Proceedings of TDWG: Abstracts of the 2006 Annual Conference of Biodiversity Information Standards (TDWG)

15-22 October 2006

Missouri Botanical Garden

St. Louis, Missouri, U.S.A.

Edited by Lee Belbin, Adrian Rissoné and Anna Weitzman

Published by the Missouri Botanical Garden and the Taxonomic Databases Working Group

TDWG 2006 sponsored by:

To be cited as:

Belbin, L., Rissoné, A. and Weitzman, A. (eds.). Proceedings of TDWG (2006), St Louis, MI.

This book contains abstracts of the selected papers, posters and computer demonstrations presented at the Annual Conference of the Taxonomic Databases Working Group held 15-22 October 2006 at the Missouri Botanical Garden in St. Louis, Missouri, U.S.A. The meeting attracted more than 160 participants from 22 countries and 94 prestigious scientific research institutions, museums and companies.

Presentations from the conference can be found at the address

http://tdwg2006.tdwg.org/programme/presentations/

The editors acknowledge with thanks the work of the peer reviewers and the contribution of Stan Blum, Alex Chapman, Donald Hobern, Charles Hussey, Rebecca Shapley and Arthur Chapman in its production.

ISBN 1-930723-56-3

Contents

Proceedings of TDWG

1. New TDWG Infrastructure

1.1. A Technical Architecture for TDWG Standards

Roger Hyam
TDWG Infrastructure Team

The TDWG Infrastructure Project was given the remit in 2005 to devise an umbrella architecture for TDWG standards. The purpose of this architecture is to:

- Provide a unified vision of the existing and proposed TDWG standards. It is important for the credibility of TDWG that its proposals are seen as part of an integrated whole;
- Suggest how TDWG standards should evolve so that they are interoperable with each other and external standards in the future and
- Maximise the effect of the limited resources of TDWG.

A meeting (TAG-1) of representative from groups currently active within TDWG was held in April 2006 in Edinburgh. This meeting produced a series of recommendations for the TDWG architecture in a report (http://wiki.tdwg.org/twiki/pub/TAG/TagMeeting1Report/TAG-1_Report_Final.pdf) that was widely circulated and adopted by the TDWG executive in Madrid in June 2006.

TAG-1 established two foundational principles:

- "The architecture is concerned with shared data." The TDWG architecture applies to data that are shared between entities. Only when data crosses boundaries does the format matter. The architecture should not dictate internal structures for Data Providers or Data Consumers. A successful architecture should enable the interoperability of providers and consumers with radically different internal implementations.
- "Biodiversity data will be modelled as a graph of identifiable objects." Exchange of literals (strings and numbers) in unlabelled packages is of no value. The number '55.7' has no meaning to a data consumer on its own. If it is combined with other literals in a labelled package then it is useful. For example:

  18439279
  -2.7
  55.7
  Lauder

But what is a SamplingStation? There is no written description here. What is the datum used for the longitude and latitude? This is an instance of an object of type SamplingStation but if we are to wrap literals in objects, we need a type catalogue or ontology where information about the semantics of objects can be stored and retrieved – both by humans and increasingly, by machines. Much of this modeling has already been done but may need to be presented in another form.

If we want to refer to this particular instance of a SamplingStation we could use the contents of its id field but the scope of the id cannot be known without further information. Is it unique to all sampling stations or just those from one particular data provider or perhaps all objects in the entire network? We need a system of Globally Unique Identifiers (GUIDs) if we are to refer to instances of objects across all Data Providers.

How do we find out more about this SamplingStation? If the id was resolvable then we could

use it to get a response. Alternatively we could run a query against one or more Data Providers. To do this we need well-defined data exchange protocols that our client software can exploit.

Modeling biodiversity data as a graph of identifiable objects implies an architecture that stands on 3 legs:

- A type catalogue or ontology based on current models;
- A system of Globally Unique Identifiers and
- Well-defined data exchange protocols.

A three legged stool is a useful metaphor because the legs are all equally important: remove one and the architecture fails. There are multiple dependencies between the legs. This model is the focus of the presentation.

