DATA AND INFORMATION MANAGEMENT AND COMMUNICATION

Walter G. Berendsohn
Freie Universität Berlin, Botanic Garden and Botanical Museum Berlin-Dahlem, Germany

Keywords: Biodiversity informatics, biological informatics, bioinformatics, taxonomic databases, systematics, taxonomy, biological collections, natural history collections, herbaria, botanical gardens, zoological gardens, culture collections.

Contents

Summary

Developments in data and information management and communications for biodiversity research are exemplified by an account of progress and perspectives in biodiversity informatics. The information domain, its networking techniques and applications are described and an attempt is made to deduce perspectives from current developments in this fast moving field.

1. Introduction

The topic of data and information management and communication for biodiversity research is here exemplified by a description of the emergence, current state and

perspectives of "biodiversity informatics", a relatively new branch of applied informatics. As a component of "biological informatics", it focuses on individual organisms, populations, taxa (named groups of related organisms), and their interaction. This is the organismic level of biodiversity, as opposed to the molecular level, for which the term "bioinformatics" is commonly used, while "environmental informatics" places the focus on ecosystems and higher-scale interaction with abiotic factors.

Biodiversity informatics is most closely allied with systematics and taxonomy, the biological sub-discipline studying the variation of organisms with its causes and consequences, and which use the result to provide the classification system for organisms. Taxonomists are thus enabling us to identify organisms and to use the system as an index to data on a wide variety of properties of organisms.

A wide variety of methods are used by systematists, and research on species interactions or functional aspects of taxa in ecosystems as well as (molecular) bioinformatics techniques are playing an increasing role, e.g. in the elucidation of higher level relationships between organism groups. However, the core data are obtained by the investigation and observation either of organisms in the field or of samples of organisms housed in natural history museums and living collections. Biodiversity informatics encompasses the bulk of the data and information used and generated by systematics research:

- primary biodiversity data, such as the primary occurrence data of organisms in the form of geo-referenced observations of species in the field, or in the form of samples in biological collections (and thus representing verifiable information as to the nature of the species), and all descriptive, experimental or analytical data derived directly thereof, and
- synthesized taxonomic information such as the definition, description and circumscription of taxa, groups of organisms (mostly) thought of as representing groups of related individuals based on the evolution of organisms, representing nodes or endpoints in phylogenetic trees.

This is a highly complex information domain, characterized by enormous and highly inconsistent ontologies that have evolved over centuries, complex nomenclatural traditions that lead to an unclear distinction between hypothesis-driven classification and the index system, the general lack of unique identifiers, and dependence on a number of auxiliary information domains some of which are highly complex in themselves (e.g. geographical information). Maybe it is because of these difficulties that biodiversity informatics has come to provide some innovative technical solutions.

2. Scope of the information domain in biodiversity informatics

Although biodiversity informatics focuses on the organismic level of biodiversity, both the molecular and the ecosystem levels represent strongly overlapping information domains. The ultimate aim will be to overcome any apparent division between these disciplines. Equally, the following subdivision of the domain itself is largely artificial and has been established for pragmatic reasons (such as the delimitation of standardization efforts) rather than representing a real subdivision. However, it has

proven useful for the discussion of the biodiversity information infrastructure to distinguish collection data, descriptive data, nomenclatural, and taxonomic data.

2.1. Primary biodiversity records: biological collection data

The term "Biological Collection" was coined to group together living collections (e.g. botanical or zoological gardens, and microbial culture collections), natural history collections (mainly in museums and universities), and data collections representing occurrence records, such as used in faunistic and floristic mapping projects, surveys and species-level monitoring. Information structure research conducted in the early 1990s concluded that with respect to data structures the similarities between these domains by far outnumber the differences.

Biological collections have been called the Archives of Biodiversity. An estimated 2-3 billion specimens are held in natural history collections world-wide. Each specimen is a substantiation of past occurrence of an organism at a defined time and geographic place. With their label or field book data, many specimens provide information ranging from ecological and morphological details to alimental, medicinal, and cultural uses.

Species occurrence records, such as those created by floristic or faunistic mapping projects, environmental impact studies, ecological research, etc. may represent a valuable source of such data as well (although they are mostly restricted to presence/absence data and their taxonomic information content cannot be verified by inspection of a deposited specimen). The amount of observation records in existence has yet to be established.

Biological collection data have been thoroughly analyzed and modeled. The Taxonomic Databases Working Group (TDWG), a community driven standardization body working on the topic since 1985, has presented two standards for collection data, which are in use to provide access to collection data: the "ABCD schema" (Access to Biological Collection Data) provides a comprehensive set of data elements aiming at full cover of collection data, including living collections (e.g. zoological and botanical gardens), natural history collections (e.g. herbaria), and observation records and it provides a detailed treatment of provider rights, IPR, and copyright statements.

Furthermore it allows alternative text representations for some of the highly structured data items to encourage potential providers to take part in information networks even if their collection databases are less atomized.

The second schema, the "Darwin Core", presents a much reduced but compatible set of data items mainly geared at specimen collections, and which is already widely used. In its development, ABCD has drawn on several exchange formats which have already been in use in some communities (plant genetic resources, culture collections, herbaria, botanical gardens), and on information models such as the comprehensive one published by the BioCISE project (Resource Identification for a Biological Collection Information Service for Europe).

The main features of this model may serve to illustrate the data dealt with in this part of

the biodiversity informatics domain: At the center of the model is the "unit", i.e. the individual specimen or observed organism. The term "specimen" can often be used as a synonym of unit, however, it lacks a precise, context independent definition, and a single specimen can represent a number of units.

