroniana
Thyasira poroniana

Neolamprologus leleupi
Neolamprologus leloupi

Antisolarium egenum
Antisolarium ogenum

Total: 8
---------------------------------
e <-> s
Conus sponealis
Conus sponsalis

Total: 1
---------------------------------
e <-> t
Hirtella eriandra
Hirtella triandra

Total: 1
---------------------------------
e <-> u
Cylichna thetides
Cylichna thetidus

Millettia peguensis
Millettia puguensis

Total: 2
---------------------------------
f <-> t
Stauronereis incerfus
Stauronereis incertus
```

```
Laeviliforina caliginosa
Laevilitorina caliginosa

Total: 2
---------------------------------
f <-> v
Sacciolepis vilfoides
Sacciolepis vilvoides

Total: 1
---------------------------------
g <-> q
Philodendron propinguum
Philodendron propinquum

Adelascopora jegolqa
Adelascopora jeqolqa

Fenestrulina exigua
Fenestrulina exiqua

Terebratella sanguinea
Terebratella sanquinea

Vibilia propingua
Vibilia propinqua

Total: 5
---------------------------------
h <-> i
Astraea heliotrophum
Astraea heliotropium

Total: 1
---------------------------------
h <-> l
Andira handroana
Andira landroana

Total: 1
---------------------------------
h <-> n
Amalda bathami
Amalda batnami

Carahgolia puliciformis
Carangolia puliciformis

Comissia horfolkensis
Comissia norfolkensis

Microsetella horvegica
Microsetella norvegica

Scyphax orhatus
Scyphax ornatus

Total: 5
---------------------------------
i <-> j
```

```
Xymene plebeius
Xymene plebejus

Artediellus gomoiunovi
Artediellus gomojunovi

Total: 2
---------------------------------
i <-> l
Bauhinia vespertilio
Bauhinia vespertillo

Corycaeus fiaccus
Corycaeus flaccus

Rhombosolea piebeia
Rhombosolea plebeia

Total: 3
---------------------------------
i <-> n
Divaricella huttoniaia
Divaricella huttoniana

Total: 1
---------------------------------
i <-> o
Epiginichthys hectori
Epigonichthys hectori

Macrochiridothea uncinata
Macrochirodothea uncinata

Monodilepas monilifera
Monodilepas monolifera

Pallium cinvexum
Pallium convexum

Planispirinoides bucculenta
Planispironoides bucculenta

Torridiharpinia hurleyi
Torridoharpinia hurleyi

Urothoe wellingtonensis
Urothoe wellongtonensis

Total: 7
---------------------------------
i <-> r
Capiomimus abbreviatus
Capromimus abbreviatus

Total: 1
---------------------------------
i <-> s
Tephrosia clementii
Tephrosia clementis

Total: 1
```

```
----------------------------------
i <-> t
Phonicosia circinata
Phonicosta circinata

Synodontis budgetii
Synodontis budgetti

Total: 2
----------------------------------
i <-> u
Sparattanthelium tipiniquinorum
Sparattanthelium tupiniquinorum

Amphithyris bispinosus
Amphithyrus bispinosus

Argobuccinum timidum
Argobuccinum tumidum

Balanus trigonis
Balanus trigonus

Clibanarius humilis
Clibanarius humilus

Cylichna thetidis
Cylichna thetidus

Haliris setosa
Halirus setosa

Hemiaulis hauckii
Hemiaulus hauckii

Hemiaulis sinensis
Hemiaulus sinensis

Prionocrangon curvicaulis
Prionocrangon curvicaulus

Soletellina siliqua
Soletellina suliqua

Trochus viridis
Trochus viridus

Total: 12
----------------------------------
i <-> y
Eugenia tinguiensis
Eugenia tinguyensis

Acacia piauhiensis
Acacia piauhyensis

Aeschynomene histrix
Aeschynomene hystrix

Bauhinia cuiabensis
Bauhinia cuyabensis
```

```
Beyrichia ocimoides
Beyrichia ocymoides

Clitoria guianensis
Clitoria guyanensis

Hieronima oblonga
Hyeronima oblonga

Maprounea guianensis
Maprounea guyanensis

Mouriri guianensis
Mouriri guyanensis

Pourouma guianensis
Pourouma guyanensis

Psidium guaiava
Psidium guayava

Pycreus polistachyos
Pycreus polystachyos

Rapanea guianensis
Rapanea guyanensis

Stylosanthes guianensis
Stylosanthes guyanensis

Capparis ico
Capparis yco

Vicia silvatica
Vicia sylvatica

Total: 16
----------------------------------
j <-> k
Cellarinella rogicjae
Cellarinella rogickae

Total: 1
----------------------------------
j <-> t
Sertella projecta
Sertella protecta

Total: 1
----------------------------------
k <-> l
Haurakia huttoni
Hauralia huttoni

Total: 1
----------------------------------
k <-> s
Pimelodella meeki
Pimelodella meesi
```

```
Total: 1
---------------------------------
l <-> m
Streptomyces albofaciens
Streptomyces ambofaciens

Total: 1
---------------------------------
l <-> r
Streptomyces limosus
Streptomyces rimosus

Total: 1
---------------------------------
l <-> t
Lyreidus tridentalus
Lyreidus tridentatus

Mytilus edulis
Mytilus edutis

Astragalus coltonii
Astragalus cottonii

Astragalus loanus
Astragalus toanus

Total: 4
---------------------------------
l <-> v
Inga lallensis
Inga vallensis

Total: 1
---------------------------------
m <-> n
Epidendrum cinnabarimum
Epidendrum cinnabarinum

Corbulella spinosissima
Corbulella spinosissina

Dosinia lambata
Dosinia lanbata

Escalima regularis
Escalina regularis

Trematomus loennbergi
Trematonus loennbergi

Lithophaga masuta
Lithophaga nasuta

Total: 6
---------------------------------
m <-> s
Solanum stipulaceum
Solanum stipulaceus