Support is acknowledged from: Gordon and Betty Moore Foundation, GBIF, TDWG

1.2. Globally Unique Identifiers (GUID) for Biodiversity Informatics

Ricardo Scachetti Pereira[1], Donald Hobern[2], Roger Hyam[1], Lee Belbin[3], Stanley Blum[4]
[1] TDWG Infrastructure Team, [2] Global Biodiversity Information Facility (GBIF),
[3] Blatant Fabrications Pty Ltd, [4] California Academy of Sciences

In spite of the increased availability of biodiversity data on the Internet, scientists and managers still spend significant resources acquiring, integrating, and processing data to achieve useful outcomes. Inefficiencies arise because data in diverse formats cannot be easily combined into a single homogeneous dataset.

The Taxonomic Databases Working Group has been developing a common architecture to increase system interoperability and to improve data integration. The main components of this architecture are a semantics description framework and a system of Globally Unique Identifiers (GUIDs). The former describes the meaning attached to data items and allows agents to transform datasets from one representation to another. The GUID system is used to name data items shared on the network.

The TDWG-GUID group first met in February 2006, in Durham, NC, USA to discuss the requirements for a GUID system in biodiversity informatics, and evaluate GUID technologies. The group recommended the adoption of Life Sciences Identifiers (LSIDs) for naming objects in biodiversity informatics as appropriate, created a plan to address outstanding technology issues, and organized subgroups to develop prototypes to test LSID software.

The group met a second time in June 2006, in Edinburgh, UK, to report on the activities performed since the first meeting and finalise an implementation plan. The group confirmed the adoption of the LSID standard for biodiversity informatics. Since the second meeting, the group has been developing an implementation plan, documentation and LSID resources.

Our presentation will report on the status of LSID work in TDWG and justify the selection of technologies.

Support is acknowledged from: The Gordon and Betty Moore Foundation, Global Biodiversity Information Facility (GBIF), The National Evolutionary Synthesis Center (NESCent)

1.3. A Documentation Strategy for TDWG

Roger Hyam
TDWG Infrastructure Team

Documentation is the recording of information to define and support standards in a permanent format.

Study of the Internet Engineering Task Force (IETF), the World Wide Web Consortium (W3C), the Institute of Electrical and Electronics Engineers (IEEE), the Open Geospatial Consortium (OGC) and the Object Management Group (OMG) indicates the following is best practice in relation to documentation:

1. The organisation uses documents as primary outputs.
2. The organisation has clearly specified its documentation process.
3. The specification of documentation is included within the standards process itself to allow for controlled evolution.
4. Clear documentation templates and style guidelines are provided.
5. Clear IP and copyright policies are used.

It would be to the benefit of TDWG to have a documentation strategy that supports best practice. Such a strategy has been developed. It is based on three kinds of document.

Type 1 Documents are the normative parts of a standard. Examples of Type 1 documents are XML Schemas, human readable specifications that must be followed for compliance and controlled vocabularies. Type 1 documents are controlled by the TDWG standards process and are highly stable once ratified.

Type 2 Documents are parts of the standard that are non-normative (informative). Examples of Type 2 documents include examples, code and illustrations that accompany and clarify the standard. As parts of a standard they are also controlled by the TDWG standards process and are highly stable but not normative. The normative documents have precedence over them.

Type 3 Documents are those that fall outside the standard. Examples of Type 3 documents are tutorials, guides, primers, Wikis and discussion forums.

A “TDWG Standards Documentation Specification” has been developed and is proposed as a new TDWG standard to govern TDWG standards (Type 1 and Type 2 documents) going forward. It stipulates that standards take the form of a logical folder or directory that may contain any number of files and may be distributed as a zip or tar archive file. Human readable parts of standards should follow a specified layout and best practice style guidelines. They should be in XHTML format. Normative documents must be in English.

At a minimum, each standard must contain:

The normative (prescriptive) form of the standard itself (e.g., XML Schema or human readable text);

A 'Cover Page' document that summarizes the content of the standard but should also contain information on the 'Motivation' for the existence of the standard and the 'Rationale' for why the standard takes the form it does.

Documents should contain or link to copyright and other legal statements as provided in the “TDWG Standards Documentation Specification”. Copyright notices are required because:

- The copyright gives TDWG the right to publish the whole document as-is in perpetuity;

- The copyright allows others to republish the whole document as-is without obtaining permission (e.g., a document repository or mirror site);
- The copyright permits translation of the whole document into other languages;
- The copyright permits the development of derivative works within the TDWG process and
- All other rights are retained by the authors.

