DATABASE MODEL FOR TAXONOMIC AND OBSERVATION DATA

Jan Lindström

Oracle Corp./ Innobase Oy
Aleksanterinkatu 17
FIN-00101 HELSINKI
FINLAND
jan.lindstrom@oracle.com

ABSTRACT
This article describes a comprehensive information model for the recording of taxonomic and observation data from literature, field collecting and other sources. The model is based on an approach using hierarchical decomposition of data areas into atomic data elements and abstraction into an entity relationship model. It encompasses taxa of all ranks, nothotaxa and hybrid formulae, unnamed taxa, cultivars, full synonymy, misapplied names, basionyms, nomenclatural data, and different taxonomic concepts as well as alternative taxonomies to any extent desired. It can help designers of biological information systems to avoid the widely made error of over-simplification of taxonomic data and the resulting loss in data accuracy and quality.

KEY WORDS
Biological databases, database modelling, taxonomy.

1. Introduction

Databases used in fields as diverse as archaeology, biodiversity and environmental impact studies, gene sequencing, collection management, and pharmacognosy, to name but a few, make use of scientific species names [4]. These names are thought to stand for taxonomic groups (taxa). Biological information is, in a general way, linked to taxa which in turn are designated by name. However, the notion of the essence of a taxon varies greatly, depending context. The taxonomist correctly understands a taxon as a hypothesis, as a set of biological objects within a classification unit supposedly linked by phylogenetic descent, or as a set of criteria applying to such objects [2]. Phylogenetic systematic is the way that biologists reconstruct the pattern of events that have led to the distribution and diversity of life.

In contrast, the user of taxonomic information tends to understand and use a taxon as if it represented a foregone conclusion. This is a misunderstanding because a scientific hypothesis must be testable and, depending on the result of testing, is bound to change. Specimens are the basic operational taxonomic units. Consequently, 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 [3, 5].

Due to the inherent limitations of nomenclature a name may correctly designate several perhaps equally well-founded concepts of a taxon [2] (see Figure 1). For the purpose of information handling, a way has to be found to differentiate between different taxa bearing the same name in an information system [4], this can be achieved by introducing a data element or data area which mirrors alternative taxonomies, and allows for the inclusion of all information-bearing individual taxonomic concepts, including misnomers.

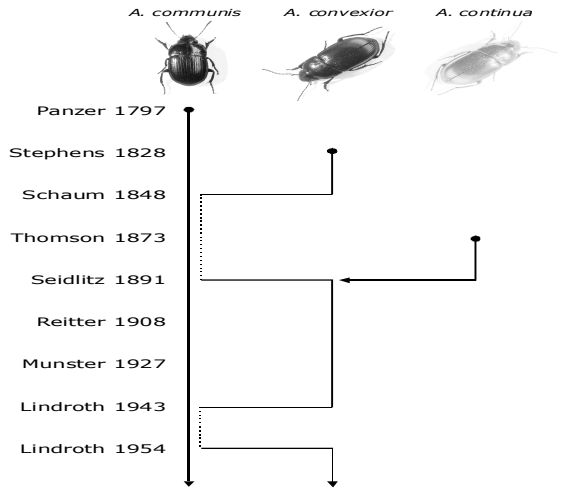

Figure 1: Evolution of one taxon.

Biodiversity has been defined as: The variability among living organisms from all sources including, among others, terrestrial, marine and other aquatic ecosystems, and the ecological complexes of which they are a part; this includes diversity among genes, individuals, and

populations within and between species, and of ecosystems.

In ecosystem level we have habitats, biotypes and associations of organisms. Organism level contains species and infraspecific taxa (a taxonomic group within a species, such as a subspecies.). Finally, molecular level consist genes and metabolic products. This distinction somewhat hinders information access because it tends to obscure the common elements important for information access, which are based on classification systems covering all levels of biodiversity information [6]. Geographic information and the scientific names for groups of organisms (taxa) are key access points as well as the most important information output provided by biodiversity information systems. Designations for taxa form the principal index for large parts of the existing scientific information about life on earth (past and present).

