Organizing Our Knowledge of Biodiversity

by Hilmar Lapp, Robert A. Morris, Terry Catapano, Donald Hobern and Norman Morrison

EDITOR'S SUMMARY

Though natural history collections are long established and numerous, data on biodiversity is sparse, poorly developed, inconsistent and rarely digitally preserved, and their providers are often inaccessible. The Global Biodiversity Information Facility (GBIF) and the Taxonomic Databases Working Group (TDWG) for Biodiversity Information Standards are leading efforts to overcome such barriers. The GBIF has collected over 200 million records formally describing natural history specimens from hundreds of sources, while also serving the rapidly growing online community of amateur field observers. The two organizations face challenges arising from ambiguity of taxonomic names and the need to make voluminous and critical historical collections available and to develop standard metadata vocabularies that can be understood and used by all. The combination of science informatics tools, including a robust vocabulary and linked data conventions, with a networked, active user community will enable effective biodiversity data management over the long term.

KEYWORDS

taxonomies

collaborative indexing

metadata

biology

informatics

Classifying the diversity of life on earth, as well as collecting specimens of the living world to document it, are old endeavors. The Swedish botanist Carl Linnaeus first published his seminal work *Systema Naturae*, which established the binomial nomenclature of living things still valid to this day, in 1735. Long before digital record-keeping was invented, generations of naturalists, including Linnaeus himself, collected specimens, archived them for perpetuity in natural history museums around the world and painstakingly described in natural language what they had found and where. According to recent estimates [1], the world's natural history museums hold an estimated three billion museum specimens, most with no digital record of any kind. Until only a few years ago, the digital records that had been created were not accessible from outside the institutions that kept them. Thus, so far most of the efforts in developing knowledge organization systems (KOS) for biodiversity have focused on digitally mobilizing the most fundamental data on biodiversity, such as which species has been observed where and when. By contrast, other fields of biology, for example those concerned with investigating genetic model organisms (such as mouse and fruitfly) and human disease, have continuously been at the forefront of building and applying knowledge discovery-oriented KOS.

Corresponding author Hilmar Lapp is assistant director of informatics, National Evolutionary Synthesis Center (NESCent). He can be reached by email at hlapp<at>nescent.org.
Corresponding author Robert A. Morris is professor emeritus of computer science, University of Massachusetts/Boston and biodiversity informatics staff, Harvard University Herbaria.
He can be reached by email at morris.bob@gmail.com.
Terry Catapano is librarian, Columbia University Libraries and vice president, Plazi.
Donald Hobern is director, Atlas of Living Australia and former TDWG chair.
Norman Morrison is ontologies and data standards manager, Natural Environment Research Council, Environmental Bioinformatics Centre (NEBC) & The University of Manchester, United Kingdom.

CONTENTS

< PREVIOUS PAGE

NEXT PAGE >

NEXT ARTICLE >

In this article, we highlight some of the challenges that the nature and history of biodiversity data has brought about for digital knowledge organization, and we suggest paths forward towards taking better advantage of the full array of KOS. The present work arises from a more comprehensive report [2] that the authors presented to the Global Biodiversity Information Facility (GBIF), one of the world's foremost aggregators of biodiversity occurrence data.

The Role of KOS in Mobilizing Biodiversity Knowledge

The world's biodiversity knowledge in the form of specimen records is preserved in a highly distributed manner in natural history museums around the globe. Most of these are not digitized, and few of those that are were born digital. Early efforts in developing KOS to mobilize this wealth of information for use in a digital world centered on enabling the keepers of natural history collections to digitally expose their specimen information records and on promoting consistent databasing of botanical and zoological information within and between different institutions. Many of these early resources are documented by the Biodiversity Information Standards Organization's (TDWG) [3] Subgroup on Biological Collection Data [4]. Even though some of these resources are still in use by biodiversity informatics projects, very few are available as machine-accessible vocabularies. The notable exceptions are several important web-accessible authority files and databases, which include names of species, journal titles, natural history collections, authors and institutions.

