INOTAXA — INtegrated Open TAXonomic Access and the "*Biologia Centrali-Americana*"

Anna L. Weitzman
The National Museum of Natural History, Smithsonian Institution

Christopher H. C. Lyal
The Natural History Museum, London, UK

"*Although the 'Biologia' contains the record of such a large number of species, it is but a fragment of what may yet be obtained. The whole work must be looked upon as only a contribution to our knowledge of the subject, and I hope it may be an incentive to others to carry it further.*" - F. Ducane Godman, F.R.S., F.Z.S. (Proc. Zool. Soc. London, Sept. 1916)

"*The current biodiversity crisis is very much an information crisis*" - Christoph Häuser (Sustainable use and conservation of biological diversity – a challenge for society, 2004).

Summary

The collaborating institutions[1] are creating a model for global access to the data necessary for understanding the world's biota. INOTAXA ('INtegrated Open TAXonomic Access') will be a web workspace in which taxonomic descriptions, identification keys, catalogues, names, specimen data, images and other resources can be accessed simultaneously according to user-defined needs. As it will use a distributed data model, it will allow access to data held in multiple servers globally if indexed through a registry such as operated by GBIF (Global Biodiversity Information Facility[2]). If, in the future, the various nomenclatural Codes permit web publication of new taxonomic names and acts, these could be integrated with the rest of the body of taxonomic knowledge through INOTAXA.

INOTAXA is built on a set of interoperable XML schemas. To ensure access to data wherever they are held, INOTAXA is working with TDWG (Taxonomic Databases Working Group[3]) to ensure that it uses, and is interoperable with, globally-accepted standard schemas. These will allow external interoperability with, for example, GBIF and access to GBIF-mediated data. INOTAXA will also provide seamless access from the content to other systems holding biodiversity and taxonomic information as these become available.

The INOTAXA project, although newly-named, was conceived and identified as a priority in a Mellon-funded meeting in 2002 at which a number of major museums and herbaria determined to demonstrate the potential of combining information, literature and research data held within their collections[4]. As a testbed for their ideas they determined to focus on Mesoamerican biodiversity, building on a major literature resource, the important and out of print scientific work the *Biologia Centrali-Americana* (BCA). The BCA was derived from scientific surveys and

---

[1] **A joint project of:** Smithsonian Institution (National Museum of Natural History, Smithsonian Institution Libraries, and Smithsonian Tropical Research Institute); Natural History Museum (London); Missouri Botanical Garden; National Commission for the Knowledge and Use of Biodiversity, Mexico (CONABIO); Instituto Nacional de Biodiversidad, Costa Rica (INBio); American Museum of Natural History; Harvard University (Museum of Comparative Zoology); Royal Botanic Gardens, Kew; Museo Entomologico de Leon, Nicaragua; Global Biodiversity Information Facility.

[2] http://www.gbif.org

[3] http://www.tdwg.org

[4] AMNH, NHM, RBGK, Missouri Botanical Gardens, NMNH, STRI, Smithsonian Institution Libraries. See http://www.sil.si.edu/digitalcollections/bca/documentation/proposal.pdf

explorations conducted during the latter part of the 19th and early 20th centuries. Many of the leading biologists of the time provided specimens and descriptions for the many volumes, which includes descriptions of more than 50,000 species of animals and plants. The illustrations are, in many cases, the only images that exist of the biota of the region.

In the first project phase the BCA was digitized and made public, and the 'electronic *Biologia Centrali-Americana*' (eBCA) now provides the single body of digitized taxonomic work available through the Internet[5]. At the same time, the project team developed an XML schema for taxonomic literature, 'taXMLit'[6], which is now being developed as a TDWG standard.

The current project phase is the development of the INOTAXA prototype. This will use the resources developed in phase one, and demonstrate the potential of interoperable XML schemas to link data of different types and from different sources, including different taxonomic treatments of the same taxa, specimen data, classifications and images. While in the prototype the data will all be maintained on a single server, in phase 3, the full implementation of INOTAXA, they will be accessed in a truly distributed system. The prototype will serve data for only a subset of Mesoamerican taxa; the full version will include all animal and plant groups world-wide.