Tripogon spicatum
```

Tripogon spicatus

Arcoscalpellum pedunculatum
Arcoscalpellum pedunculatus

Fusitriton magellanicum
Fusitriton magellanicus

Malluvium calcareum
Malluvium calcareus

Cichlasoma ornatum
Cichlasoma ornatus

Chaetoceros secundum
Chaetoceros secundus

Total: 7
---
m <-> u
Corella emmyota
Corella eumyota

Total: 1
---
n <-> p
Mimosa nigra
Mimosa pigra

Total: 1
---
n <-> r
Heteronandoa carinata
Heteronardoa carinata

Maurea turneranum
Maurea turnerarum

Miconia nimalis
Miconia rimalis

Portulaca elation
Portulaca elatior

Isocladus magellanicus
Isocladus magellaricus

Notoplites vanhoffeni
Notoplites vanhofferi

Macrochiridotea uncinata
Macrochiridotea urcinata

Orchestia waipuna
Orchestia waipura

Parawaldeckia schellenbergi
Parawaldeckia schellerbergi

Limanda beani
Limanda beari

Total: 10
---
n <-> u
Combretum anfractuosum
Combretum aufractuosum

Lupinus sericens
Lupinus sericeus

Landeria annulata
Lauderia annulata

Clausocalanus ingens
Clausocalanus ingeus

Isognomon perna
Isognomon perua

Total: 5
---
o <-> r
Aglaophamus macoura
Aglaophamus macrura

Total: 1
---
o <-> u
Notopandulus magnocolus
Notopandulus magnoculus

Astragalus ankylotos
Astragalus ankylotus

Tilapia cameronensis
Tilapia camerunensis

Total: 3
---
p <-> r
Cantharidus capilaceus
Cantharidus carilaceus

Total: 1
---
p <-> s
Machaerium paraense
Machaerium saraense

Total: 1
---
r <-> s
Tetraclitella purpurarcens
Tetraclitella purpurascens

Total: 1
---
r <-> t
Campylaspis antarctica
Campylaspis antatctica

Leptomya retiaria
Leptomya retiatia

Cellaria aurorae
Cellaria aurotae

Smittina crenocondyla
Smittina ctenocondyla

Astragalus amarus
Astragalus amatus

Streptomyces violarus
Streptomyces violatus

Total: 6
---
r <-> v
Hippomonarella flexuosa
Hippomonavella flexuosa

Sthenolepis laeris
Sthenolepis laevis

Total: 2
---
s <-> t
Lupinus soratensis
Lupinus toratensis

Total: 1
---
s <-> z
Gallesia gorasema
Gallesia gorazema

Centrosema brasilianum
Centrosema brazilianum

Total: 2
---
t <-> v
Astragalus curtipes
Astragalus curvipes

Total: 1
---
t <-> y
Triptertgion varium
Tripterygion varium

Total: 1
---
u <-> y
Fleurya aestuans
Fleurya aestyans

Rhipsalis cassutha
Rhipsalis cassytha

Microcytherura crassa

Microcytheryra crassa

Total: 3
---
v <-> y
Alveolophragmium jeffrevsi
Alveolophragmium jeffreysi

Total: 1

Transposition
---
ac <-> ca
Cadulus deliactulus
Cadulus delicatulus

Ciliacea dolorosa
Cilicaea dolorosa

Onacea conifera
Oncaea conifera

Total: 3
---
ae <-> ea
Ebalia laevis
Ebalia leavis

Pholadidea acherontae
Pholadidea acherontea

Total: 2
---
ag <-> ga
Longimactra elonagta
Longimactra elongata

Total: 1
---
ai <-> ia
Macomona lilaina
Macomona liliana

Tellina lilaina
Tellina liliana

Trochosodon mosaicus
Trochosodon mosiacus

Total: 3
---
al <-> la
Atrina zealndica
Atrina zelandica

Barbatia novaezealndiae
Barbatia novaezelandiae

Chalmys celator
Chlamys celator

Chalmys taiaroa
Chlamys taiaroa

Endeis australis
Endeis austrlais

Lucicutia falvicornis
Lucicutia flavicornis

Monia zealndica
Monia zelandica

Pecten novaezealndiae
Pecten novaezelandiae

Sigapatella novaezealndiae
Sigapatella novaezelandiae

Triplophysa alticeps
Triplophysa laticeps

Total: 10
---
ar <-> ra
Ommatocarcinus macgillivaryi
Ommatocarcinus macgillivrayi

Total: 1
---
au <-> ua
Ostrea sinnauta
Ostrea sinnuata

Total: 1
---
cr <-> rc
Nectocarcinus antacrticus
Nectocarcinus antarcticus

Hiatella acrtica
Hiatella arctica

Macropora grandis
Marcopora grandis

Total: 3
---
ei <-> ie
Mesosagitta decipeins
Mesosagitta decipiens

Total: 1
---
el <-> le
Styela plicata
Stylea plicata

Trachyleberis lytteltonensis
Trachyleberis lyttletonensis

Total: 2
---
en <-> ne
Biddulphia mobiliensis
Biddulphia mobilinesis

Total: 1
---
er <-> re
Modiolus aerolatus
Modiolus areolatus

Total: 1
---
es <-> se
Cirolana woodjonesi
Cirolana woodjonsei

Total: 1
---
hp <-> ph
Lumbrineris sphaerocehpala
Lumbrineris sphaerocephala

Total: 1
---
il <-> li
Pulleniatina obliquiloculata
Pulleniatina obliquliоculata

Total: 1
---
il <-> pi
Cantharidus capillaceus
Cantharidus cappilaceus

Total: 1
---
in <-> ni
Terebratella sanguinea
Terebratella sanguniea

Total: 1
---
io <-> oi
Fusitriton retiolus
Fusitriton retoilus

Total: 1
---
ir <-> ri
Neoguraleus sinclairi
Neoguraleus sinclarii

Total: 1
---
is <-> si
Harpecia spinosissima
Harpecia spinossisima

Osthimosia notialis

Osthimosia notialsi

Otionella australis
Otionella australsi

Total: 3
---
it <-> ti
Apseudes seitferus
Apseudes setiferus

Glycymeris laitcostata
Glycymeris laticostata

Total: 2
---
iu <-> ui
Syzygium jambolanum
Syzyguim jambolanum

Total: 1
---
lo <-> ol
Poroleda lanceloata
Poroleda lanceolata

Total: 1
---
lp <-> pl
Scalpomactra scalpellum
Scalpomactra scaplellum

Total: 1
---
lu <-> ul
Cibicides reflugens
Cibicides refulgens

Total: 1
---
ny <-> yn
Periclimenes yaldwnyi
Periclimenes yaldwyni

Total: 1
---
os <-> so
Cirostrema zelebori
Cirsotrema zelebori

Total: 1
---
ou <-> uo
Pecten raoulensis
Pecten rauolensis

Total: 1
---
pr <-> rp
Emballotheca monomoprha
Emballotheca monomorpha

```
Total: 1
----------------------------------
ps <-> sp
Apseudes diversus
Aspeudes diversus

Apseudes spinifer
Aspeudes spinifer

Total: 2
----------------------------------
rt <-> tr
Gonimyrtea concinna
Gonimytrea concinna

Total: 1
----------------------------------
ru <-> ur
Balanus decorus
Balanus decours

Venericardia purpruata
Venericardia purpurata

Total: 2
----------------------------------
```

Appendix III - Final analysis reports of the spelling error experiments

a) Summarized report

This table shows the number of errors detected by each algorithm (G2, G3 etc.), both in total and classified by error type (EX, T etc.).

Glossary:

EX – Extra/Missing Letter
T – Transposition of Letters
W – Wrong Letter
MU – Multi-error
FA – False Alarm

```
**************************
51645 pair of names detected
**************************

-------------------------------------------
To G2 (bigram):
-------------------------------------------
Total Names to G2 :4063
Total Names to G2 @ EX :354
Total Names to G2 @ T :26
Total Names to G2 @ W :259
Total Names to G2 @ MU :200
Total Names to G2 @ FA :3224
-------------------------------------------
To G3 (trigram):
-------------------------------------------
Total Names to G3 :1980
Total Names to G3 @ EX :303
Total Names to G3 @ T :13
Total Names to G3 @ W :164
Total Names to G3 @ MU :119
Total Names to G3 @ FA :1381
-------------------------------------------
To SNDX (soundex):
-------------------------------------------
Total Names to SNDX :5008
Total Names to SNDX @ EX :285
Total Names to SNDX @ T :45
Total Names to SNDX @ W :181
Total Names to SNDX @ MU :224
Total Names to SNDX @ FA :4273
-------------------------------------------
To PHX (phonix):
-------------------------------------------
Total Names to PHX :11909
Total Names to PHX @ EX :302
Total Names to PHX @ T :42
Total Names to PHX @ W :188
Total Names to PHX @ MU :254
Total Names to PHX @ FA :11123
```

```
-------------------------------------------
To SKL (skeleton):
-------------------------------------------
Total Names to SKL :432
Total Names to SKL @ EX :226
Total Names to SKL @ T :39
Total Names to SKL @ W :45
Total Names to SKL @ MU :73
Total Names to SKL @ FA :49
-------------------------------------------
To LV (levenshtein):
-------------------------------------------
Total Names to LV :39150
Total Names to LV @ EX :367
Total Names to LV @ EX = 1 : 354
Total Names to LV @ EX = 2 : 10
Total Names to LV @ EX = 3 : 1
Total Names to LV @ EX = 4 : 2
Total Names to LV @ T :54
Total Names to LV @ T = 1 : 0
Total Names to LV @ T = 2 : 54
Total Names to LV @ T = 3 : 0
Total Names to LV @ T = 4 : 0
Total Names to LV @ W :276
Total Names to LV @ W = 1 : 276
Total Names to LV @ W = 2 : 0
Total Names to LV @ W = 3 : 0
Total Names to LV @ W = 4 : 0
Total Names to LV @ MU :301
Total Names to LV @ MU = 1 : 0
Total Names to LV @ MU = 2 : 238
Total Names to LV @ MU = 3 : 53
Total Names to LV @ MU = 4 : 10
Total Names to LV @ FA :38152
Total Names to LV @ FA = 1 : 0
Total Names to LV @ FA = 2 : 1178
Total Names to LV @ FA = 3 : 6726
Total Names to LV @ FA = 4 : 30248
===========================================
-------------------------------------------
Total to EX: 367
-------------------------------------------
-------------------------------------------
Total to T: 54
-------------------------------------------
-------------------------------------------
Total to W: 276
-------------------------------------------
-------------------------------------------
Total to MU: 301
-------------------------------------------
-------------------------------------------
Total to FA: 50647
-------------------------------------------
```

b) Detailed Report

The number of errors summarized in appendix III (a) are here listed for each database (CNIP, PMA etc.), and the name pairs themselves are also listed.

Glossary:

EX – Extra/Missing Letter
T – Transposition of Letters
W – Wrong Letter
MU – Multi-error
FA – False Alarm

G2 – Bigram
G3 – Trigram
SNDX – Soundex
PHX – Phonix
SKL – Skeleton
LV – Levenshtein

CNIP - Brazilian Northeast checklist Database
PMA – Brazilian Atlantic Rain Forest Database
ILDIS – International Legume Database & Information Service Database
SP2K – Species 2000 Database
MARINE - Marine Invertebrates of New Zealand Database

```
-------------------------------------------
To CNIP:
-------------------------------------------
Total Names :2642
Total Names to EX :38
Total Names to T :0
Total Names to W :53
Total Names to MU :43
Total Names to FA :2508

Total Names to G2 :260
Total Names to G3 :148
Total Names to SNDX :245
Total Names to PHX :526
Total Names to SKL :46
Total Names to LV :2227

