Lagascalia 19(1-2): 667-672 (1997)

A FRAMEWORK FOR BIOLOGICAL INFORMATION MODELS

WALTER G. BERENDSOHN, ANASTASIOS ANAGNOSTOPOULOS, JASMIN JAKUPOVIC, PIER-LUIGI NIMIS & BENITO VALDÉS

Abstract

A general reference system for biological information processing needs to address information structures of the biological research process as well as those of the materials used in the study. The present article presents a framework for such a reference system by means of a categorization of biological data structures into "biological study", "descriptor", and "biological object". An entity relation model is presented to clarify the internal structure of biological objects, which may stand for scientific names, taxa, units in a natural history collection, or even geographical sites or ecological categories.

Introduction

Botanists as well as researchers in many other branches of science use materials obtained from organisms as the base of their studies and subsequently store their research results in databases. Their source material often consists of samples taken from natural history collections, or it is vouchered in such collections to ensure proper identification of the organisms. This material includes plant, animal, or paleontological specimens in natural history collections, microbial strains in cultive collections, living plants or animals in botanical or zoological gardens, chemicals in natural product collections, etc.

In almost all of these fields, Europe owns the most extensive collections of such specimens worldwide. To facilitate access to these resources, electronic inventories must be created. To ensure interoperability of present and future databases, common data models and standards are needed.

At the outset of the project, the objective of CDEFD ("A Common Datastructure for European Floristic Databases") was to develop project-independent structures to be used in the design of floristic databases and databases including floristic data. In the course of the project, this was extended to include biological collections in general, because it was realized that all objects or samples obtained from organisms share the same core data structure. The resulting datamodel will be published (BERENDSOHN & al., in press) and it is available on the World Wide Web.

The present article is to describe the wider context of this model. Some consideration is given to the basic types of electronic data produced by biological investigations ("Studies") and to the basic classification of source materials ("Biological Objects") they act upon.

Modelling categories

Studies and descriptors.- In essence, any biological study, be it experimental or observational, involves examination of groups, individuals, or parts of organisms, or of materials originally derived from organisms. Results of studies may take the form either of a series of values for defined parameters, or of unstructured textual information conveying a law or abstraction derived from the facts revealed during the investigation. Although unstructured textual information may also be stored, research databases lend themselves principally to the storage of information which may be expressed as parameter and value (characters and character states, "Descriptors"), because in this area the strength of electronic processing of large datasets takes effect. For the purpose of this article, comparative studies may be regarded as processing of information gathered –and stored– as the result of individual studies.

From the point of view of the information modeller, the "Study" provides the framework to link descriptors with the organismic object of the study. Methods, persons, or bibliographical data related to the investigation are here registered. CDEFD used karyological investigations as an example to illustrate a complex type of study by means of a detailed information model (Berendsohn & al. 1996). In contrast to the extensive descriptor structures described there, Study structures may also take a rather simple form, e.g. in the recording of presence/absence data for floristic mapping. The function of the Study - entity may even be reduced to a link to a bibliographic reference. In any case, the Study represents the description and the result of an investigative process, which acts upon Biological Objects.

The "Biological Object". - A Biological Object is here defined as an entity-type or supertype in an information model. It provides a gateway between investigative or descriptive data and the organismic objects a defined study investigates. In a distributed database environment, the Biological Object may be used to provide a simplified view of the model to people who want to use a system based on it without knowing the intricacies of its design. Of course, in a relational system external information may be linked to many points, i.e., any of the entities may contain a key which can be used to link information in another, external entity to it. However, a defined interface has to be provided to people who either do not want to dive too deeply into the model's design, or who do not want to link their information too intimately with the underlying system. For these, the Biological Object serves as a "switched" interface to link their information e.g. to the collection and taxonomic information covered by the IOPI and CDEFD models.

In the course of the investigation, the object of the study is initially always a material one: The animal which is observed, the soil sample containing microorganisms, the cell culture, or the tree in the forest under investigation. In the course of formulating results, the biological object may become an abstraction, e.g. a plant name representing a taxon, or an ecological category (a site investigated, a syntaxon, a biogeographical classification unit, etc.). Fig. 1 depicts the principal subdivision of Biological Objects.