Background

A repeated message from those interested in conservation of biodiversity around the world, especially those in biodiversity-rich but resource-poor countries, is the need for taxonomic information (see Blackmore 2002; Raven, 2004; Wilson, 2003 and references). This is necessary for a wide range of environmental management and conservation purposes (e.g., managing endangered and protected species; biodiversity conservation; sustainable development; managing agricultural and other pests, invasive species, disease vectors and pathogens; monitoring hazards such as bird strikes on airplanes); as well as being a basic tool for education and enjoyment of the natural world. This issue has been identified as a part of the 'taxonomic impediment' - the lack of taxonomic information, skills, personnel and capacity inhibiting many developing countries from implementing policies and practices of sustainable management and conservation of biodiversity. In particular, under the Convention on Biological Diversity (CBD[7]), the Global Taxonomy Initiative (GTI) Work Programme[8] highlights the need to make available the contents of taxonomic literature and details of material held in collections outside the countries of origin.

Taxonomy is an accumulation of information and expertise about plants and animals. It includes the names of organisms which are governed by Codes of Nomenclature, methods of identifying them, and hypotheses of their evolutionary relationships. Evolutionary relationships are understood by analyzing shared similarities in morphology, gene sequences and other available data. Some 15 million species are believed to exist on earth (estimates vary between 5 and 100 million), but only about 1.7 million are currently known to science (Blackmore, 2002; Blaxter, 2003; May, 1999; Raven, 2004; Wilson, 2003). Knowledge of species is largely based on museum collections, which are estimated to hold between 1.3-3 billion specimens. The science of taxonomy continues to change as new techniques are discovered. Most recently, this includes "DNA barcoding" (Blaxter, 2003; Tautz et al., 2003). A DNA barcode is a short gene sequence taken from a standard portion of the genome, used to identify species quickly. DNA barcoding has the potential to be particularly important for non-taxonomists to identify species for the reasons listed above. For DNA barcoding to be truly useful, it is dependent on good taxonomy and access to all taxonomic data.

Taxonomy is unusual in scientific fields in using its accumulated published literature from the past and present. Unlike many sciences whose relevant literature may only include the past 5-15 years, taxonomists regularly use literature from some 300 years—much of which is only available in select libraries in developed countries. In order

[5] The eBCA is on the web at http://www.sil.si.edu/digitalcollections/bca/
[6] Schema, explanatory introduction and sample of marked up text from the BCA available at: http://www.sil.si.edu/digitalcollections/bca/status.cfm
[7] http://www.biodiv.org
[8] http://www.biodiv.org/programmes/cross-cutting/taxonomy/default.asp

to make it easier to use that accumulated knowledge and break down the taxonomic impediment, it is necessary to make it accessible and usable to all taxonomists, conservation biologists, decision makers and many others.

Natural history museums and similar large biological repositories and their libraries hold a wealth of inadequately accessible resources that describe and explain the diversity and depth of life on earth. Mining these data for research, conservation, drug discovery, protected area management, disease control, etc., is difficult, time consuming, and often leads to redundant efforts. What should be a seamless, open "book" of knowledge consists, instead, of disparate, unintegrated sets of data - some in electronic form but most still on paper, published and unpublished. Museum data center on the following types of biological datasets:

1. *Specimen collections*. Many biological repositories are converting manual records about their collections of biological specimens into integrated electronic collections information systems.
2. *Taxonomic databases* that record the names, classification, synonymy, geographic distributions and relationships of biological organisms.
3. *Published taxonomic literature*, including journal articles, monographs, and other forms of publication that name and describe taxa, details of collection, and other information.
4. *Geographical information systems* (GIS) that link geographic place names and other geographic data elements with precise geospatial coordinates. Once large numbers of specimens have been georeferenced, aggregate studies may be performed, such as species distribution over time.
5. *Unpublished archival materials*, including field notebooks, correspondence, and research files hold a wealth of untapped information that relates to biodiversity.
6. *Gene sequence databases.*

Unfortunately, even many taxonomists, much less other researchers and 'users' do not know how or where to find all of the relevant data and/or they cannot afford the time or money to access them. The consequence is that only a limited subset of appropriate literature and potential data are used in most analyses, limiting the adequacy of their scientific results.

The project collaborators propose a solution of leveraging existing technology to address the above issues by creating a taxonomic web space. A taxonomic web space must:

- allow access to all related data in interoperable formats;
- be accessible from anywhere in the world;
- be distributed and accessible through multiple portals;
- be flexible so that users may access the data they need in the way they expect to see it;
- be analyzable by web and other tools as they are developed;
- have ownership and intellectual property rights retained by the contributors; and
- accommodate full taxonomic treatments and single species descriptions.

To enable this, standards must be developed to allow interoperability between different data sets.