Total Names to EX @ G2 :38
Total Names to EX @ G3 :37
Total Names to EX @ SNDX :29
Total Names to EX @ PHX :31
Total Names to EX @ SKL :23
Total Names to EX @ LV :38
Total Names to T @ G2 :0
Total Names to T @ G3 :0
Total Names to T @ SNDX :0
Total Names to T @ PHX :0
Total Names to T @ SKL :0
Total Names to T @ LV :0
Total Names to W @ G2 :50
Total Names to W @ G3 :34
Total Names to W @ SNDX :38
```

```
Total Names to W @ PHX :39
Total Names to W @ SKL :5
Total Names to W @ LV :53
Total Names to MU @ G2 :38
Total Names to MU @ G3 :25
Total Names to MU @ SNDX :35
Total Names to MU @ PHX :35
Total Names to MU @ SKL :7
Total Names to MU @ LV :43
Total Names to FA @ G2 :134
Total Names to FA @ G3 :52
Total Names to FA @ SNDX :143
Total Names to FA @ PHX :421
Total Names to FA @ SKL :11
Total Names to FA @ LV :2093

Total Names to EX @ G :3
Total Names to EX @ E :35
Total Names to EX @ N :0
Total Names to T @ G :0
Total Names to T @ E :0
Total Names to T @ N :0
Total Names to W @ G :3
Total Names to W @ E :50
Total Names to W @ N :0
Total Names to MU @ G :1
Total Names to MU @ E :42
Total Names to MU @ N :0
Total Names to FA @ G :916
Total Names to FA @ E :1592
Total Names to FA @ N :0

Total Names to EX @ G2 @ G :3
Total Names to EX @ G2 @ E :35
Total Names to EX @ G2 @ N :0
Total Names to EX @ G3 @ G :3
Total Names to EX @ G3 @ E :34
Total Names to EX @ G3 @ N :0
Total Names to EX @ SNDX @ G :3
Total Names to EX @ SNDX @ E :26
Total Names to EX @ SNDX @ N :0
Total Names to EX @ PHX @ G :3
Total Names to EX @ PHX @ E :28
Total Names to EX @ PHX @ N :0
Total Names to EX @ SKL @ G :3
Total Names to EX @ SKL @ E :20
Total Names to EX @ SKL @ N :0
Total Names to EX @ LV @ G :3
Total Names to EX @ LV @ E :35
Total Names to EX @ LV @ N :0
Total Names to T @ G2 @ G :0
Total Names to T @ G2 @ E :0
Total Names to T @ G2 @ N :0
Total Names to T @ G3 @ G :0
Total Names to T @ G3 @ E :0
Total Names to T @ G3 @ N :0
Total Names to T @ SNDX @ G :0
Total Names to T @ SNDX @ E :0
Total Names to T @ SNDX @ N :0
Total Names to T @ PHX @ G :0
Total Names to T @ PHX @ E :0
```

```
Total Names to T @ PHX @ N :0
Total Names to T @ SKL @ G :0
Total Names to T @ SKL @ E :0
Total Names to T @ SKL @ N :0
Total Names to T @ LV @ G :0
Total Names to T @ LV @ E :0
Total Names to T @ LV @ N :0
Total Names to W @ G2 @ G :2
Total Names to W @ G2 @ E :48
Total Names to W @ G2 @ N :0
Total Names to W @ G3 @ G :2
Total Names to W @ G3 @ E :32
Total Names to W @ G3 @ N :0
Total Names to W @ SNDX @ G :2
Total Names to W @ SNDX @ E :36
Total Names to W @ SNDX @ N :0
Total Names to W @ PHX @ G :2
Total Names to W @ PHX @ E :37
Total Names to W @ PHX @ N :0
Total Names to W @ SKL @ G :0
Total Names to W @ SKL @ E :5
Total Names to W @ SKL @ N :0
Total Names to W @ LV @ G :3
Total Names to W @ LV @ E :50
Total Names to W @ LV @ N :0
Total Names to MU @ G2 @ G :1
Total Names to MU @ G2 @ E :37
Total Names to MU @ G2 @ N :0
Total Names to MU @ G3 @ G :1
Total Names to MU @ G3 @ E :24
Total Names to MU @ G3 @ N :0
Total Names to MU @ SNDX @ G :1
Total Names to MU @ SNDX @ E :34
Total Names to MU @ SNDX @ N :0
Total Names to MU @ PHX @ G :1
Total Names to MU @ PHX @ E :34
Total Names to MU @ PHX @ N :0
Total Names to MU @ SKL @ G :0
Total Names to MU @ SKL @ E :7
Total Names to MU @ SKL @ N :0
Total Names to MU @ LV @ G :1
Total Names to MU @ LV @ E :42
Total Names to MU @ LV @ N :0
Total Names to FA @ G2 @ G :15
Total Names to FA @ G2 @ E :119
Total Names to FA @ G2 @ N :0
Total Names to FA @ G3 @ G :5
Total Names to FA @ G3 @ E :47
Total Names to FA @ G3 @ N :0
Total Names to FA @ SNDX @ G :45
Total Names to FA @ SNDX @ E :98
Total Names to FA @ SNDX @ N :0
Total Names to FA @ PHX @ G :185
Total Names to FA @ PHX @ E :236
Total Names to FA @ PHX @ N :0
Total Names to FA @ SKL @ G :2
Total Names to FA @ SKL @ E :9
Total Names to FA @ SKL @ N :0
Total Names to FA @ LV @ G :723
Total Names to FA @ LV @ E :1370
Total Names to FA @ LV @ N :0
```

```
Total Names to G2 unicos : 14
Total Names to G3 unicos : 0
Total Names to SNDX unicos : 44
Total Names to PHX unicos : 303
Total Names to SKL unicos : 1
Total Names to LV unicos : 1914

Total Names to G2@ EX unicos : 0
Total Names to G3@ EX unicos : 0
Total Names to SNDX@ EX unicos : 0
Total Names to PHX@ EX unicos : 0
Total Names to SKL@ EX unicos : 0
Total Names to LV@ EX unicos : 0
Total Names to G2@ T unicos : 0
Total Names to G3@ T unicos : 0
Total Names to SNDX@ T unicos : 0
Total Names to PHX@ T unicos : 0
Total Names to SKL@ T unicos : 0
Total Names to LV@ T unicos : 0
Total Names to G2@ W unicos : 0
Total Names to G3@ W unicos : 0
Total Names to SNDX@ W unicos : 0
Total Names to PHX@ W unicos : 0
Total Names to SKL@ W unicos : 0
Total Names to LV@ W unicos : 0
Total Names to G2@ MU unicos : 0
Total Names to G3@ MU unicos : 0
Total Names to SNDX@ MU unicos : 0
Total Names to PHX@ MU unicos : 0
Total Names to SKL@ MU unicos : 0
Total Names to LV@ MU unicos : 1
Total Names to G2@ FA unicos : 14
Total Names to G3@ FA unicos : 0
Total Names to SNDX@ FA unicos : 44
Total Names to PHX@ FA unicos : 303
Total Names to SKL@ FA unicos : 1
Total Names to LV@ FA unicos : 1913

Total Names detected in all tests : 31

Total Names detected in all tests @ EX : 22
 Agonandra brasilensis
 Agonandra brasiliensis

 Andropogon leucostachys
 Andropogon leucostachyus

 Cleome diffusa
 Cleome difusa

 Coccoloba ilheensis
 Coccoloba ilhensis

 Copaifera langsdorffii
 Copaifera langsdorfii

 Erechtites hieracifolia
 Erechtites hieraciifolia

 Galium hypocarpium
```

Gallium hypocarpium

Galium noxium
Gallium noxium

Herpetacanthus melancholicus
Herpetacanthus melancholius

Hibiscus furcelatus
Hibiscus furcellatus

Ichnanthus nemoralis
Ichnanthus nemorallis

Ilex theezans
Ilex thezans

Jatropha gossypifolia
Jatropha gossypiifolia

Ludwigia octovalis
Ludwigia octovalvis

Mandevilla illustris
Mandevilla ilustris

Ouratea castanaefolia
Ouratea castaneaefolia

Phaseolus pandurantus
Phaseolus panduratus

Pilocarpus spicatus
Pilocarpuss spicatus

Stenorrhynchos orchidoides
Stenorrhynchos orchioides

Stryphnodendron purpurem
Stryphnodendron purpureum

Turnera hermanniodes
Turnera hermannioides

Zornia flemingioides
Zornia flemmingioides

Total Names detected in all tests @ T: 0
Total Names detected in all tests @ W : 4
Centrolobium microchaeta
Centrolobium microchaete

Epidendrum cinnabarimum
Epidendrum cinnabarinum

Hybanthus ipecacuanha
Hybanthus ipepacuanha

Lagenocarpus guianenses
Lagenocarpus guianensis

```
Total Names pegos detected in all tests @ MU : 4
 Chamaecrista nictans
 Chamaecrista nictitans