Natural history collections are not only numerous, but also highly heterogeneous in the resources available to them. As a consequence, bringing biodiversity information online on a purely distributed basis has proved to be fraught with issues including inconsistent application of standards and metadata vocabularies as well as providers simply being out of service on an unpredictable basis. These barriers to accessing biodiversity information on a large scale have been dramatically lowered by web-based biodiversity data resources that aggregate, validate and cache data from the distributed network of original data providers. One of the most prominent of these aggregators is the Global Biodiversity Information Facility (GBIF) [5], an intergovernmental agency with secretariat in Copenhagen. GBIF presently serves over 200 million metadata records for natural history specimens and field observations, which are aggregated, indexed and cached from over 300 sources.

While online data aggregators have helped tremendously in increasing the amount of biodiversity information that is digitally accessible, the meaning of the aggregated data is often ambiguous to users of such resources. There is typically no guarantee for the consistency of how various metadata elements get mapped from the source data models to the aggregated model. These metadata include elements that are critical to the semantics of a data point, such as the specification of time, place, individual, organization, data gathering protocols and taxonomic identification. KOS in the form of controlled vocabularies, domain ontologies and authority databases could prove to be pivotal in alleviating the semantic ambiguities of data aggregation and its result. Consequently, approaches to developing such KOS started in TDWG as early as 1990, with an initial focus on achieving community agreement on the definition and use of controlled vocabularies.

Data aggregation at the breadth and scale of GBIF has also increased awareness among biodiversity information scientists that the balance of sources for biodiversity information is shifting towards born-digital field observations provided by a large and worldwide, distributed network of skilled lay observers (citizen scientists). Traditionally, species observations were tied to archived physical specimens, called vouchers, and were primarily the occupation of professional researchers. More recently, online community sites such as eBird and eNaturalist have allowed any nature enthusiast who owns a smartphone to share, describe and discuss biodiversity field observations. Field observation records already account for about 40% of GBIF's aggregated holdings and are its fastest growing component. In consequence, the KOS developed for biodiversity information will not only need to be usable by professional scientists, but also by citizen scientists, or otherwise a significant fraction of the online biodiversity knowledge may be precluded from full scientific exploitation due to unresolved issues of data quality in the absence of a physical voucher for validation.

CONTENTS TOP OF ARTICLE < PREVIOUS PAGE NEXT PAGE > NEXT ARTICLE >

The biodiversity informatics community is also concerned about describing and accessing a variety of ancillary information that might be associated with a particular specimen or occurrence record. This ancillary information will often be critical for analyzing processes and patterns in biodiversity and enlarges the scope of potentially relevant data to include a broad range of observed measurements about the biotic and abiotic aspects of the environment. For example, when examining patterns in the global abundance of some species, information regarding the co-situated precipitation, frost-days, soil type, land use and so forth could be important as parameters for analysis. Thus, biodiversity informatics needs will ultimately converge with the activities of other earth and environmental sciences that rely on multidisciplinary data for integrated or holistic understanding.

Challenges

Taxonomic names. Datasets about species frequently use the Latin binomial scientific name as a natural primary key for species. However, the names of species change over time due to taxonomic revisions. Such name changes have to be published in literature to become valid, and nowadays published name changes are also digitally represented in authoritative online databases of names. Thus, it is the combination of the Latin scientific name and the authority or publication that made it valid that unambiguously identifies a taxon, or more precisely, the taxon concept. Yet, a large proportion of species data lacks the metadata necessary to identify the taxonomic authority or publication implied by the data author when the taxon was assigned. This deficiency presents formidable challenges to determining if two records using the same scientific name are actually referencing the same species.

In practice this problem is compounded by the fact that identifying organisms is a non-trivial act and open to considerable debate. Taxonomic judgment on the boundaries of taxa (species or higher level taxonomic classifications, such as genus, family or order) varies already. The ability to then identify an organism to a recognized taxon concept and to provide an unequivocal reference to that concept is equally difficult. As a result, most of the data available for integration have an unreliable connection to the taxon to which they have been identified, even though taxonomy should be the core axis on which scientists rely for subsequent inquiry. Without robust solutions to these issues, many practitioners invariably trust their experience more than computer-based inferences.