Of course there are other major biological fields which use ecosystems, taxa and/or molecules as their objects of research, analysing their composition, structure, processes and organisation. However, all scientific fields develop classification systems and terminologies to express their results, so the principal problems encountered in the indexing of biodiversity information are by no means restricted to that subject. The interdisciplinary approach of this paper is based on the realization that these problems exist in information retrieval for any field to which human descriptive and research activities are directed [1].

The name of a taxon describes a class of items that are not exactly alike. Classes are formed to subdivide the almost infinite amount of variation found in the objects surrounding us. A taxon is a class of organisms, defined by a selection of criteria which are (at least in their combination) unique to the class. Taxa are named in accordance with the international codes of nomenclature.

As soon as a class of organisms has been formed, the international codes of nomenclature provide clear rules on how to name the taxon. Separate codes exist for Animals [10], Plants [9], Bacteria [12], and Viruses [8]. The names formed in accordance with the codes are not descriptive, which has proven to be a practical requirement of the system.

The rest of the paper is organised as follows. In Section 2 taxonomic part of the database is presented. Section 3 presents the observation part of the database. Section 4 presents the user extendable parts of the database. Finally Section 5 concludes this paper.

2. Taxonomic database

A comprehensive information model for the recording of taxonomic data from literature and other sources is presented (see Figure 2 and [15]). The model is based on an approach using hierarchical decomposition of data areas into atomic data elements and abstraction into an entity relationship model. It encompasses taxa of all ranks, nothotaxa (named hybrids) and hybrid formulae, unnamed taxa, cultivars, full synonymy, misapplied names, basionyms, nomenclatural data, and differing taxonomic concepts as well as alternative taxonomies to any extend desired.

To preserve the history of the name editing process, names are stored in two tables *Taxon* (see Table 1) and *TaxonHist*. The *Taxon* table contains the set of names currently in use by a system. The *TaxonHist* table represents the editing history of names. The *TaxonHist* table is structured very similarly with a few exceptions: a self-referential pointer (*SuccTaxId*) links a name to its successor within the editing history. An additional pointer (*CurrTaxId*) gives for every name the name currently in use.

Rank relation is container for a taxonomic rank or its abbreviation e.g. division, phylum, subphylum, super class, class, subclass, order, family, genus, species, subspecies, variety, form. *HigherRank* attribute is link to higher taxonomic rank.

Relations *Dataset*, *Contact* and *Rights* are container elements for one to many unit data records and their shared metadata elements. The metadata relates to the source, provision, the ownership and rights etc. pertaining to the entire dataset. *Rights* relation contains elements expressing legal rights. *Dataset* relation holds properties of the original database or other data sources from which the data is derived. *Contact* relation identifies an organization and/or a person as contact and provides contact information of the data source.

Table 1. Fields of the Taxon relation.

Column	Data type	Description
TaxonId (PK, I1)	int	Primary key
Name (I2)	varchar(255)	Full name of the taxon
Description	text	Description of the taxon
InfSpepi	text	Infraspecific epithet
Year	int	Year
Status	varchar(40)	Status of this taxon
Notes	text	Comments
Updated	datetime	Date of the last update
Created	datetime	Date created
DataSetId (FK,I3)	int	Link to dataset
RankId (FK,I4)	int	Link to rank of this taxon
Editor (FK, I5)	int	Link to author of this record

The *RelTaxon* table is used to express directed binary relations between any two entries of the *Taxon* table (See Table 2). This includes taxonomic inclusions (i.e. the classification system) and any kind of synonym and concept relations. The type of the relation is specified in the attribute *Qualifier*. The source of the relationship is indicated by the attribute *Relation*. Holding all kinds of binary relations in a single table provides a convenient and transparent way to query the links between taxa, and it greatly facilitates the implementation of client-sided navigation functions.

Table 2: Fields of the RelTaxon relation.