 Coccoloba parimenensis
 Coccoloba parimensis

 Manilkara salzmanii
 Manilkara salzmanni

 Pithecellobium avaremotemo
 Pithecellobium avaremoto

Total Names detected in all tests @ FA : 1
 Aechmea blanchetiana
 Aechmea blanchetii


-------------------------------------------
To PMA:
-------------------------------------------
Total Names :156
Total Names to EX :8
Total Names to T :0
Total Names to W :4
Total Names to MU :1
Total Names to FA :143

Total Names to G2 :22
Total Names to G3 :15
Total Names to SNDX :22
Total Names to PHX :40
Total Names to SKL :10
Total Names to LV :129

Total Names to EX @ G2 :8
Total Names to EX @ G3 :8
Total Names to EX @ SNDX :8
Total Names to EX @ PHX :8
Total Names to EX @ SKL :8
Total Names to EX @ LV :8
Total Names to T @ G2 :0
Total Names to T @ G3 :0
Total Names to T @ SNDX :0
Total Names to T @ PHX :0
Total Names to T @ SKL :0
Total Names to T @ LV :0
Total Names to W @ G2 :4
Total Names to W @ G3 :3
Total Names to W @ SNDX :4
Total Names to W @ PHX :4
Total Names to W @ SKL :1
Total Names to W @ LV :4
Total Names to MU @ G2 :1
Total Names to MU @ G3 :0
Total Names to MU @ SNDX :1
Total Names to MU @ PHX :1
Total Names to MU @ SKL :0
Total Names to MU @ LV :1
Total Names to FA @ G2 :9
Total Names to FA @ G3 :4
```

```
Total Names to FA @ SNDX :9
Total Names to FA @ PHX :27
Total Names to FA @ SKL :1
Total Names to FA @ LV :116

Total Names to EX @ G :3
Total Names to EX @ E :5
Total Names to EX @ N :0
Total Names to T @ G :0
Total Names to T @ E :0
Total Names to T @ N :0
Total Names to W @ G :0
Total Names to W @ E :4
Total Names to W @ N :0
Total Names to MU @ G :0
Total Names to MU @ E :1
Total Names to MU @ N :0
Total Names to FA @ G :51
Total Names to FA @ E :92
Total Names to FA @ N :0

Total Names to EX @ G2 @ G :3
Total Names to EX @ G2 @ E :5
Total Names to EX @ G2 @ N :0
Total Names to EX @ G3 @ G :3
Total Names to EX @ G3 @ E :5
Total Names to EX @ G3 @ N :0
Total Names to EX @ SNDX @ G :3
Total Names to EX @ SNDX @ E :5
Total Names to EX @ SNDX @ N :0
Total Names to EX @ PHX @ G :3
Total Names to EX @ PHX @ E :5
Total Names to EX @ PHX @ N :0
Total Names to EX @ SKL @ G :3
Total Names to EX @ SKL @ E :5
Total Names to EX @ SKL @ N :0
Total Names to EX @ LV @ G :3
Total Names to EX @ LV @ E :5
Total Names to EX @ LV @ N :0
Total Names to T @ G2 @ G :0
Total Names to T @ G2 @ E :0
Total Names to T @ G2 @ N :0
Total Names to T @ G3 @ G :0
Total Names to T @ G3 @ E :0
Total Names to T @ G3 @ N :0
Total Names to T @ SNDX @ G :0
Total Names to T @ SNDX @ E :0
Total Names to T @ SNDX @ N :0
Total Names to T @ PHX @ G :0
Total Names to T @ PHX @ E :0
Total Names to T @ PHX @ N :0
Total Names to T @ SKL @ G :0
Total Names to T @ SKL @ E :0
Total Names to T @ SKL @ N :0
Total Names to T @ LV @ G :0
Total Names to T @ LV @ E :0
Total Names to T @ LV @ N :0
Total Names to W @ G2 @ G :0
Total Names to W @ G2 @ E :4
Total Names to W @ G2 @ N :0
Total Names to W @ G3 @ G :0
```

```
Total Names to W @ G3 @ E :3
Total Names to W @ G3 @ N :0
Total Names to W @ SNDX @ G :0
Total Names to W @ SNDX @ E :4
Total Names to W @ SNDX @ N :0
Total Names to W @ PHX @ G :0
Total Names to W @ PHX @ E :4
Total Names to W @ PHX @ N :0
Total Names to W @ SKL @ G :0
Total Names to W @ SKL @ E :1
Total Names to W @ SKL @ N :0
Total Names to W @ LV @ G :0
Total Names to W @ LV @ E :4
Total Names to W @ LV @ N :0
Total Names to MU @ G2 @ G :0
Total Names to MU @ G2 @ E :1
Total Names to MU @ G2 @ N :0
Total Names to MU @ G3 @ G :0
Total Names to MU @ G3 @ E :0
Total Names to MU @ G3 @ N :0
Total Names to MU @ SNDX @ G :0
Total Names to MU @ SNDX @ E :1
Total Names to MU @ SNDX @ N :0
Total Names to MU @ PHX @ G :0
Total Names to MU @ PHX @ E :1
Total Names to MU @ PHX @ N :0
Total Names to MU @ SKL @ G :0
Total Names to MU @ SKL @ E :0
Total Names to MU @ SKL @ N :0
Total Names to MU @ LV @ G :0
Total Names to MU @ LV @ E :1
Total Names to MU @ LV @ N :0
Total Names to FA @ G2 @ G :0
Total Names to FA @ G2 @ E :9
Total Names to FA @ G2 @ N :0
Total Names to FA @ G3 @ G :0
Total Names to FA @ G3 @ E :4
Total Names to FA @ G3 @ N :0
Total Names to FA @ SNDX @ G :4
Total Names to FA @ SNDX @ E :5
Total Names to FA @ SNDX @ N :0
Total Names to FA @ PHX @ G :11
Total Names to FA @ PHX @ E :16
Total Names to FA @ PHX @ N :0
Total Names to FA @ SKL @ G :0
Total Names to FA @ SKL @ E :1
Total Names to FA @ SKL @ N :0
Total Names to FA @ LV @ G :39
Total Names to FA @ LV @ E :77
Total Names to FA @ LV @ N :0

Total Names to G2 unicos : 1
Total Names to G3 unique : 0
Total Names to SNDX unique : 2
Total Names to PHX unique : 18
Total Names to SKL unique : 0
Total Names to LV unique : 106

Total Names to G2@ EX unique : 0
Total Names to G3@ EX unique : 0
Total Names to SNDX@ EX unique : 0
```

```
Total Names to PHX@ EX unique : 0
Total Names to SKL@ EX unique : 0
Total Names to LV@ EX unique : 0
Total Names to G2@ T unique : 0
Total Names to G3@ T unique : 0
Total Names to SNDX@ T unique : 0
Total Names to PHX@ T unique : 0
Total Names to SKL@ T unique : 0
Total Names to LV@ T unique : 0
Total Names to G2@ W unique : 0
Total Names to G3@ W unique : 0
Total Names to SNDX@ W unique : 0
Total Names to PHX@ W unique : 0
Total Names to SKL@ W unique : 0
Total Names to LV@ W unique : 0
Total Names to G2@ MU unique : 0
Total Names to G3@ MU unique : 0
Total Names to SNDX@ MU unique : 0
Total Names to PHX@ MU unique : 0
Total Names to SKL@ MU unique : 0
Total Names to LV@ MU unique : 0
Total Names to G2@ FA unique : 1
Total Names to G3@ FA unique : 0
Total Names to SNDX@ FA unique : 2
Total Names to PHX@ FA unique : 18
Total Names to SKL@ FA unique : 0
Total Names to LV@ FA unique : 106

Total Names detected in all tests : 9

Total Names detected in all tests @ EX : 8
 Beilschmedia emarginata
 Beilschmiedia emarginata

 Beilschmedia taubertiana
 Beilschmiedia taubertiana

 Brosimum glaziovi
 Brosimum glaziovii

 Cecropia glaziovi
 Cecropia glaziovii

 Chomelia estrelana
 Chomelia estrellana

 Ormosia fastgiata
 Ormosia fastigiata

 Thamniopisis incurva
 Thamniopsis incurva

 Thamniopsis langsdorffii
 Thamniopsis langsdorfii

Total Names detected in all tests @ T: 0
Total Names detected in all tests @ W : 1
 Maytenus brasilienses
 Maytenus brasiliensis

Total Names detected in all tests @ MU : 0
```

```
Total Names detected in all tests @ FA : 0

-------------------------------------------
To ILDIS:
-------------------------------------------
Total Names :33178
Total Names to EX :24
Total Names to T :0
Total Names to W :43
Total Names to MU :25
Total Names to FA :33086

Total Names to G2 :1848
Total Names to G3 :637
Total Names to SNDX :2618
Total Names to PHX :5779
Total Names to SKL :31
Total Names to LV :26631