Standards are being developed by GBIF and TDWG members to support this work. Currently there are standards or standards are being developed in the following areas:

- Taxonomic Names and Taxonomic Concepts (Linnean Core[9], TCS[10])
- Specimens (Darwin Core[11], ABCD[12])
- Taxonomic Literature (metadata[13] and content[14], and see below)
- Taxonomic Descriptions (SDD[15])
- Spatial Data Standards (SDS[16])

[9] http://bdei.cs.umb.edu/twiki/bin/view/UBIF/LinneanCore
[10] http://tdwg.napier.ac.uk/index.php?pagename=HomePage
[11] http://darwincore.calacademy.org/
[12] http://www.bgbm.org/TDWG/CODATA/Schema/
[13] http://lists.tdwg.org/mailman/listinfo/tdwg-lit_lists.tdwg.org
[14] http://www.sil.si.edu/digitalcollections/bca/status.cfm
[15] http://wiki.cs.umb.edu/twiki/bin/view/SDD/WebHome

In order for interoperability to be possible, global unique identifiers of some sort are necessary. The taxonomic community is leaning toward the use of Life Science Identifiers[17].

These datasets are part of a larger, worldwide effort to enable easy access to the complete range of data required to understand individual species and their environmental and evolutionary relationships. This will require the establishment of cross-linkages between, and simultaneous access to, data sets from such information sources throughout the world. The relationships between these kinds of data are very complex as is they way that taxonomists use them in their work (Figure 1). Fortunately, most of the published data are highly structured and amenable to automation.