The author proposes that a firm base for TDWG standards efforts can be built by differentiating between the three core document types and adoption of the "TDWG Standards Documentation Specification".

Support is acknowledged from: Gordon and Betty Moore Foundation, GBIF, TDWG

1.4. New Website and Online Collaboration Infrastructure for TDWG

Ricardo Scachetti Pereira[1], Lee Belbin[1], Roger Hyam[1], Stan Blum[2], Donald Hobern[3]
[1] TDWG Infrastructure Team, [2] California Academy of Sciences, San Francisco, CA, USA,
[3] Global Biodiversity Information Facility

The Gordon and Betty Moore Foundation (GBMF: http://www.moore.org) has awarded a grant for the Taxonomic Databases Working Group (TDWG) to strengthen its standards development and management processes. The TDWG Infrastructure Project (TIP) has been conducted by the authors of this article in collaboration with the TDWG Executive Committee, TDWG subgroup conveners and TDWG members.

Our assessment is that for TDWG to efficiently provide high quality biodiversity information standards, it needs:

- To improve all aspects of communication in the organization, including the communication within and between subgroups, between TDWG and other standards development bodies, and with its main clients;
- To establish a simple, structured and well documented standards development process supported by state-of-the-art online collaboration tools.

To address these requirements, we have developed a new website and have set up several online collaboration tools, which together form the TDWG Online Environment. Its main components are: the new website, a Wiki, a Blog, the Proceedings of TDWG and the Standards Track System.

The new TDWG Website (http://www.tdwg.org) is the main communication channel between TDWG, its members and its target audience. It effectively publishes information about TDWG standards and the groups involved in their development. The site also keeps members and public informed about relevant activities.

The TDWG Wiki (http://wiki.tdwg.org) is a flexible web-based authoring system used by subgroup members to collaboratively develop drafts of standards, documentation and policy. The TDWG Blog (http://www.tdwg.org/blog) is a web logging application that allows members to informally share ideas in the form online articles. The Proceedings of TDWG (http://www.tdwg.org/ojs) is an online journal powered by the Open Journal System (OJS). Currently, the Proceedings comprise the abstracts approved for the TDWG Annual Meeting.

The TDWG Standards Track System automates most of the process associated with the development and approval of TDWG standards. The System helps subgroup conveners to prepare, submit, track, and receive feedback on their standards. The System also assists the TDWG Executive Committee in reviewing, approving, and publishing its standards.

The TDWG Online Environment also includes e-mail based mailing lists, the Subversion version control system, a Schema Repository (http://rs.tdwg.org) and a Web Archive (http://archive.tdwg.org).

We will provide an overview of the TDWG Online Environment, instructions on how to use the system, and references to additional information.

Support is acknowledged from: The Gordon and Betty Moore Foundation, The Global Biodiversity Information Facility, The Natural History Museum

1.5. TDWG Ongoing Support

Lee Belbin
TDWG Infrastructure Team

One of the five main activities of the TDWG Infrastructure Project (TIP) was to devise a strategy that would provide sustainable support for TDWG. The bulk of the work was to be undertaken within the last third of the project. One of the first problems to be recognized was the inappropriateness of the name 'Taxonomic Databases Working Group'. While ‘TDWG’ did reflect the original nature of the group, it no longer communicated effectively the significance of TDWG’s activities. The group had moved from taxonomy and databases to biodiversity and interoperability standards. ‘TDWG’ made it more difficult for members to gain support of their institutions for TDWG-related activities. The name ‘TDWG’ would also make it difficult to recruit a broader institutional membership.

The TDWG survey (http://wiki.tdwg.org/twiki/bin/view/TIP/TipSurveyResults) and subsequent interviews also helped to identify issues related to TDWG support. The taxonomy-related and IT-related membership is a strength and a potential problem for TDWG. 65% of respondents identified ‘biointeroperability standards’ as TDWG core business. 46% suggested that TDWG needed a more professional approach to standards development and more effective communication. The latter aligned with the greatest weakness seen by respondents: lack of promotion. Respondents identified data aggregators and bioinformaticists as the most important recruitment targets. Quality documentation was identified as the greatest need for respondent’s organisations. The (poor) TDWG name was identified as the most significant ‘other’ issue.