Total Names to EX @ G2 :24
Total Names to EX @ G3 :19
Total Names to EX @ SNDX :10
Total Names to EX @ PHX :12
Total Names to EX @ SKL :8
Total Names to EX @ LV :24
Total Names to T @ G2 :0
Total Names to T @ G3 :0
Total Names to T @ SNDX :0
Total Names to T @ PHX :0
Total Names to T @ SKL :0
Total Names to T @ LV :0
Total Names to W @ G2 :41
Total Names to W @ G3 :32
Total Names to W @ SNDX :25
Total Names to W @ PHX :28
Total Names to W @ SKL :2
Total Names to W @ LV :43
Total Names to MU @ G2 :19
Total Names to MU @ G3 :10
Total Names to MU @ SNDX :12
Total Names to MU @ PHX :21
Total Names to MU @ SKL :3
Total Names to MU @ LV :25
Total Names to FA @ G2 :1764
Total Names to FA @ G3 :576
Total Names to FA @ SNDX :2571
Total Names to FA @ PHX :5718
Total Names to FA @ SKL :18
Total Names to FA @ LV :26539

Total Names to EX @ G :1
Total Names to EX @ E :23
Total Names to EX @ N :0
Total Names to T @ G :0
Total Names to T @ E :0
Total Names to T @ N :0
Total Names to W @ G :1
Total Names to W @ E :42
Total Names to W @ N :0
Total Names to MU @ G :0
Total Names to MU @ E :25
```

```
Total Names to MU @ N :0
Total Names to FA @ G :1636
Total Names to FA @ E :31450
Total Names to FA @ N :0

Total Names to EX @ G2 @ G :1
Total Names to EX @ G2 @ E :23
Total Names to EX @ G2 @ N :0
Total Names to EX @ G3 @ G :0
Total Names to EX @ G3 @ E :19
Total Names to EX @ G3 @ N :0
Total Names to EX @ SNDX @ G :0
Total Names to EX @ SNDX @ E :10
Total Names to EX @ SNDX @ N :0
Total Names to EX @ PHX @ G :0
Total Names to EX @ PHX @ E :12
Total Names to EX @ PHX @ N :0
Total Names to EX @ SKL @ G :0
Total Names to EX @ SKL @ E :8
Total Names to EX @ SKL @ N :0
Total Names to EX @ LV @ G :1
Total Names to EX @ LV @ E :23
Total Names to EX @ LV @ N :0
Total Names to T @ G2 @ G :0
Total Names to T @ G2 @ E :0
Total Names to T @ G2 @ N :0
Total Names to T @ G3 @ G :0
Total Names to T @ G3 @ E :0
Total Names to T @ G3 @ N :0
Total Names to T @ SNDX @ G :0
Total Names to T @ SNDX @ E :0
Total Names to T @ SNDX @ N :0
Total Names to T @ PHX @ G :0
Total Names to T @ PHX @ E :0
Total Names to T @ PHX @ N :0
Total Names to T @ SKL @ G :0
Total Names to T @ SKL @ E :0
Total Names to T @ SKL @ N :0
Total Names to T @ LV @ G :0
Total Names to T @ LV @ E :0
Total Names to T @ LV @ N :0
Total Names to W @ G2 @ G :1
Total Names to W @ G2 @ E :40
Total Names to W @ G2 @ N :0
Total Names to W @ G3 @ G :1
Total Names to W @ G3 @ E :31
Total Names to W @ G3 @ N :0
Total Names to W @ SNDX @ G :0
Total Names to W @ SNDX @ E :25
Total Names to W @ SNDX @ N :0
Total Names to W @ PHX @ G :0
Total Names to W @ PHX @ E :28
Total Names to W @ PHX @ N :0
Total Names to W @ SKL @ G :0
Total Names to W @ SKL @ E :2
Total Names to W @ SKL @ N :0
Total Names to W @ LV @ G :1
Total Names to W @ LV @ E :42
Total Names to W @ LV @ N :0
Total Names to MU @ G2 @ G :0
Total Names to MU @ G2 @ E :19
```

```
Total Names to MU @ G2 @ N :0
Total Names to MU @ G3 @ G :0
Total Names to MU @ G3 @ E :10
Total Names to MU @ G3 @ N :0
Total Names to MU @ SNDX @ G :0
Total Names to MU @ SNDX @ E :12
Total Names to MU @ SNDX @ N :0
Total Names to MU @ PHX @ G :0
Total Names to MU @ PHX @ E :21
Total Names to MU @ PHX @ N :0
Total Names to MU @ SKL @ G :0
Total Names to MU @ SKL @ E :3
Total Names to MU @ SKL @ N :0
Total Names to MU @ LV @ G :0
Total Names to MU @ LV @ E :25
Total Names to MU @ LV @ N :0
Total Names to FA @ G2 @ G :29
Total Names to FA @ G2 @ E :1735
Total Names to FA @ G2 @ N :0
Total Names to FA @ G3 @ G :11
Total Names to FA @ G3 @ E :565
Total Names to FA @ G3 @ N :0
Total Names to FA @ SNDX @ G :214
Total Names to FA @ SNDX @ E :2357
Total Names to FA @ SNDX @ N :0
Total Names to FA @ PHX @ G :715
Total Names to FA @ PHX @ E :5003
Total Names to FA @ PHX @ N :0
Total Names to FA @ SKL @ G :1
Total Names to FA @ SKL @ E :17
Total Names to FA @ SKL @ N :0
Total Names to FA @ LV @ G :807
Total Names to FA @ LV @ E :25732
Total Names to FA @ LV @ N :0

Total Names to G2 unique : 253
Total Names to G3 unique : 0
Total Names to SNDX unique : 1157
Total Names to PHX unique : 4107
Total Names to SKL unique : 0
Total Names to LV unique : 24440

Total Names to G2@ EX unique : 0
Total Names to G3@ EX unique : 0
Total Names to SNDX@ EX unique : 0
Total Names to PHX@ EX unique : 0
Total Names to SKL@ EX unique : 0
Total Names to LV@ EX unique : 0
Total Names to G2@ T unique : 0
Total Names to G3@ T unique : 0
Total Names to SNDX@ T unique : 0
Total Names to PHX@ T unique : 0
Total Names to SKL@ T unique : 0
Total Names to LV@ T unique : 0
Total Names to G2@ W unique : 0
Total Names to G3@ W unique : 0
Total Names to SNDX@ W unique : 0
Total Names to PHX@ W unique : 0
Total Names to SKL@ W unique : 0
Total Names to LV@ W unique : 1
Total Names to G2@ MU unique : 0
```

Total Names to G3@ MU unique : 0
Total Names to SNDX@ MU unique : 0
Total Names to PHX@ MU unique : 0
Total Names to SKL@ MU unique : 0
Total Names to LV@ MU unique : 1
Total Names to G2@ FA unique : 253
Total Names to G3@ FA unique : 0
Total Names to SNDX@ FA unique : 1157
Total Names to PHX@ FA unique : 4107
Total Names to SKL@ FA unique : 0
Total Names to LV@ FA unique : 24438

Total Names detected in all tests : 17

Total Names detected in all tests @ EX : 6
 Astragalus oxyrhynchus
 Astragalus oxyrrhynchus

 Caragana gerardiana
 Caragana gerrardiana

 Gastrolobium calcycinum
 Gastrolobium calycinum

 Lessertia acanthorachis
 Lessertia acanthorhachis

 Peltophorum dasyrhachis
 Peltophorum dasyrrhachis

 Vicia monardi
 Vicia monardii

Total Names detected in all tests @ T: 0
Total Names detected in all tests @ W : 2
 Bauhinia vespertilio
 Bauhinia vespertillo

 Inga brachystachya
 Inga brachystachys

Total Names detected in all tests @ MU : 1
 Jacksonia rhadinoclada
 Jacksonia rhadinoclona

Total Names detected in all tests @ FA : 8
 Adesmia trifoliata
 Adesmia trifoliolata

 Chamaecrista benthamiana
 Chamaecrista benthamii

 Copaifera jacquiniana
 Copaifera jacquinii

 Dalbergia foliolosa
 Dalbergia foliosa

 Dalbergia gossweileri
 Dalbergiella gossweileri

```
 Dalea foliolosa
 Dalea foliosa

 Indigofera evansiana
 Indigofera evansii

 Inga obtusa
 Inga obtusata


---------------------------------------------
To SP2K:
---------------------------------------------
Total Names :14062
Total Names to EX :14
Total Names to T :1
Total Names to W :23
Total Names to MU :10
Total Names to FA :14014

Total Names to G2 :1162
Total Names to G3 :647
Total Names to SNDX :1441
Total Names to PHX :4654
Total Names to SKL :11
Total Names to LV :8846