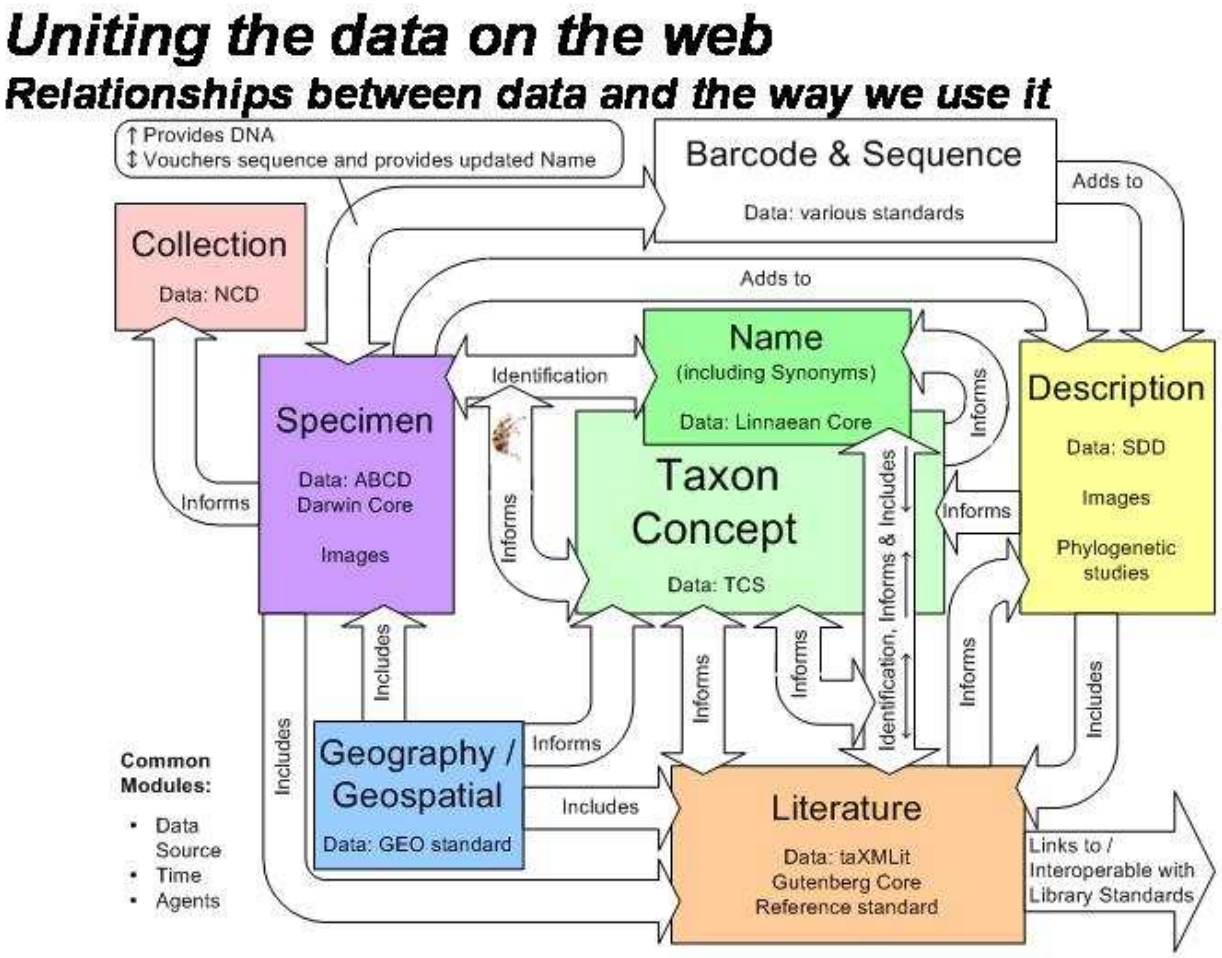

Figure 1. Relationships between taxonomic data types and they way that taxonomists use information to inform their science.

Taxonomic literature is being made available sporadically by a wide variety of museums, universities, botanic gardens, and others (e.g., American Museum of Natural History[18]; Landcare Research, New Zealand;[19] Australian

[16] http://www.tdwg.org/tdwg2000/SpatialData.htm
[17] http://www.tdwg.org/TDWG_GUID.htm
[18] http://library.amnh.org/diglib/index.html
[19] http://floraseries.landcareresearch.co.nz/pages/index.aspx

Biological Resources Study[20], New York Botanical Garden[21], Missouri Botanical Garden[22]). Most of these works are being made available are as images of pages, e.g., jpeg and pdf files. Most importantly, the Biodiversity Heritage Library[23] is a consortium of libraries that has been formed to create indexed, scanned legacy literature (as copyright allows), which will facilitate access. While this is a huge step forward in providing access to the information, it is no easier to find the needed content—essentially, unless someone knows what is in a work already, having images of pages will not help them find what they need. In addition, images of pages cannot be searched (in some cases pdfs may have been captured in such a way that unstructured, general text search may be possible) and cannot be made interoperable with other data.

The partner institutions designed the eBCA and INOTAXA as a means of demonstrating a new manner of accessing and working with these data, and addressing the Taxonomic Impediment. The INOTAXA project aims to deliver full, searchable text which can be searched and made interoperable with and link to other datasets as needed by taxonomists and others, especially those making biodiversity-related decisions. To do this, the partners have created an XML standard that can hold the complete content of all taxonomic literature, taXMLit[24].

The Electronic *Biologia Centrali-Americana* (eBCA)

The *Biologia Centrali-Americana* (Godman & Salvin, 1879-1915) is a fundamental work for the study of neotropical flora and fauna and includes nearly everything known about the biological diversity of Mexico and Central America at the time of publication. The BCA was privately issued in installments between 1879 and 1915 by F. Ducane Godman and Osbert Salvin of The Natural History Museum (London). "The work consists of 63 volumes containing 1677 plates (of which more than 900 are coloured) depicting 18,587 subjects. The total number of species described is 50,263 of which 19,263 are described for the first time."[25] The illustrations are, in many cases, the only images that exist of the biota of the region and as such could be compiled for use in an electronic field guide if available in a digital and portable format. The specimens described are deposited in many places including The Natural History Museum (London), Royal Botanic Gardens (Kew, UK), Missouri Botanical Garden, American Museum of Natural History, Harvard University, and the National Museum of Natural History (Smithsonian). Since the BCA appeared, a few select volumes have been republished but never the entire series.

The entire 63-volume BCA is believed to be held by only a few libraries world-wide. Some Central American countries lack a complete set; thus the BCA is not generally accessible to taxonomists working in the region.

In phase one of the project the electronic *Biologia Centrali-Americana* has been created. The eBCA has replicated the full BCA and made images[26] available in jpeg and pdf format[27], making the BCA available to any researcher with an Internet connection anywhere in the world. It is the largest single body of digitized taxonomic work on the World Wide Web. Researchers who formerly had to travel significant distances to use these texts can access them from their desks. Downloaded elements or parts captured on disk and used on non-connected machines are enabling workers in remote locations to study and identify animals and plants.