Given this information, a likely strategy that would ensure better ongoing support would include the following- a) improving the quality of TDWG’s standards, b) effectively communicating TDWG activities to a broader audience, c) increasing institutional membership and d) identifying a more appropriate name for TDWG.

Support is acknowledged from: The Gordon and Betty Moore Foundation

2. New and Emerging Standards

2.1. A Web Services API for Fundamental Niche Modeling

Tim Sutton, Renato De Giovanni
CRIA, Campinas, SP, Brazil

The openModeller project aims to provide a flexible, user friendly, cross-platform environment where the entire process of conducting a fundamental niche modeling experiment can be carried out. The software includes facilities for reading species occurrence and environmental data, selection of environmental layers on which the model should be based, creating a fundamental niche model and projecting the model into an environmental scenario. A number of fundamental niche modeling algorithms are provided as plug-ins, including GARP, Climate Space Model and Bioclimatic Envelopes. Additional algorithms are planned for the future. The

submission of alternative algorithms is always welcome.

The basis of openModeller is a software library that provides all of the processing logic associated with niche modeling. Programmatic, command line and graphical user interfaces provide access to the functionality available in the library.
The openModeller graphical user interface (OMGUI) is written in Qt4/C++. It is designed to be re-usable by making use of a component-based architecture. This re-usability allows the software to be, for example, embedded into other applications such as Quantum GIS and in the future, TerraView. We are extending the OMGUI to include the facility for conducting experiments for multiple species, multiple modeling algorithms and multiple environmental layer sets into which the models should be projected. We plan to offer tools in OMGUI for pre-modeling activities for example searching for species occurrence data on speciesLink and GBIF and importing esoteric environmental data formats. We also plan to offer tools in OMGUI for post-processing the model outputs such as computing probability 'hotspots', comparing model outputs and presenting a detailed report of the experiment once it has been completed.

In its simplest form, the OMGUI software performs all of its computation and data management locally on the user's workstation. Through the use of 'modeler adapter plug-ins' we are also enabling the OMGUI to act as a controller for carrying out experiments distributed across one or more remote systems, for example using Web Services and Condor.

We will describe the openModeller project and propose a standard Web Services Application Programming Interface (API) for interoperation with other environmental niche modeling applications. This Web Services API defines a minimal set of operations that need to be implemented in order to provide remote invocation capability for a modeling application. This API is designed to promote interoperability between different modeling applications, even if they have been developed in different programming languages and with different application architectures.

The openModeller project is an Open Source project (published under the GNU General Public License), currently being funded by FAPESP. For more information please visit http://openmodeller.sf.net.

Support is acknowledged from: We acknowledge support from FAPESP

2.2. The EFG extension to the ABCD schema

Wolfgang Kiessling[1], Charles Copp[2], Adrian Rissoné[3], Markus Döring[4], Heike Mewis[1]
[1] Museum für Naturkunde der Humboldt-Universität zu Berlin, [2] Environmental Information Management, [3] Natural History Museum, London, [4] Botanic Garden and Botanical Museum Berlin-Dahlem

Large museum and university collections are as widespread in the geosciences as they are in the biological sciences. Even more than in the biological sciences, the geosciences are faced with a large array of different objects, which are described in a heterogeneous fashion. We propose that building on the existing ABCD schema will be a fast and at the same time the most comprehensive way to mobilize information in the earth science disciplines. This is done with due regard to several individual initiatives attempting to database and mobilize geoscience data.

In the framework of the European SYNTHESYS project, an international workshop was organised in Berlin (July 2005) to develop comprehensive data models for the earth sciences building on existing models in biology (GeoCASE – http://projects.naturkundemuseum-berlin.de/synthesys_activity_d/). The Extension For Geosciences (EFG) was translated into an XML schema by Charles Copp and integrated into the ABCD schema by Charles Copp and Markus Döring. The resulting schema is compatible in principle with the geological extension proposed for DarwinCore2 (DarwinCoPE (DarwinCore Paleontology Extension) - http://darwincope.museum.state.il.us/) but significantly broader in scope. We anticipate the two

schemas being able to operate side by side through the unified TAPIR protocol (http://ww3.bgbm.org/protocolwiki/).

The draft ABCDEFG schema is now being utilized to map the large palaeontological, geological and mineralogical collections of the Berlin Museum für Naturkunde. This demonstrates the functionality of the new schema to a broad audience. We anticipate the schema will encourage other large museums to adopt it as a standard. The databases are expected to go online by October 2006. A second workshop to be held in Berlin in January 2007 will be used to fine-tune the schema and to instruct colleagues from other European museums who are willing to use the schema for the databases at their home institutions.