Total Names to EX @ G2 :13
Total Names to EX @ G3 :10
Total Names to EX @ SNDX :5
Total Names to EX @ PHX :5
Total Names to EX @ SKL :5
Total Names to EX @ LV :13
Total Names to T @ G2 :1
Total Names to T @ G3 :1
Total Names to T @ SNDX :0
Total Names to T @ PHX :0
Total Names to T @ SKL :0
Total Names to T @ LV :1
Total Names to W @ G2 :20
Total Names to W @ G3 :17
Total Names to W @ SNDX :12
Total Names to W @ PHX :11
Total Names to W @ SKL :0
Total Names to W @ LV :23
Total Names to MU @ G2 :7
Total Names to MU @ G3 :3
Total Names to MU @ SNDX :3
Total Names to MU @ PHX :7
Total Names to MU @ SKL :0
Total Names to MU @ LV :10
Total Names to FA @ G2 :1121
Total Names to FA @ G3 :616
Total Names to FA @ SNDX :1421
Total Names to FA @ PHX :4631
Total Names to FA @ SKL :6
Total Names to FA @ LV :8799

Total Names to EX @ G :1
Total Names to EX @ E :13
Total Names to EX @ N :0
```

```
Total Names to T @ G :0
Total Names to T @ E :1
Total Names to T @ N :0
Total Names to W @ G :0
Total Names to W @ E :23
Total Names to W @ N :0
Total Names to MU @ G :2
Total Names to MU @ E :8
Total Names to MU @ N :0
Total Names to FA @ G :3928
Total Names to FA @ E :10086
Total Names to FA @ N :0

Total Names to EX @ G2 @ G :0
Total Names to EX @ G2 @ E :13
Total Names to EX @ G2 @ N :0
Total Names to EX @ G3 @ G :0
Total Names to EX @ G3 @ E :10
Total Names to EX @ G3 @ N :0
Total Names to EX @ SNDX @ G :0
Total Names to EX @ SNDX @ E :5
Total Names to EX @ SNDX @ N :0
Total Names to EX @ PHX @ G :1
Total Names to EX @ PHX @ E :4
Total Names to EX @ PHX @ N :0
Total Names to EX @ SKL @ G :0
Total Names to EX @ SKL @ E :5
Total Names to EX @ SKL @ N :0
Total Names to EX @ LV @ G :0
Total Names to EX @ LV @ E :13
Total Names to EX @ LV @ N :0
Total Names to T @ G2 @ G :0
Total Names to T @ G2 @ E :1
Total Names to T @ G2 @ N :0
Total Names to T @ G3 @ G :0
Total Names to T @ G3 @ E :1
Total Names to T @ G3 @ N :0
Total Names to T @ SNDX @ G :0
Total Names to T @ SNDX @ E :0
Total Names to T @ SNDX @ N :0
Total Names to T @ PHX @ G :0
Total Names to T @ PHX @ E :0
Total Names to T @ PHX @ N :0
Total Names to T @ SKL @ G :0
Total Names to T @ SKL @ E :0
Total Names to T @ SKL @ N :0
Total Names to T @ LV @ G :0
Total Names to T @ LV @ E :1
Total Names to T @ LV @ N :0
Total Names to W @ G2 @ G :0
Total Names to W @ G2 @ E :20
Total Names to W @ G2 @ N :0
Total Names to W @ G3 @ G :0
Total Names to W @ G3 @ E :17
Total Names to W @ G3 @ N :0
Total Names to W @ SNDX @ G :0
Total Names to W @ SNDX @ E :12
Total Names to W @ SNDX @ N :0
Total Names to W @ PHX @ G :0
Total Names to W @ PHX @ E :11
Total Names to W @ PHX @ N :0
```

```
Total Names to W @ SKL @ G :0
Total Names to W @ SKL @ E :0
Total Names to W @ SKL @ N :0
Total Names to W @ LV @ G :0
Total Names to W @ LV @ E :23
Total Names to W @ LV @ N :0
Total Names to MU @ G2 @ G :1
Total Names to MU @ G2 @ E :6
Total Names to MU @ G2 @ N :0
Total Names to MU @ G3 @ G :0
Total Names to MU @ G3 @ E :3
Total Names to MU @ G3 @ N :0
Total Names to MU @ SNDX @ G :1
Total Names to MU @ SNDX @ E :2
Total Names to MU @ SNDX @ N :0
Total Names to MU @ PHX @ G :2
Total Names to MU @ PHX @ E :5
Total Names to MU @ PHX @ N :0
Total Names to MU @ SKL @ G :0
Total Names to MU @ SKL @ E :0
Total Names to MU @ SKL @ N :0
Total Names to MU @ LV @ G :2
Total Names to MU @ LV @ E :8
Total Names to MU @ LV @ N :0
Total Names to FA @ G2 @ G :167
Total Names to FA @ G2 @ E :954
Total Names to FA @ G2 @ N :0
Total Names to FA @ G3 @ G :94
Total Names to FA @ G3 @ E :522
Total Names to FA @ G3 @ N :0
Total Names to FA @ SNDX @ G :694
Total Names to FA @ SNDX @ E :727
Total Names to FA @ SNDX @ N :0
Total Names to FA @ PHX @ G :2975
Total Names to FA @ PHX @ E :1656
Total Names to FA @ PHX @ N :0
Total Names to FA @ SKL @ G :0
Total Names to FA @ SKL @ E :6
Total Names to FA @ SKL @ N :0
Total Names to FA @ LV @ G :517
Total Names to FA @ LV @ E :8282
Total Names to FA @ LV @ N :0

Total Names to G2 unique : 174
Total Names to G3 unique : 0
Total Names to SNDX unique : 546
Total Names to PHX unique : 3624
Total Names to SKL unique : 0
Total Names to LV unique : 7758

Total Names to G2@ EX unique : 0
Total Names to G3@ EX unique : 0
Total Names to SNDX@ EX unique : 0
Total Names to PHX@ EX unique : 1
Total Names to SKL@ EX unique : 0
Total Names to LV@ EX unique : 0
Total Names to G2@ T unique : 0
Total Names to G3@ T unique : 0
Total Names to SNDX@ T unique : 0
Total Names to PHX@ T unique : 0
Total Names to SKL@ T unique : 0
```

```
Total Names to LV@ T unique : 0
Total Names to G2@ W unique : 0
Total Names to G3@ W unique : 0
Total Names to SNDX@ W unique : 0
Total Names to PHX@ W unique : 0
Total Names to SKL@ W unique : 0
Total Names to LV@ W unique : 1
Total Names to G2@ MU unique : 0
Total Names to G3@ MU unique : 0
Total Names to SNDX@ MU unique : 0
Total Names to PHX@ MU unique : 0
Total Names to SKL@ MU unique : 0
Total Names to LV@ MU unique : 1
Total Names to G2@ FA unique : 174
Total Names to G3@ FA unique : 0
Total Names to SNDX@ FA unique : 546
Total Names to PHX@ FA unique : 3623
Total Names to SKL@ FA unique : 0
Total Names to LV@ FA unique : 7756

Total Names detected in all tests : 4

Total Names detected in all tests @ EX : 4
 Cottus kessleri
 Cottus kesslerii

 Nemacheilus strauchi
 Nemacheilus strauchii

 Paracottus kneri
 Paracottus knerii

 Percarina demidoffi
 Percarina demidoffii

Total Names detected in all tests @ T: 0
Total Names detected in all tests @ W : 0
Total Names detected in all tests @ MU : 0
Total Names detected in all tests @ FA : 0

-------------------------------------------
To MARINE:
-------------------------------------------
Total Names :1772
Total Names to EX :275
Total Names to T :52
Total Names to W :151
Total Names to MU :211
Total Names to FA :1083

Total Names to G2 :763
Total Names to G3 :525
Total Names to SNDX :669
Total Names to PHX :908
Total Names to SKL :325
Total Names to LV :1459

Total Names to EX @ G2 :263
Total Names to EX @ G3 :222
Total Names to EX @ SNDX :224
Total Names to EX @ PHX :238
```

```
Total Names to EX @ SKL :176
Total Names to EX @ LV :275
Total Names to T @ G2 :25
Total Names to T @ G3 :12
Total Names to T @ SNDX :44
Total Names to T @ PHX :41
Total Names to T @ SKL :39
Total Names to T @ LV :52
Total Names to W @ G2 :142
Total Names to W @ G3 :78
Total Names to W @ SNDX :100
Total Names to W @ PHX :103
Total Names to W @ SKL :36
Total Names to W @ LV :151
Total Names to MU @ G2 :127
Total Names to MU @ G3 :77
Total Names to MU @ SNDX :163
Total Names to MU @ PHX :179
Total Names to MU @ SKL :60
Total Names to MU @ LV :209
Total Names to FA @ G2 :206
Total Names to FA @ G3 :136
Total Names to FA @ SNDX :138
Total Names to FA @ PHX :347
Total Names to FA @ SKL :14
Total Names to FA @ LV :772