---

[20] http://www.deh.gov.au/biodiversity/abrs/online-resources/flora/50/index.html
[21] http://image2.nybg.org/cgi-bin/nybg.exe
[22] http://www.illustratedgarden.org/mobot/rarebooks/
[23] http://www.bhl.si.edu/
[24] http://www.sil.si.edu/digitalcollections/bca/documentation/taXMLit-v2.xsd
[25] Prospectus, Biologia Centrali-Americana, p. 4.
[26] Original images 8-bit images captured at 400 dpi on a Zeutschel Model OS 7000 AO HX-4265.01 by Preservation Resources (Bethlehem, Pennsylvania). Additional 24-bit images captured at 400 dpi on a BetterLight Super6K at the Smithsonian Institution Libraries Imaging Center. Images stored as TIFFs. High resolution images are reduced to jpeg format for the web at 96 dpi and rescaled to 1000 pixels on the short side. Images which do not contain significant color information are displayed in grayscale.
[27] http://www.sil.si.edu/digitalcollections/bca/

INOTAXA (INtegrated Open TAXonomic Access): Mesoamerican Portal

INOTAXA will mediate access to (1) published taxonomic texts (descriptions, identification keys, discussion etc), using the BCA as an initial 'backbone' but including other older and more recent treatments of Mesoamerican biota; (2) specimen data from collections in different museums and herbaria; (3) taxonomic catalogues of the animals and plants included; (4) images, both from the publications and other sources; and (5) gazetteer data. It will also provide directed links to key resources on the internet, and web tools such as mapping software. It will be a vital resource for anyone wishing to study the flora or fauna of Mesoamerica (and ultimately the world). The initial prototype, whilst meeting the needs of taxonomists for the groups covered, will also be a testbed for linkages with other data.

taXMLit

In order to accomplish the goals of INOTAXA, it is necessary to create a single format that can hold all of the data in taxonomic literature. This format must be flexible while still allowing users to search for all of the kinds of data that they expect and need to search for. We have proposed an XML schema to fill this purpose, taXMLit. At a high level, it allows for capture of publication metadata; taxonomic name content (names, synonyms, author, literature citations, type information); specimen citations; geographic content (in the form of general distribution information and geographic information within the specimen citations); character content (in the form of both descriptions and keys); and more general kinds of text in the form of discussions, various kinds of introductory and back matter, including indices, tables of contents, bibliographies and text paragraphs. It is very detailed, but it is this level of detail that will allow for:

- reconstruction of taxonomic text in various formats including species pages, keys, and checklists for specific geographic areas or specific taxa;
- reconstruction of monographs;
- automatic links to updated taxonomy (dependent on working name authority files being created via GBIF and other projects, including INOTAXA itself);
- automatic links to updated place names using a variety of gazetteers; and
- linkage to all available related biodiversity, literature, and other databases.

taXMLit is in the process of becoming a standard through TDWG and GBIF. taXMLit and explanatory documentation are is available at http://www.sil.si.edu/DigitalCollections/bca/status.cfm and is currently in an open review phase.

Prototype

The prototype INOTAXA portal will contain the fully-searchable text of two BCA volumes, one for plants and another for weevils, and marked up in taXMLit. It will also serve the text of several other taxonomic treatments of the same taxa as are covered in the volumes, to demonstrate access to multiple treatments simultaneously. Digitized data from the labels of specimens (again of species covered in the taxonomic text) held in the Smithsonian Institution's National Museum of Natural History and the Natural History Museum, London, will be provided in XML format, as will names of the taxa in modern catalogues. Images of some specimens will also be served. A gazetteer of Mesoamerican insect localities will be provided, and a simple mapping tool. External links will be provided to GBIF, *Flora Mesoamericana*[28], Harvard University botanical reference databases[29] and Google. The Taxon Concept Schema (TCS[30]) will be used. The prototype will, other than the external links mentioned, hold all data on a single server. The functions of the prototype will include:

[28] http://www.mobot.org/MOBOT/fm/
[29] http://cms.huh.harvard.edu/databases/publication_index.html
[30] http://www.soc.napier.ac.uk/tdwg/index.php

(1) Test architecture[31], methodologies and interoperability of different schemas;
(2) Provide proof of concept to underpin further grant applications;
(3) Identify user needs through a workshop, user feedback and advisory groups;
(4) Demonstrate the value of the approach to the user community; and
(5) Provide a real resource for taxonomists working on the groups covered.

The prototype will be available on the Internet in late 2006. Its functionality is discussed in a document at http://www.sil.si.edu/digitalcollections/bca/status.cfm, with indicative screens.