Support is acknowledged from: SYNTHESYS

2.3. TAPIR 1.0

Renato De Giovanni[1], Markus Döring[2], Javier de la Torre[3]
[1] CRIA, Brazil, [2] Botanic Garden and Botanical Museum Berlin-Dahlem,
[3] Museo Nacional de Ciencias Naturales, Madrid

The TDWG Access Protocol for Information Retrieval (TAPIR) is a next generation of query protocols that can be used by biodiversity information networks. It was initially proposed as a new protocol unifying DiGIR and BioCASE during the TDWG 2004 meeting. After that, many changes were incorporated to add new functionalities and to allow different levels of provider implementations. A fully functional provider software (PyWrapper) has been developed and is ready to be used.

The TAPIR protocol consists of five operations – Metadata, Capabilities, Ping, Inventory and Search – that can be invoked either through XML or simple KVP (key-value pairs) requests. The Metadata response has been refactored recently to make use of elements from well-known namespaces like DublinCore, VCARD and the W3C Basic Geo vocabulary, and also to include additional data such as any number of related entities, multi-language support, and indexing preferences, among others. Capabilities is a separate operation to retrieve technical metadata, allowing providers to have different levels of functionality. Ping can be used to monitor providers. Inventory operations now accept more than one concept. Both Inventory and Search can now make use of new filtering capabilities and can be represented by query templates.

This session will introduce TAPIR, explaining the basic concepts behind it, including output models, query templates and the different ways of processing them. The final TAPIR 1.0 specification will be presented together with new perspectives and future directions.

TAPIR Wiki: http://ww3.bgbm.org/protocolwiki/
PyWrapper home page: http://www.pywrapper.org/

Support is acknowledged from: TDWG Infrastructure Project, Gordon and Betty Moore Foundation, GBIF

3. Integrating Standards

3.1. Exchange of Germplasm Datasets with PyWrapper/BioCASE

Dag T. F. Endresen[1], Johan Bäckman[1], Helmut Knüpffer[2], Samy Gaiji[3]
[1] Nordic Gene Bank (NGB), [2] Institute of Plant Genetics and Crop Plant Research (IPK Gatersleben),
[3] International Plant Genetic Resources Institute (now 'Bioversity International')

There are more than six million ex situ germplasm accessions of agricultural and horticultural crops conserved worldwide by genebanks (seed banks), according to the FAO. These germplasm collections share most of their attributes, but database systems and data models implemented may differ substantially. The International Plant Genetic Resources Institute, IPGRI, has developed standards for data exchange and data integration, which are implemented by many genebanks. Germplasm collections also share many attributes with other biodiversity collections, such as natural history museums, botanical gardens or herbaria. Today there is no single point of access allowing the discovery of germplasm samples across all genebank collections worldwide. Germplasm data portals like EURISCO (European genebanks), SINGER (CGIAR genebanks), USDA-GRIN (USA) and NGB (Northern Europe) successfully demonstrate that distributed data on germplasm accessions (genebank seed samples) can be mapped to common standards and thus accessed from global and regional data portals. These regional portals have so far been implemented as classical data warehouses. The attributes of the source datasets have been transformed to the agreed data exchange standard and included in a central database or index.

GBIF supports data flow from simple web services implemented by DiGIR or BioCASE/PyWrapper data provider software installed locally at each data source node. Such wrappers can be implemented for different database systems and do not require modification of the local database structure. Any update of contents of the local database will immediately be visible for search portals.