Total Names to EX @ G :70
Total Names to EX @ E :205
Total Names to EX @ N :0
Total Names to T @ G :10
Total Names to T @ E :42
Total Names to T @ N :0
Total Names to W @ G :47
Total Names to W @ E :104
Total Names to W @ N :0
Total Names to MU @ G :21
Total Names to MU @ E :162
Total Names to MU @ N :28
Total Names to FA @ G :538
Total Names to FA @ E :545
Total Names to FA @ N :0

Total Names to EX @ G2 @ G :60
Total Names to EX @ G2 @ E :203
Total Names to EX @ G2 @ N :0
Total Names to EX @ G3 @ G :54
Total Names to EX @ G3 @ E :168
Total Names to EX @ G3 @ N :0
Total Names to EX @ SNDX @ G :52
Total Names to EX @ SNDX @ E :172
Total Names to EX @ SNDX @ N :0
Total Names to EX @ PHX @ G :55
Total Names to EX @ PHX @ E :183
Total Names to EX @ PHX @ N :0
Total Names to EX @ SKL @ G :40
Total Names to EX @ SKL @ E :136
Total Names to EX @ SKL @ N :0
Total Names to EX @ LV @ G :70
Total Names to EX @ LV @ E :205
Total Names to EX @ LV @ N :0
```

```
Total Names to T @ G2 @ G :2
Total Names to T @ G2 @ E :23
Total Names to T @ G2 @ N :0
Total Names to T @ G3 @ G :0
Total Names to T @ G3 @ E :12
Total Names to T @ G3 @ N :0
Total Names to T @ SNDX @ G :6
Total Names to T @ SNDX @ E :38
Total Names to T @ SNDX @ N :0
Total Names to T @ PHX @ G :3
Total Names to T @ PHX @ E :38
Total Names to T @ PHX @ N :0
Total Names to T @ SKL @ G :6
Total Names to T @ SKL @ E :33
Total Names to T @ SKL @ N :0
Total Names to T @ LV @ G :10
Total Names to T @ LV @ E :42
Total Names to T @ LV @ N :0
Total Names to W @ G2 @ G :46
Total Names to W @ G2 @ E :96
Total Names to W @ G2 @ N :0
Total Names to W @ G3 @ G :30
Total Names to W @ G3 @ E :48
Total Names to W @ G3 @ N :0
Total Names to W @ SNDX @ G :33
Total Names to W @ SNDX @ E :67
Total Names to W @ SNDX @ N :0
Total Names to W @ PHX @ G :34
Total Names to W @ PHX @ E :69
Total Names to W @ PHX @ N :0
Total Names to W @ SKL @ G :17
Total Names to W @ SKL @ E :19
Total Names to W @ SKL @ N :0
Total Names to W @ LV @ G :47
Total Names to W @ LV @ E :104
Total Names to W @ LV @ N :0
Total Names to MU @ G2 @ G :15
Total Names to MU @ G2 @ E :109
Total Names to MU @ G2 @ N :3
Total Names to MU @ G3 @ G :5
Total Names to MU @ G3 @ E :72
Total Names to MU @ G3 @ N :0
Total Names to MU @ SNDX @ G :16
Total Names to MU @ SNDX @ E :132
Total Names to MU @ SNDX @ N :15
Total Names to MU @ PHX @ G :18
Total Names to MU @ PHX @ E :144
Total Names to MU @ PHX @ N :17
Total Names to MU @ SKL @ G :8
Total Names to MU @ SKL @ E :48
Total Names to MU @ SKL @ N :4
Total Names to MU @ LV @ G :19
Total Names to MU @ LV @ E :162
Total Names to MU @ LV @ N :28
Total Names to FA @ G2 @ G :107
Total Names to FA @ G2 @ E :99
Total Names to FA @ G2 @ N :0
Total Names to FA @ G3 @ G :73
Total Names to FA @ G3 @ E :63
Total Names to FA @ G3 @ N :0
Total Names to FA @ SNDX @ G :70
```

```
Total Names to FA @ SNDX @ E :68
Total Names to FA @ SNDX @ N :0
Total Names to FA @ PHX @ G :246
Total Names to FA @ PHX @ E :101
Total Names to FA @ PHX @ N :0
Total Names to FA @ SKL @ G :10
Total Names to FA @ SKL @ E :4
Total Names to FA @ SKL @ N :0
Total Names to FA @ LV @ G :283
Total Names to FA @ LV @ E :489
Total Names to FA @ LV @ N :0

Total Names to G2 unique : 10
Total Names to G3 unique : 0
Total Names to SNDX unique : 37
Total Names to PHX unique : 195
Total Names to SKL unique : 0
Total Names to LV unique : 591

Total Names to G2@ EX unique : 0
Total Names to G3@ EX unique : 0
Total Names to SNDX@ EX unique : 0
Total Names to PHX@ EX unique : 0
Total Names to SKL@ EX unique : 0
Total Names to LV@ EX unique : 5
Total Names to G2@ T unique : 0
Total Names to G3@ T unique : 0
Total Names to SNDX@ T unique : 0
Total Names to PHX@ T unique : 0
Total Names to SKL@ T unique : 0
Total Names to LV@ T unique : 5
Total Names to G2@ W unique : 0
Total Names to G3@ W unique : 0
Total Names to SNDX@ W unique : 0
Total Names to PHX@ W unique : 0
Total Names to SKL@ W unique : 0
Total Names to LV@ W unique : 6
Total Names to G2@ MU unique : 0
Total Names to G3@ MU unique : 0
Total Names to SNDX@ MU unique : 1
Total Names to PHX@ MU unique : 1
Total Names to SKL@ MU unique : 0
Total Names to LV@ MU unique : 17
Total Names to G2@ FA unique : 10
Total Names to G3@ FA unique : 0
Total Names to SNDX@ FA unique : 36
Total Names to PHX@ FA unique : 194
Total Names to SKL@ FA unique : 0
Total Names to LV@ FA unique : 558

Total Names detected in all tests : 206

Total Names detected in all tests @ EX : 153
 Aglaophamus verrilli
 Aglaophamus verrillii