Column	Data type	Description
TaxonId (PK, I1)	int	Fist taxon id
RelTaxonId (PK,I2)	int	Second taxon id
Qualifier	varchar(80)	Type of relation
Relation	varchar(80)	Source of relationship
Notes	text	Notes
Created	datetime	Date when created
Update	datetime	Date of the last update

As an example consider query where user wants to select all synonyms for *Larus Argentatus.*

```
select Name from Taxon T, RelTaxon R
where T.Name = 'Larus Argentatus' and
T.TaxonId = R.TaxonId and
R.Qualifier = 'Synonym';
```

References of any kind (e.g. nomenclatural references, references for taxonomic opinions, factual data references) are accommodated in a single recurrent structure *Publications*. *Author* relation is a container for e.g. taxon authors, publication authors. *Users* relation is a container for e.g. record editors, database users, and project members. *Users* relation is very important because different authors, database users, record editors and project members have different needs and privileges to use and edit the data stored to the database.

3. Observation part of the database

Natural history collections and observation data sets represent sets of observations (see Figure 2), with each record detailing the observation of an organism, ideally at a specific geo-temporal location. In the case of collections, the observation is permanent in that the organism was collected from the field and preserved in a coated collection intended to last indefinitely.

Collected specimens can be prepared in various ways, and several preparations from a single organism are not unusual (skin, skeleton, and perhaps microscope slides), thus there may be several records for a single organism, each representing the organism prepared using different techniques, but all records referring to a single observation event.

Conversely, some collection records may represent a collection object that contains many organisms. For example in ichthyology the contents of a trawl may be sorted by taxon and lumped into a single collection container. Observation data sets catalogue the observation of an organism, also at a specific geo-temporal location, but in this case the organism observed is not collected, and hence the observation record is the only information recorded about the organism.

In both cases a taxonomic identification of the organism is attempted, with obvious consequences for accuracy of identification (a specimen available for identification to several experts compared with a potentially fleeting glimpse of an organism in the field).

In field collections and observations we distinguish four main data areas:

- The gathering site provides all geographical and ecological data deemed necessary to describe the locality where the organisms were collected or observed.
- The gathering event refers to the act of collecting and/or observing organisms at a given site, i.e. time, person, and project data.
- The gathering or field unit represents the non-descriptive field information specific to each item which has been distinguished by the collector or observer during the gathering event (e.g. items distinguished by means of different collection numbers).
- Field descriptors are the results of observations made in the field on the item specified in the entity *Gathering*.

Specimens are collected during periodically scheduled visits to the field. These visits are called gatherings and information obtained from them is stored to *Gathering* relation. Each gathering has several items of associated data, among these are: the site, the period of time, the name of the collector, the method of collection, etc. *BiotypeData* relation stores classification what kind of biotype was found from gathering site and its name and description.

Geography relation stores geographic information of the gathering site i.e. the location of the gathering site. *Iscurrent* attribute tells if this row in the relation is the valid at the moment. This is needed because geography changes at time. New countries form, their names can change and other natural changes can happen. *CurrentId* links the row to the current geographical information.

A detailed and complete coverage of all the data items which may be incorporated into the geographical and ecological site description would clearly exceed the scope of this model. The US Federal Geographic Data Committee [8] lists more than 300 individual data elements and compound elements for geospatial data alone. We came to the conclusion that for extensive coverage of geographical and other data related to field

collections a database system should rely on one of the available commercial GIS (geographic information system) programs. However, site data cannot be excluded from this model, because they are too important or even form essential part of the objective of the task. The model as presented here is intended to aid in the definition of requirements for a linked GIS or for the basic structures to be implemented in a proprietary database design. Roughly, four data areas may be distinguished:

- Geospatial co-ordinate data, either in the form of point locations (geographical co-ordinates and altitude), or in the form of grid locations. The former are represented by a "flat" data structure, while grid data may include a hierarchical element. Co-ordinate data are unequivocal as long as the base system and/or the method of measurement is cited.
- Gazetteer data include a diverse throng of areas, i.e. bounded, continuous pieces of the earth's surface which are designated with a name and which are delimited by political, administrative, traditional, geomorphologic and/or ecological criteria. Named areas are often part of a more or less well-defined hierarchy, they may change over time, and many synonyms as well as homonyms may exist.
- Geo-ecological classification units are named classes of areas distinguished by some more or less well-defined climatic, edaphically, geomorphologic, or synecological characteristics. Outside of published systems, little standardization exists and consequently, many common terms may be equivocal.
- Ecological site descriptors are individual measurements or observations of ecological parameters at the collection site itself.