A number of genebanks have already joined GBIF as data providers. This process was initiated by IPK Gatersleben, Germany. The first genebank to provide its accession data records to GBIF was the Nordic Gene Bank (North Europe) in March 2004. IHAR (Poland) and IPK Gatersleben (Germany) followed soon after. Later also USDA-GRIN (USA, 2005) and WUR, CGN Wageningen (The Netherlands, 2006) became GBIF data providers. The CGIAR genebanks through SINGER and EURISCO representing most European genebanks have also joined GBIF (2006) and will soon provide data records to the GBIF index. The GBIF data portal provides a new and valuable channel to promote germplasm datasets. The adoption of the PyWrapper software has proven relatively simple, and a genebank providing data to GBIF will, with little additional effort, be able to provide the same dataset to EURISCO or SINGER, using the same data standards. Exchange of germplasm data with PyWrapper has successfully been tested with the ABCD, Darwin Core, GCP Passport, and MCPD data standards. Work is in progress to further implement PyWrapper as the preferred data exchange tool for the genebanks providing data to EURISCO and SINGER. Development of the new TAPIR protocol will soon provide many important and promising improvements to the data harvesting and indexing routines of germplasm data portals. Improved data harvesting routines and a time-to-live attribute for datasets and individual records are also under development. The TDWG standards on GUIDs will also play an important role.

Some germplasm data portals: EURISCO (http://eurisco.ecpgr.org/); SINGER (http://singer.grinfo.net/), USDA-GRIN (http://www.ars-grin.gov/); SESTO (http://www.ngb.se/sesto/); FAO (http://apps3.fao.org/wiews/).

Support is acknowledged from: Nordic Gene Bank (NGB), International Plant Genetic Resources Institute (IPGRI), IPK Gatersleben

3.2. PyWrapper v2: Toward a Real Open Source Community

Javier de la Torre[1], Markus Döring[2]
[1] Museo Nacional de Ciencias Naturales - CSIC, [2] BGBM - FU Berlin

PyWrapper v2 is a major revision of the previous BioCASe Provider Software. It has been redeveloped to become the first TAPIR implementation. During the last year several projects have contributed to its development and extension. At the same time the project has moved into a new development environment, outside of any institution, to promote its development by an open source community.

PyWrapper has evolved into a multiprotocol middleware software. Projects are demanding that their data providers operate with different protocols, some not TDWG related. The Python based software has been modularized and support for the BioMOBY protocol has been implemented.

Plans include providing support for LSID resolution and WFS. The goal is to provide a single interface for providers to map their databases once and share their data using multiple protocols.

It is also envisioned that complementary tools will be bundled with PyWrapper. The first tool will be the QueryTool; a generic client to create web interfaces for providers databases based on AJAX technology. We hope the number of additional modules will grow as different communities contribute to the open source project.

http://www.pywrapper.org

Support is acknowledged from: IPGRI, Synthesys, GBIF, BioCASE, ENBI

3.3. Biodiversity Informatics and the GeoWeb: Toward an Integration of TDWG and OGC Standards

Javier de la Torre[1], Patricia Mergen[2], Jorge M. Lobo[1]
[1] Museo Nacional de Ciencias Naturales, Madrid, [2] Royal Museum for Central Africa, Belgium

The geospatial aspect of biodiversity data is very prominent for research in ecology, biogeography, as well as for planning, conservation and management. Most use cases for biodiversity primary data involves the geospatial analysis of data using GIS tools. Therefore, facilitating the access of GIS users to primary data is an important task in fulfilling many user requirements for biodiversity information networks.

The best way to meet users demands is through the use of open standards like the ones being promoted by the Open Geospatial Consortium (OGC, http://www.opengeospatial.org/). OGC has been working in open standards for more than a decade and has created several widely deployed specifications, like WMS/WFS/WCS and GML. These efforts are creating an interoperable environment where "geodata" are consumed, analyzed, integrated and published in what is starting to be called the GeoWeb.

OGC and TDWG standards together can provide the building blocks for a "BiogeoWeb", where biodiversity data can be visualized and analyzed together with other "geodata" sources thanks to interface and semantic interoperability. This process has already been initiated by the TDWG Spatial Data Standards subgroup and it will gain force with the creation of an agreement between OGC and TDWG. The inclusion of TDWG standards in the OGC world will also guarantee further integration of our community in spatial initiatives, like GEOSS (http://www.earthobservations.org) or INSPIRE (http://www.ec-gis.org/inspire/), that have biodiversity data within their scope.

A description of how the different existing standards can be used in biodiversity informatics, together with practical results from the setup of SYNTHESYS (http://www.biocase.org/products/geo_services/core_gis/) project services will be presented in the context of a future Biogeography Spatial Data Infrastructure: “BiogeoSDI”.