Ecological categories.- Because of the great diversity of investigative approaches and classification systems involved, it may prove impossible to provide a generalized information model for ecological categories. However, as in comparative studies, the

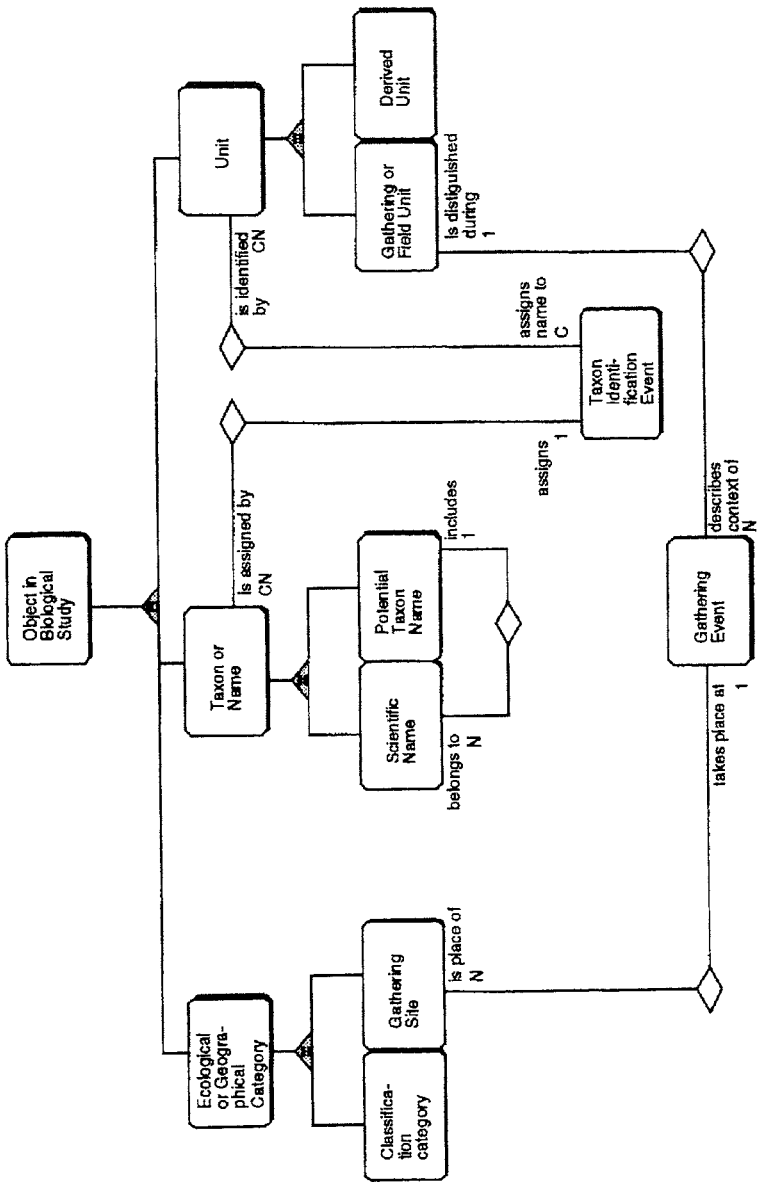

Fig. 1. Extended Entity-Relation Diagram for Biological Objects. In the diagram, entity-types are represented by rectangles. They may be thought of as representing a table or a system of tables in a relational database. The triangles represent exclusive classification relationships (subtyping). An Object in a Biological Study (supertype) may be either of the following subtypes: an Ecological or Geographic Category, a Taxon or Name, or a Unit. Other relationships are read along the connecting lines, starting with the entity-type name, followed by the descriptive text nearest to it, then the cardinality (i.e., how many instances of the second entity-type are referred to the instance in the first one) and, finally, the name of the second entity-type. The cardinality may be "1" (exactly 1), "C" (0 or 1), "N" (1 to many), or "CN" (0 to many). The "C" is for conditional relationship, i.e. it is possible that no instance is referred to. For further details on modelling techniques please refer to Berendsohn & al. (in press).

basic data needed for such an investigation may well take the form of individual studies on organisms, related to one another, or related to a specific position in space and time (a site). This area is in urgent need of more basic research.

Names and taxa.- Using taxon names as biological objects in a study may be problematic, because a name may represent several concepts of a taxon. At least two cases must be distinguished: the name is provided on its own, or in the form of a "Potential Taxon" (BERENDSOHN, 1995), i.e. bibliographic or other references are provided which clarify the taxonomic concept represented by the name. Much thought has been given to the structure of taxon name information (e.g. BEACH & al., 1993), standards (BISBY, 1994) and detailed information models have been developed (e.g. BERENDSOHN, 1994).