 Amalda novaezealandiae
 Amalda novaezelandiae

 Ampelisca bouvieri
 Ampelisca bouvierii
```

Amphidesma subtriangulatum
Amphidesma subtriangulaum

Amphidesma subtriangulatum
Amphidesma subtringulatum

Amphilochus filidactylus
Amphilocus filidactylus

Antisolarium egenum
Antisolarium eugenum

Apseudes novaezealandiae
Apseudes novaezelandiae

Arachnopusia latiavicularis
Arachnopusia lativicularis

Argobuccinum tumidum
Argoobuccinum tumidum

Astraea heliotrophium
Astraea heliotrophum

Astraea heliotrophium
Astraea heliotropium

Astraea heliotrophum
Astraea heliotropum

Astraea heliotropium
Astraea heliotropum

Balanus vestitus
Blanus vestitus

Barbatia novaezealandiae
Barbatia novaezealndiae

Barbatia novaezealandiae
Barbatia novaezelandiae

Barbatia novaezelandiae
Barbatia novazelandiae

Baryspira novaezealandiae
Baryspira novaezelandiae

Baryspira novaezealandiae
Baryspira novazealandiae

Basina yatei
Bassina yatei

Bassina yatei
Bassinia yatei

Biddulpha chinensis
Biddulphia chinensis

Caberea darwini
Caberea darwinii

Canitharidus capilaceus
Cantharidus capilaceus

Cantharidus capilaceus
Cantharidus capillaceus

Cantharidus capilaceus
Cantharidus cappilaceus

Cerataulina pelagica
Ceratulina pelagica

Chirundina streetsi
Chirundina streetsii

Chlamys kiwaensis
Chlamys kiwiaensis

Cibicides wuelleerstorfi
Cibicides wuellerstorfi

Cirolana quadripustlata
Cirolana quadripustulata

Cnemidocarpa bicornuata
Cnemidocarpa bicornuta

Coelorhynchus oliveranus
Coelorhynchus oliverianus

Comanthus wahlbergi
Comanthus wahlbergii

Cominella nassoides
Cominella nassoiodes

Corbula zealandica
Corbula zelandica

Corethron criophilium
Corethron criophilum

Coscinodiscus grani
Coscinodiscus granii

Crassimarginatella inconstanta
Crassimarginatella inconstantia

Ctenocalanus vanus
Ctenoclanus vanus

Cuspidaria trailei
Cuspidaria traillei

Cycloberris zealandica
Cycloberris zelandica

Cycloleberis zealandica

Cyclolebris zealandica

Dentalium zealandicum
Dentalium zelandicum

Dinophysis forti
Dinophysis fortii

Ditylum brightwelli
Ditylum brightwellii

Dosina zelandica
Dosinia zelandica

Dosinia zealandica
Dosinia zelandica

Dosinula zealandica
Dosinula zelandica

Duplicara tristis
Duplicaria tristis

Dynamenella cordiforaminalis
Dynamenella cordiforminalis

Ennucula strangei
Ennucula strangeri

Epitonium philippinarium
Epitonium philippinarum

Escharina waiparaensis
Escharina waiparensis

Eucomina nassoides
Eucominia nassoides

Exosphaeroma chilensis
Exosphaeroma chiliensis

Flaccisaagitta hexaptera
Flaccisagitta hexaptera

Fragilaria oceanica
Fragillaria oceanica

Galeodea trigancea
Galeodea triganceae

Gari strangei
Gari strangeri

Glycera tesselata
Glycera tessellata

Golfingia canatabriensis
Golfingia cantabriensis

Gondogeneia danai
Gonodogeneia danai

Gonimyrtea concinna
Goniomyrtea concinna

Halicarcinus cooki
Halicarcinus cookii

Harpecia spinosissima
Harpecia spinossissima

Harpecia spinossisima
Harpecia spinossissima

Hoplostethus gilchristi
Hoplosthethus gilchristi

Hoplostethus intermedius
Hoplosthethus intermedius

Hyperammina novaezealandiae
Hyperammina novaezelandiae

Hyperia sibaginis
Hyperia sibaginsis

Kolostoneura novaezealandiae
Kolostoneura novaezelandiae

Lagena meridionalis
Lagena meridonalis

Ligia novaezealandiae
Ligia novaezelandiae

Lima zealandica
Lima zelandica

Liothyrella neozealanica
Liothyrella neozelanica

Lithosoma novaezealandiae
Lithosoma novaezelandiae

Lucinoma galathea
Lucinoma galatheae

Lumbrineris tetraua
Lumbrineris tetraura

Macomona lilian
Macomona liliana

Macrochiridotea uncinata
Macrochiridothea uncinata

Maera masteri
Maera mastersi

Metridia gerlachei
Metridia gerlachi

Microporella ciliata
Microporella cilliata

Modiolus aereolatus
Modiolus aerolatus

Modiolus aereolatus
Modiolus areolatus

Modiolus areolataus
Modiolus areolatus

Myadora antipodium
Myadora antipodum

Myadora novaezealandiae
Myadora novaezelandiae

Nemocardium pulchellum
Nemocardium pulchelum

Nitzschia closterium
Nitzschiza closterium

Nitzschia seriata
Nitzschiza seriata

Notocallista mulitistriata
Notocallista multistriata

Notocorbula zealandica
Notocorbula zealnandica

Notocorbula zealandica
Notocorbula zelandica

Notocorbula zelandica
Notocorbula zelanndica

Nucula nitida
Nucula nitidia

Nucula strangei
Nucula strangeri

Oncaea mediteranea
Oncaea mediterranea

Onychoteuthis banksi
Onychoteuthis banksii

Oolina heteromorpha
Oolina heterormorpha

Ophionotus victoriae
Ophiontus victoriae

Oplophorus novaezeelandiae
Oplophorus novaezelandiae

Orthoporida compacta

Orthoporidra compacta

Orthoporida stenorhyncha
Orthoporidra stenorhyncha

Orthoscuticella fissurata
Orthoscuticella fisurata

Osthimosia milleporides
Osthimosia milleporoides

Ostrea sinnuata
Ostrea sinuata

Pagurus novaezealandia
Pagurus novaezelandia

Pagurus novaezelandia
Pagurus novaezelandiae

Panthalis novaezealandiae
Panthalis novaezelandiae

Paphirus largillerti
Paphirus largillierti

Paranarthrura similis
Paranarthrura simils

Pecten novaezealandiae
Pecten novaezealndiae

Pecten novaezealandiae
Pecten novaezelandiae

Pecten novaezealandiae
Pecten novazealandiae

Penion cuvierana
Penion cuvieriana

Pervicacia tristis
Pervicacia tristris

Petrolisthes lamarcki
Petrolisthes lamarckii

Phenatoma zealandica
Phenatoma zelandica

Phenatoma zelandica
Phenatoma zelanidica

Phylo novazealandiae
Phylo novazelandiae

Pleuromeris zealandica
Pleuromeris zelandica

Pogonophryne scotti
Pogonphryne scotti

Poiriera zelandica
Poirieria zelandica

Poirieria zelandica
Poririeria zelandica

Potamopyrgus antipodium
Potamopyrgus antipodum

Proharpina antipoda
Proharpinia antipoda

Proharpina hurleyi
Proharpinia hurleyi

Pterocirrus magalhaensis
Pterocirrus magalhensis

Pyramimonas grossi
Pyramimonas grossii

Pyura bouvetensis
Pyura bouvettensis

Quinqueloculina colleenae
Quinqueloculina collenae

Racovitzia harrisoni
Racovitzia harrissoni

Rhysoplax canaliculata
Rhyssoplax canaliculata

Ruvettus prometheus
Ruvettus promethus

Scrupocellaria ornithorhynchus
Scrupocellaria ornithorhyncus

Similium acutum
Smilium acutum

Smilium acutum
Smillium acutum

Smittina akaroaensis
Smittina akaroensis

Splendrilla aoteana
Splendrillia aoteana

Syringammina fragilissima
Syringammina fragillissima

Tellina hutoni
Tellina huttoni

Terebellides stroemi
Terebellides stroemii

Textularia tennuissima
Textularia tenuissima

Thalassiothrix nitzschioides
Thalassiothrix nitzschoides

Trypanosyllis taeniaeformis
Trypanosyllis taeniaformis

Venerupis largillerti
Venerupis largillierti

Vibilia stebbingi
Vibilia stebbingii

Waitangi brevirostis
Waitangi brevirostris

Zethalia zealandica
Zethalia zelandica

Zygophlax sibogae
Zygophylax sibogae

Total Names detected in all tests @ T: 12
Barbatia novaezealndiae
Barbatia novaezelandiae

Cantharidus capillaceus
Cantharidus cappilaceus

Harpecia spinosissima
Harpecia spinossisima

Lucicutia falvicornis
Lucicutia flavicornis

Lumbrineris sphaerocehpala
Lumbrineris sphaerocephala

Osthimosia notialis
Osthimosia notialsi

Otionella australis
Otionella australsi

Pecten novaezealndiae
Pecten novaezelandiae

Pholadidea acherontae
Pholadidea acherontea

Pulleniatina obliquiloculata
Pulleniatina obliquliоculata

Sigapatella novaezealndiae
Sigapatella novaezelandiae

Trachyleberis lytteltonensis
Trachyleberis lyttletonensis

Total Names detected in all tests @ W : 21
Astraea heliotrophum
Astraea heliotropium

Astrononian novozealandicum
Astrononion novozealandicum

Cylichna thetides
Cylichna thetidis

Divaricella huttoniaia
Divaricella huttoniana

Escharoides praestita
Escharoides praestite

Eunice tridentata
Eunice tridentate

Globigerinella aequilateralis
Globigerinella aequilaterelis

Hemiaulis hauckii
Hemiaulus hauckii

Hemiaulis sinensis
Hemiaulus sinensis

Heteromolpadia marenzelleri
Heteromolpadia marenzelliri

Heteronandoa carinata
Heteronardoa carinata

Kolestoneura novaezelandiae
Kolostoneura novaezelandiae

Macrochiridothea uncinata
Macrochirodothea uncinata

Melarhapha cincta
Melarhaphe cincta

Melarhapha oliveri
Melarhaphe oliveri

Notopandalus magnoculus
Notopandulus magnoculus

Parawaldeckia parata
Paraweldeckia parata

Parawaldeckia thomsoni
Paraweldeckia thomsoni

Planispirinoides bucculenta
Planispironoides bucculenta

Prionocrangon curvicaulis
Prionocrangon curvicaulus

Torridiharpinia hurleyi
Torridoharpinia hurleyi

Total Names detected in all tests @ MU : 20
Acitellina urinatoni
Acitellina urinatoria

Astraea heliotrophium
Astraea heliotropum

Cadulus delicatulus
Cadulus delicatus

Crassimarginatella valdemunita
Crassimarginatella valdemunitella

Divaricella huttonia
Divaricella huttoniaia

Divaricella huttonia
Divaricella huttoniana

Eucampia balaustima
Eucampia balaustium

Haplophragmoides canariense
Haplophragmoides canariensis

Haploscoloplos kerguelenensis
Haploscoloplos kerguelensis

Marikellia rotunda
Marikellia rotundata

Molgula mortenseni
Molgula mortensi

Pagurus novaezealandia
Pagurus novaezelandiae

Paphies subtriangula
Paphies subtriangulata

Penion adusta
Penion adustus

Protothaca crassicosta
Protothaca crassicostata

Reophax sabulosa
Reophax sabulosus

Solemya parkinsoni
Solemya parkinsoniana

Sternaspis scuta
Sternaspis scutata

Symplectoscyphus subarticulata
Symplectoscyphus subarticulatus

Uberella barriensis
Uberella barrierensis

Total Names detected in all tests @ FA : 0