Mesoamerican Portal

The portal will provide access to a fully-searchable version of the BCA and other taxonomic works, with the potential to refer back to images of the original pages (including eBCA) to see the original context and format. It will contain other data types listed above for the prototype as well as glossaries, bibliographies and other resources. These data will be served not only from the same server but also from other servers worldwide in a true distributed fashion. Data (specimens, observations and names) will be available dynamically through GBIF. This linkage will enable interoperability with other GBIF-mediated data, including digitized literature such as in the Biodiversity Heritage Library (BHL[32]) project, where the metadata and parsing do not permit the full access as shown in INOTAXA. A tool-set facilitating mark-up of taxonomic texts into the taXMLit schema will be made available.

Taxonomy is an accumulation of expertise about plants and animals, nomenclature, and published literature over time. In order to make it easier to understand that accumulated knowledge, INOTAXA will allow for expert interpretation of published knowledge. For example, matching a specimen to a listing in the BCA, is not always simple, because the citation may be incomplete and the specimen not clearly labeled. An expert who has worked with the specimens and the work, may be able to provide expert knowledge about the most likely linkage between the two. Similarly, collection localities are often ambiguous. An expert may be able to gain further information from field notes or itineraries, which can be provided to other users of the system. Since both of these are not primary data, but interpretation, they will be held in a different layer within INOTAXA. By providing interpreted information and linkages, INOTAXA will provide other key information to understanding biodiversity and speeding taxonomic work.

INOTAXA is envisioned as a project to which many will contribute, and which will have a very wide 'ownership'. The partners in this process are expected to include those listed but also other organizations from around the world. INOTAXA will solicit contributions from the taxonomic and wider community (not only including taxa included in the original BCA, but expanding to all other groups, including marine taxa and to other regions). No one team will be able to provide all of the information that will serve to complete it, and indeed it may never be 'complete', in the sense that it covers 100% of all of the world's species and all of the possible information and analytical tools. However, as increasing numbers of workers use it and contribute to it, not only will it grow in content, but also more uses of it will be devised and developed. The possibility of establishing INOTAXA as an electronic journal for the publication of taxonomic information is being discussed.

With the development of tools to facilitate global access to specimen databases and taxonomic name servers, such as are being established by GBIF, dynamic linkages will be added to INOTAXA. In combination with appropriate tools, these will enable a number of additional possibilities, such as the "on the fly" preparation of check-lists of fauna and flora at all levels from local to regional and distribution maps. Such maps could be time-sliced, or be linked to climatic and ecological data enabling predictive analysis of species ranges. These products will be based on the most recent identifications in all available collections databases and, where possible, with updated information available from taxonomic name servers and/or electronic gazetteers.

[31] See http://www.sil.si.edu/digitalcollections/bca/documentation/draft-EBCA-BCAC-HLA-final.pdf
[32] http://www.bhl.si.edu/

Literature Cited

Blackmore, S. 2002. Biodiversity Update – Progress in Taxonomy. *Science* 298: 365.

Blaxter, M. 2003. Counting Angels with DNA. *Nature* 421: 122-124.

Godman, F.D. & O. Salvin, eds. 1879-1915. *Biologia Centrali-Americana*. 63 volumes. The Natural History Museum , London.

May, R., 1999. The dimensions of life on Earth. Pp. 30-45 in P.H. Raven & T. Williams (eds.) Nature and human society: the quest for a sustainable world. *Proceedings of the 1997 forum on biodiversity: National Academy Press, Washington, D.C.*

Raven, P. H. 2004. Taxonomy: Where are we now? *Phil. Trans. R. Soc. London* B. DOI 10.1098/rstb.2004.1462.

Tautz, D., P. Arctander, A. Minelli, R.H. Thomas & A.P. Vogler. 2003. A plea for DNA taxonomy. *Trends in Ecology and Evolution*, 18, 70-74 .

Wilson, E. O. 2003. The Encyclopedia of Life. *Trends in Ecology and Evolution*, 18, 77-80.

Acknowledgements

We would like to thank the following for their contributions to making this paper possible:

- Courtney Shaw for bringing INOTAXA to the Special Libraries Association and the opportunity to include this paper.
- Tom Garnett for driving the initial eBCA project and spearheading the funding effort.
- Tom Garnett, Martin Kalfatovic, and Suzanne Pilsk for the work and contribution to the work, especially on eBCA, but also constant contributions and support to the project.
- Scott Miller, Leonard Hirsch, Kristian Fauchald, and Sandy Knapp for their pivotal roles in the idea for and foundation of the project and for their continued input and support.
- Members of the BCA and INOTAXA advisory committees and participating organizations for their support and review of documents and progress.