Criticality and size of historical data. As mentioned above, in contrast to many other branches of natural science, centuries-old data and publication is critical to disambiguating biodiversity knowledge, specifically the names and descriptions of species. The specimen records held by the world's more than 4,000 natural history collections number in the billions, and the vast majority of them have not been digitized in any form. Producing digital metadata and images even for only those specimens held in U.S. institutions is expected by the U.S. National Science Foundation (NSF) to require a well-funded 10-year effort [6].

A vast store of legacy biodiversity literature has until very recently also not been accessible in digital form. In the last four years, a worldwide effort under the leadership of the Biodiversity Heritage Library has produced scans of nearly 33 million pages of biodiversity publications that are out of copyright [7]. Nevertheless, reliable OCR and automated semantic markup of literature digitized in this way remains a subject of active informatics research. To lessen the difficulty of future semantic markup for prospective publishing, a taxonomic markup extension to the Journal Archiving and Interchange Tag Suite of the U.S. National Library of Medicine and National Center for Biotechnology Information is being developed [8].

Design and management of common vocabularies. Standard metadata vocabularies are key to effective data exchange and reuse. These benefits can be achieved without much or any hierarchical structure, and by simply describing the intended semantics in the textual definitions of concepts. Flat vocabularies are also familiar to many scientists from the routine activity of publishing, for example in the form of glossaries. However, software tools cannot use textual descriptions to understand, let alone enforce, any semantics. This is true even for simple semantics such as a requirement that two data fields both either be present or absent because they form necessary parts of a piece of information. For example, the Darwin Core [9] vocabulary, a widely used metadata standard for specimen and observation records, cannot express in a machine-interpretable way that a record with a value for latitude but none for longitude cannot be meaningfully compared to others by geographic location. Despite such limitations for KOS tools, documentation of the intended relations among terms can still significantly enhance the usability for humans. Moreover, if vocabularies are developed from the outset such that the semantics of their terms are extensively documented even if they are otherwise flat, they can still be reused later as the terminological foundation for semantically richer domain ontologies.

Our knowledge of biodiversity is changing rapidly. While it may seem fully obvious that ontologies that capture this knowledge therefore need to have mechanisms that allow them to change accordingly and track the input from experts, this requirement is in fact equally valid for flat vocabularies. Even if there is no hierarchy between terms that needs updating, all other properties of terms such as meaning, definition and applicability still need to be able to change in a community-coordinated fashion. Otherwise, a vocabulary risks obsolescence, or, worse yet, "flavors" may arise that are understood by some resources but not by others. The upside of this situation is that social and technical ontology lifecycle-management tools already developed in response to these needs in other ontology-building communities (for example OBO Foundry [10] and the NCBO BioPortal [11] can likely be reused for the same purpose in biodiversity information science. Such tools can then also support modern KOS cyber-infrastructure requirements such as the management of globally unique, persistent, de-referenceable identifiers for the terms.

As a consequence of the focus in biodiversity information management on mobilizing and exchanging data, the development and use of sophisticated deeply hierarchical ontologies, and even more so the application of even simple machine inferencing to biodiversity data, has to-date been largely limited to well-funded collaborations between highly experienced informatics specialists and active biodiversity researchers.