GatheringSite stores locality where collecting event took place containing detailed description of the gathering site. Unit stores elements for all unit-level data. The concept of a Unit as a physical object in the field or in a collection is central to this model. It includes organisms observed in the field or soil samples taken, herbarium specimens, microbial strains, or even pure substances in a natural products collection. Field data, taxonomic identifications, curatorial activities, collection management data as well as all kinds of descriptive data, are linked to units.

Identifications is complex type that supports the application of a name (or concept) to a *Unit* (specimen, observation, etc.), with appropriate metadata about the Identification, also known as *Determination*. In biological terminology, classification is the process of defining and naming classes of organisms. These classes are called taxa. Identification is the process of assigning a specimen to a (pre-existing) taxon. The name of the taxon can then be used as an index to find known information about the taxon, and therefore about the specimen itself (e.g. whether it is a pest, and, if so, how it can be controlled).

Table 3. Fields of Identifications relation

Column	Data type	Description
IdentificationId (PK, I1)	int	Primary key
Name (I2)	varchar(255)	Name
Description	text	Description
Type	varchar(255)	Type
CollectionId (I3)	varchar(255)	Collection identifier
Notes	text	Notes
Created	datetime	Date created
Updated	datetime	Date of the last update

Identifications are connected to Taxon using a *TaxIdRef* relation which contains primary keys from both table i.e. *TaxonId and IdentificationId.* As an example consider Query where user request a number of observations of 'Larus Argentatus'.

```
select count(1)
from Taxon T, TaxIdRef R, Identifications I,
where T.Name = 'Larus Argentatus' and
T.TaxonId = R.TaxonId and
R.IdentificationId = I.IdentificationId;
```

Unit stores which dataset this identification belongs, who is the editor for this unit, dataset where this identification was provided, field number and other metadata. *Unit* is connected to *DataSet* relation which contain information where and when this dataset was obtained.

Stratigraphy contains the study of strata, or layers. Specifically, stratigraphy refers to the application of the Law of Superposition to soil and geological strata containing archaeological materials in order to determine the relative ages of layers. In addition, stratigraphy can tell us much about the processes affecting the deposition of soils, and the condition of sites and artefacts. *GeologicTimePeriod* stores geologic time periods while *GeologicTimeBoundary* stores boundaries for geologic time periods.

Synecology stores information what other species where found from the gathering site. Synecology is the study of entire ecosystems, as opposed to autecology. Synecology focuses on the relationships between species.

4. User extensions

The database model presented here is based on two standards Darwin Core 2 [14] and Access to Biological Collections Data (ABCD) [13] proposed to Taxonomic Database Working Group (TDWG). The Darwin Core 2 (DwC) is a simple set of data element definitions designed to support the sharing and integration of primary biodiversity data. The Access to Biological Collections Data (ABCD) Schema is an evolving comprehensive standard for the exchange of data about specimens and observations (primary biodiversity data) and is based on XML technologies.

While the proposed database attempts to be comprehensive and highly structured supporting Darwin Core fully and most important parts of ABCD, users have different needs. Therefore, it is essential that database is easily extended to users requirements. Thus, we propose extendable database where users can add their own attributes to most of the relations using *Attributes* relation (see Figure 2).