It is also normally perceived in a narrower sense compared to the unit, not including observation records, although these represent identical information structures, the principal difference consisting of the existence of additional data items relating to physical management of the specimen in a collection. Any object containing, being, or being part of a living, petrified, or conserved organism is considered a unit as soon as a record of it is created. Usually, a unit is gathered (observed or collected) in the field. Derived units may recursively emerge from it through specimen processing, breeding or cultivation.

Directed relationships between units may exist ("Associations", e.g. host/parasite), or units may be grouped together ("Assemblages", e.g. a herd, nest and eggs). Gathering events (who collected or observed, when, in what context), gathering site (location, geographical and ecological features), specimen management (acquisition, accession, storage, preservation, exchange, ownership), and taxonomic or other identifications (who identified when as what) relate to the unit. Further information on the unit may include age, stage, and gender, and other descriptive information of any kind.

The latter is not considered part of the collection sub-domain and treated only in a generalized form; descriptive data form their own sub-domain, traditionally and because they apply to both individual units and taxonomic groups (taxa). These, in turn, are included only in so far as they represent the result of an identification. Synthesized taxon-related information such as species distribution or indicator value, and synonyms etc. is not considered to be part of the unit domain.

The International Committee for Documentation of the International Council of Museums (ICOM-CIDOC) has compiled a conceptual reference model which now has become an ISO standard. This is a "formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information" and thus goes beyond the domain of biological collections.

2.2. Collection-level data

There is a second sub-domain directly related to biological collections. This is data describing entire collections of units, i.e. institutions holding collections such as herbaria and natural history museums, but also survey and mapping projects.

Collection-level data cover information about the institution (name, address, contacts), categorizations (e.g. "culture collection"), descriptive keywords (mainly taxonomic and geographic), other general content and storage characteristics, as well as IPR statements and administrative properties (e.g. access restrictions). Where already in place, means to electronically access the collection catalogues (i.e. unit-level information) may be included, and this gives the collection catalogue the quality of a web service register and an index for data discovery.

Access to collection-level data can be considered as a first step on the way to a common access system and the data has been collected in different ways. For the biodiversity community on the one hand any of the aforementioned standards for unit-level data also contains a section describing the location of the object, ownership, etc. that essentially represents collection level data. On the other hand, a first standard format to describe biological collections has been developed in the context of the BioCASE project. This is presently taken forward by TDWG and GBIF in an attempt to provide a general standard for organizations providing biodiversity data.

-
-
-

Bibliography

Anonymous 2003. GBIF Strategic Plan. Secretariat of the Global Biodiversity Information Facility. Kopenhagen. [Provides an overview of GBIF's vision, tasks and organization*]*.

Berendsohn W.G. (1997). A taxonomic information model for botanical databases: the IOPI model. Taxon 46: 283-309. [Apart from presenting a concept-oriented taxonomic information model, this article provides an overview of the relevant literature. Also available under www.bgbm.org/BioDivInf/Docs/IOPI_Model/IOPI_Model.pdf].

Berendsohn W.G., Anagnostopoulos A. Hagedorn G., Jakupovic J., Nimis P.L., Valdés B., Güntsch A., Pankhurst R.J. and White R.J. (1999). A comprehensive reference model for biological collections. Taxon 48: 511-562. [This article covers unit-level collection data and provides an extensive list of relevant information models and standards for this domain. Also under www.bgbm.org/biodivinf/docs/CollectionModel].

Güntsch A. (2004). Globale Netze der Kooperation bei der Sacherschließung im naturkundlichen Bereich. Pp. 25-33 in: Sieglerschmidt, J. (ed.): Regelwerke für die Sacherschliessung. Workshop electronic imaging and the visual arts. Berlin. [Provides an overview of the networking techniques developed for biodiversity informatics. Also available under titan.bsz-bw.de/cms/service/museen/publ/eva2004*]*.

Scoble M.J. (in press). Unitary or unified taxonomy? In: Godfray H.C.J. and Knapp S. (ed.): Taxonomy for the 21st century. Philosophical Transactions of the Royal Society (Biological Sciences). [Some ideas on the transformation of the taxonomic work process towards a web-based approach are here critically examined].

Biographical Sketch

Prof. Dr. Walter G. Berendsohn, Freie Universität Berlin, ZE Botanischer Garten & Botanisches Museum Berlin-Dahlem (BGBM)

1981-1985 Studies of Biology at the University of Hamburg, University of the West Indies (Mona) and the Freie Universität Berlin

1985 Diploma degree, Freie Universität Berlin

1987 to 1990 Research director of the Jardín Botánico La Laguna, El Salvador, C.A.

1990 PhD, Freie Universität Berlin

1990	Curator Botanical Garden and Botanical Museum Berlin-Dahlem (BGBM)
1991	Senior Curator BGBM, Head, Electronic Data Processing and Documentation
1998	Head, Dept. of Biodiversity Informatics and Laboratories, BGBM
2000	Director and Professor.

Research priorities: data structure research in biodiversity sciences; dendroflora of Central America. Active in the organisation of national and international projects in the field of biodiversity informatics and botanical inventories, inter alia in the context of the Global Biodiversity Information Facility (GBIF), the Taxonomic Databases Working Group (TDWG - IUBS Commission for Taxonomic Databases) and CETAF (Consortium of Large Scale Taxonomic Research Facilities). Coordinator of EU projects under Framework Programmes 3 to 6.