Data deluge. Owing to the decreasing costs of data acquisition and storage, widely distributed data are accumulating throughout the sciences at a prodigious rate. Some of the resulting problems could be alleviated by more broadly employing Semantic Web technologies. Specifically, integrating data from heterogeneous sources can in theory become drastically easier if all sources make their data available in the W3C's Resource Description Framework (RDF) standard, because it is by definition interoperable at least on the syntax level. The Linked Data conventions [12] expand on this potential by not only requiring data and metadata to be in RDF format but also to use URIs (Uniform Resource Identifiers) as common unique identifiers for things that connect datasets to each other and to use common vocabularies wherever possible. Linked Data presents a low barrier to entry for data producers, yet allows generic tools to be built for discovering and consuming or aggregating online data. Linked Data principles have already been adopted in prototype-level applications to expose and integrate biodiversity information and knowledge. These include a regional species occurrence record database (GeoSpecies) and a resource annotating digitized biodiversity publications with the geographical locations, taxa, authors and citations referenced therein. Nonetheless, in the absence of sophisticated domain ontologies against which to reason over the discovered data, it is still difficult for software to decide much about the fitness of such data for any particular scientific inquiry. Such ontologies are beginning to emerge, but still remain the rare exception rather than the norm in biodiversity informatics research.

Which Path Forward?

The development and subsequent management of intellectual or machine frameworks for the organization of fundamental concepts of scientific knowledge comes at a significant cost, both in terms of intellectual input and sustainability. This is no less true for biodiversity science than it is for other fields. By contrast, controlled vocabularies are typically generated with less controversy, have reduced management burdens and often have greater longevity within the domain. Consequently, data aggregators like GBIF and information standards organizations like TDWG must manage a dual velocity problem, helping promote the robust but rapid development and management of controlled vocabularies, while providing vision and support for the more complex and deliberate task of supporting scientific analysis, synthesis and inference on large heterogeneous data sets dispersed around the Internet. Fortunately, this problem is nicely congruent to the current development of science informatics tools and human resources: vocabulary development is well under way, supported by available tools and social organization systems that are usable by domain scientists with only modest support from IT professionals. For the longer term, if past experiences in software engineering are any evidence, some fraction of the domain scientists will turn to knowledge engineering, and some fraction of an increasing population of knowledge engineers will find biodiversity applications fascinating and join the fray. ■

Resources Mentioned in the Article

[1] Wheeler, Q., & Valdecasas, A.G. (2010), Cybertaxonomy and ecology. *Nature Education Knowledge 1*(11), 6.

[2] Catapano, T., Hobern, D., Lapp, H., Morris, R.A., Morrison, N., Noy, N., et al. (February 4, 2011). Recommendations for the Use of Knowledge Organisation Systems by GBIF. Copenhagen: Global Biodiversity Information Facility. Retrieved March 3, 2011, from http://links.gbif.org/gbif_kos_whitepaper_v1.pdf.

[3] *Biodiversity Information Standards Organization [aka Taxonomic Databases Working Group (TDWG)]*: http://tdwg.org/. TDWG promulgates biodiversity data and information exchange standards.

[4] *TDWG Subgroup on Biological Collection Data:* www.bgbm.org/TDWG/acc/Referenc.htm.

[5] *Global Biodiversity Information Facility (GBIF):* www.gbif.org.

[6] Mares, M. A. (August 2010). *A strategic plan for establishing a network integrated biocollections alliance* [brochure]. Available March 3, 2011, at http://digbiocol.files.wordpress.com/2010/08/niba_brochure.pdf. See also www.nsf.gov/funding/pgm_summ.jsp?pims_id=503559.

[7] Biodiversity Heritage Library. (2011). Now online. *Biodiversity Heritage Library* [web site]. Retrieved February 21, 2011 from www.biodiversitylibrary.org/.

[8] Catapano, T. (2010). TaxPub: An extension of the NLM/NCBI Journal Publishing DTD for Taxonomic Descriptions. *Proceedings of the Journal Article Tag Suite Conference 2010 [Internet]*. Bethesda, MD: National Center for Biotechnology Information (US). www.ncbi.nlm.nih.gov/books/NBK47081/.

[9] *Darwin Core:* http://rs.tdwg.org/dwc/index.htm.

[10] *The OBO (Open Biological and Biomedical Ontologies) Foundry:* www.obofoundry.org/.

[11] *The NCBO (National Center for Biomedical Ontology) BioPortal:* http://bioportal.bioontology.org/.

[12] *Linked Data:* http://linkeddata.org/.