As an example consider again bird example where user wants to add length of the wing for *Larus Argentatus* in *Identifications* relation. Thus he first adds this new attribute to the relation *Attributes:*

Column	Data
AttributeId	102
AttributeName	WingLength
AttributeType	number(4,1)
Description	Length of the wing

Now he adds connection between *Identification* where he has defined species to be *Larus Argentatus* (lets assume that *IdentificationId* is 100) and this new attribute. Thus he adds a row to *UserIdentificationData* relation:

Column	Data
IdentificationDataId	100
Value	45.2
Comment	
AttributeId	102
IdentificationId	100

Additionally, different projects can have different needs and views to the database. These can be expressed using *Projects*, *Members* and *AttributeRef* relations. As an example consider query where minimum, maximum and average wing length of the *Larus Argentatus* is selected.

5 Conclusions

The subtleties and complexities in taxonomy and collections are difficult to explain in a short space. Taxonomy has evolved over a long period and the wealth of knowledge developed into a complex area. We have presented our approach to modelling and querying taxonomic and observation data, which is based on the conceptual taxonomic model.

Proposed structure is able to handle a multitude of tasks, among them conceptual circumscription of taxa, acceptance, synonymy, classification in alternative taxonomies, and systematic taxa sequence. However, the self-set challenge that the present model is designed to meet is to present a compact unified solution to accommodate all aspects of non-descriptive, non-distributional taxonomic database.

References

[1]. Altman, R. B., Editorial: Building successful biological databases. *Briefings in Bioinformatics 5*(1), pp. 4, 2004.

[2]. Berendsohn, W. G., The concept of potential taxa in databases. *Taxon 44*, pp.207-212, 1995.

[3]. Berendsohn, W. G., A taxonomic information model for botanical databases: the IOPI model. *Taxon 46*(4), pp.283-309, 1997.

[4]. Berendsohn, W. G., Names, Taxa, and Information. In *Proceedings of the Taxonomic Authority Files Workshop*, 1998.

[5]. Berendsohn, W. G., Anagnostopoulos, A., Hagedorn, G., Jakupovic, J., Nimis, P. L., Valdes, B., Guntsch, A., Pankhurst, R. J., and White, R. J. A comprehensive reference model for biological collections and surveys, *Taxon 48*, pp. 511-562, 1999.

[6]. Birney, W., and Clamp, M. E., Biological database design and implementation. *Briefings in Bioinformatics 5*(1), pp. 31-38, 2004.

[7]. Discala, C., Benigni, X., Barillot, E., and Vaysseix, G., DBcat: a catalog of 500 biological databases. *Nucleic Acids Research 28*(1), pp. 8-9, 2000.

[8]. Francki, R. I. B. , Fauquet, C. M., Knudson, D. L., and Brown, F., Classification and nomenclature of viruses. *Archives of Virology Supplement* (2),pp. 1-445, 1990.

[9]. Greuter, W., McNeill, J., Barrie, R., Burdet, H., Demoulin, V., Filgueriasnigstein, S., Nicolson, D. H., Silva, P. C., Skog, J. E., Trehane, P., Turland, N. J., Hawksworth, D. L., International code of Botanical Nomenclature. In *Proceedings of the Sixteenth International Botanical Congress* St. Luis, Missouri. *Regnum Vegetabile 138*. Koeltz Scientific Books, 1999.

[10]. International Commission on Zoological Nomenclature (ed.), International Code of Zoological Nomenclature. Fourth Edition. The International Trust for Zoological Nomenclature, 1999.

[11]. Kazic, T., Coe, E., Polacco, M. L., and Shyu, C.-R., Whither Biological Database Research? *OMICS 7*(1), pp. 61-66, 2003.

[12]. Snealth, P. H, A. (Ed.), International Code of Nomenclature of Bacteria, 1980 Revision, 1992.

[13]. Task Group on Access to Biological Collection Data (ed.), Access to Biological Collection Data (ABCD), http://www.bgbm.org/TDWG/CODATA/ABCD-FirstReferenceImplementation.htm.

[14]. Taxonomic Databases Working Group (ed.), Taxonomic Databases Working Group: Darwin Core 2, http://darwincore.calacademy.org.

[15]. http://www.cs.helsinki.fi/u/jplindst/biodb.html

Figure 2. Database schema.