UNITS.- CDEFD's main concern were material biological objects, which are referred to as Units. Fig. 1 illustrates the principal relationships of units in the CDEFD model, details can be found in Berendsohn & al. (in press). Two main categories (subtypes) of Units are recognized:

The "Gathering or Field Unit" represents the biological object in its original location, unaltered by the investigative process. The entity-type "Gathering Event" provides information concerning the Who and When of the observation of the object, and it links units to the "Gathering Site", which in turn (directly or indirectly) provides all relevant locality data.

The second subtype is the "Derived Unit". Units may be derived from other Units, both, field and derived units. E.g. a microscopic slide may be prepared from a fungus found on a leaf which has been taken from a herbarium sheet. Each of these items (herbarium sheet, leaf with fungus spores, microscopic slide) do form a derived unit which may have distinct information attached to it (e.g. the taxonomic determination differs for fungus and herbarium sheet; the storage location may be different for all three items, etc.). By means of the gathering site which was stored with the original unit (the tree from which the herbarium sample was taken), the original location can be named for all these items. A Derived Unit is the product of a "Derived Unit Creation Event", which may be a process of curation, preparation, cultivation, or a transfer event, which creates one or more Derived Units from one (or rarely more) parent unit(s). As the model allows multiple iterations of this operation, it permits to store highly iterative processes, such as cultivation and propagation histories.

Conclusions

Databasing biological information is a highly complex task and any attempt to provide an easily comprehensible model are bound to fail - at least when attempts are made to inter-connect systems. The present article gives a highly condensed view of the CDEFD model and its embedding in general structures. We opine that biological data in general are covered by this view - in fact, the unit concept already permits extension to many other fields represented in natural history collections. A similar generalization may also be possible for taxonomic and nomenclatural systems used in other fields, e.g. in geology or synthetic chemistry. In this area, as well as concerning

ecological categories, more basic research efforts such as the one undertaken by CDEFD are needed.

References

BEACH, J. H., S. PRAMANIK & J. H. BEAMAN (1993). Hierarchic taxonomic databases, in R. FORTUNER (ed.), *Advances in computer methods for systematic biology: Artificial intelligence, databases, computer vision*, 241-256. Baltimore.

BERENDSOHN, W. G. (1994). IOPI World Vascular Plant Checklist. A CASE Model of Checklist System Data, in K. WILSON (ed.), *International Organisation for Plant Information (IOPI), Global Plant Checklist project plan, version 1.2.*, Sidney. (An updated draft version is available on the World Wide Web under http://userpage.fu-berlin.de/~wgb/7301root.htm.)

—— (1995). The concept of "potential taxa" in databases. Taxon **44**: 207-212.

——, J. GREILHUBER, A. ANAGNOSTOPOULOS, G. BEDINI, J. JAKUPOVIC, P. L. NIMIS & B. VALDÉS (1996): A comprehensive datamodel for karyological databases. *Plant Syst. Evol.*

——, A. ANAGNOSTOPOULOS, G. HAGEDORN, J. JAKUPOVIC, P. L. NIMIS & B. VALDÉS (in press): The CDEFD information model for biological collections, in *Proceedings of the European Science Foundation workshop "Disseminating Biodiversity Information", Amsterdam, 25. - 27. 3. 1996.* (Pre-print available on the WWW under http://userpage.fu-berlin.de/~wgb/CDEFD/cdefd.html.).

BISBY, F. A. (1994). *Plant Names in botanical databases.* International Working Group on Taxonomic Databases for Plant Sciences (TDWG) Plant Taxononomic Databases Standards No. 3. Pittsburgh.

Addresses of the authors:

Dr. W. G. Berendsohn, Botanical Garden and Botanical Museum Berlin-Dahlem, 14191 Berlin, Germany; Dr. A. Anagnostopoulos, 121 Stratigou Dagli str., 11145 Athens, Greece; Dr. J. Jakupovic, Institute for Organic Chemistry, Sekr. C3, Technical University of Berlin, Strasse des 17. Juni 135, 10623 Berlin, Germany; Prof. Dr. Pier-Luigi Nimis, Biology Department, University of Trieste, 10 Via Giorgieri, 34127 Trieste, Italy; Prof. Dr. Benito Valdés,Departamento de Biologia vegetal y ecología, Universidad de Sevilla, Apdo. 1095, 41080 Sevilla, Spain.