Support is acknowledged from: SYNTHESYS

3.4. An Integrative, Standards-Compliant Framework for TDWG Schemata and Services

Phillip C. Dibner
Ecosystem Associates

The ISO 19100 series of geographic information standards provide language and a set of concepts for describing abstractions of real-world entities, or Features. ISO Features as information constructs provide great generality for characterizing phenomena, while retaining a consistent, normalized underlying concept model that facilitates integration and analysis. Many objects of interest to systematists, ecologists, and field biologists can be modeled usefully as Features. In this presentation, we illustrate the application of the Feature Model and the Observation and Measurements framework (OGC Document 05-087r3, Simon Cox, 2005, http://portal.opengeospatial.org/files/?artifact_id=14034) to biological collections data and to field observations. We demonstrate how vocabularies defined by the ABCD, Darwin Core, and TCS schemata, and some emerging work from the TDWG Geospatial Interest Group, fit naturally and compatibly into this structure.

Support is acknowledged from: The Global Biodiversity Information Facility (GBIF), the Open Geospatial Consortium (OGC), The U.S. National Aeronautics and Space Agency (NASA)

3.5. TDWG and the OGC: An Update

Phillip C. Dibner
Ecosystem Associates

The author reports briefly on the status of the relationship between TDWG and the Open Geospatial Consortium (OGC), an international industry consortium of more than 300 universities, public agencies, and companies that uses a consensus process to develop publicly-available standards for interoperable geospatial web services. During the past few years, TDWG representatives have attended several OGC Technical Committee (TC) meetings.

In June, 2006, representatives of the TDWG Geospatial Interest Group (GIG) and the TDWG Technical Architecture Group (TAG) attended an OGC TC meeting in Edinburgh, Scotland. It was agreed that TDWG and the OGC would explore ways of sharing relevant standards documentation, jointly develop standards-based profiles and schemas to support consistent representation of taxonomic objects, jointly consider outreach opportunities, and potentially identify and promote joint testbed and pilot deployment activities. This agreement is currently being incorporated into a Memorandum of Understanding that will document and help guide these shared activities.

Since the Edinburgh meeting, TDWG members have already facilitated the attendance of OGC staff at a workshop on taxonomic databases for paleontology. The relationship between the two organizations will also figure in the emerging OGC Interoperability Institute (OGCII), an affiliate of the industry consortium dedicated to the promotion and development of interoperable geospatial services for scientific inquiry and basic research.

4. Ontologies and Semantics

4.1. Developing a Core Ontology for Taxonomic Data

Jessie Kennedy[1], Robert Gales[2], Robert Kukla[1], Roger Hyam[3], John R Wieczorek[4], Gregor Hagedorn[5], Markus Döring[6], Dave Vieglais[2]
[1] Napier University, [2] University of Kansas, [3] TDWG Infrastructure Team, [4] University of California, Berkeley, [5] Institute for Plant Virology, Microbiology, and Biosafety, Federal Research Center for Agriculture and Forestry, Berlin, [6] Botanic Garden and Botanical Museum, Berlin-Dahlem

Over recent years several sub-groups within the Taxonomic Databases Working Group (TDWG) have developed models and exchange standards to facilitate data sharing within the taxonomic community. These include ABCD, SDD, DwC, TCS and Spatial Data Standards. Of these, ABCD, SDD and TCS have been ratified as TDWG standards (see http://www.tdwg.org/standards). Although each group focused on different aspects of taxonomic data and its representation, the resulting standards duplicated the modeling of many aspects of taxonomy. For example, Biological Collections refer to specimens, taxonomic names or concepts, institutions, people, publications; Descriptions refer to people, specimens, publications, taxonomic names and concepts etc; and Taxonomic Concepts refer to specimens, taxonomic names, publications, people, descriptions etc. The resulting overlap across the existing standards has limited or no common terminology or model. A Core Ontology is proposed as a reference for all taxonomic domain models to facilitate more effective data sharing within the community.

Representatives from ABCD, DwC, GBIF, SDD and TCS analysed the existing data models and exchange standards and propose a draft Core Ontology for taxonomy. The result is expressed in terms of a Base Ontology, Core Ontology and Domain Ontology from which applications can be developed. The Base Ontology comprises classes representing general, non-taxonomic specific concepts which are seen as base classes from which the Core Ontology classes are derived. The Core Ontology comprises classes that correspond to the most common and therefore important concepts within the TDWG community. These classes are seen as the basis of a community vocabulary. Where possible, such classes were given textual definitions based on Oxford English dictionary to aid in the general understanding of what was intended by the class. The properties of Core Ontology classes are limited in type other classes in the Core Ontology, leaving further elaboration to the domain classes. A Domain Ontology is developed from the classes in Core Ontology. The Domain Ontology comprises sub-ontologies which have a correspondence to a single class in the Core Ontology to encourage reusability of the Domain Ontology classes. This, for example, would prevent a Specimen ontology defining Taxonomic Name. The Domain Ontology classes capture additional semantics to the core classes. This allows the community to share classes and further to compose application schemas to represent their perspective of the taxonomic domain such as museum curation, ecological surveys and nomenclature management. Relationships introduced between classes not explicit in the Core Ontology may indicate misuse or misunderstanding of the semantics of the Core Ontology. Application schemas will develop classes/data structures using the classes and the properties in the Domain Ontology.

The presentation will report on the design and development of the Core Ontology and how the Core was extended to a trial Domain Ontology. The trial Domain Ontology had an associated data repository that enabled the testing of the Core Ontology.

Support is acknowledged from: TDWG Infrastructure Project, Gordon and Betty Moore Foundation

4.2. Converting an Existing Taxonomic Data Resource to Employ an Ontology and LSIDs

Jessie Kennedy[1], Robert Gales[2], Robert Kukla[1]
[1] Napier University, [2] University of Kansas

Data sharing is fundamental to biodiversity and taxonomic data applications, however previous attempts at developing mechanisms to facilitate sharing within the community have had limited effect. Reasons for this include the lack of take up of data exchange standards (which is now slowly happening due to the TDWG standards initiative), the absence of a common terminology or vocabulary for use within taxonomic data and the lack of reference database systems for serving and referring to authoritative data. In an attempt to improve this situation, a Core Ontology for taxonomic data has been developed to model the entities widely used in taxonomy in an independent manner and allow their reuse for different taxonomic purposes. In addition Life Science Identifiers (LSIDs) have been proposed by the TDWG GUID working group as the means for uniquely identifying taxonomic data objects, such as specimens, taxonomic names, taxonomic concepts, etc. The LSIDs can make use of a Core Ontology or a Domain Ontology derived from the Core in order to define the data to be returned from resolving an LSID. These data are expressed in RDF, a language central to the semantic web.

For this approach to be effective it is essential that a mechanism exits for migrating existing data to the new technologies, e.g. LSIDs and RDF using a Core Ontology. However, using LSIDs per se will not address the issue of data sharing unless repositories reuse LSIDs to cross reference data internally and externally. It is important that taxonomists use the same LSID to refer to the same taxonomic entity rather than have multiple LSIDs identifying the same entity. If this were to happen we would need to decide if two LSIDs were really the same thing. We would be in a similar situation as we are today where we are trying to decide if two taxonomic names are really the same. Generating LSIDs for any self contained data set is trivial. It is a challenge however to allocate LSIDs to data when the LSID may be new because the data are owned by a specific repository, or to determine when an LSID should be acquired from an external database that serves as an authority for the data.

This presentation will report on the migration of the Hexacorallians of the World to a domain ontology derived from the proposed TDWG core ontology. The ontologies are represented in RDF and the data were cross-referenced using LSIDs. The focus is on the development of a tool to aid the process of converting internal database keys to LSIDs. These LSIDs may be generated automatically for data owned by the repository or appropriated from some external LSID authority. The provision of such a tool will facilitate domain scientists in publishing their data in a manner that will enable better discovery, reuse and cross referencing using LSIDs.

Support is acknowledged from: TDWG Infrastructure Project, Gordon and Betty Moore Foundation

4.3. TOM - The TDWG Ontology Metamodel

Roger Hyam
TDWG Infrastructure Team

The report of the Technical Architecture Group's first meeting identifies a need to develop Core and Base Ontologies to act as a typing mechanism for exchanged objects. Here a platform independent metamodel is proposed for the TDWG ontology. This is a specification for how the ontology should be built rather thanfor its contents.

There are many technologies available for defining ontologies. The initial candidate list includes RDFS, OWL (Lite, DL or Full), XML Schemas and OGC's Geography Mark-up Language (GML). The problem with adopting one of these technologies is that it would result in advocating the adoption of a single technology for all interactions within the TDWG

community. If the architecture were to advocate the adoption of a purely W3C Semantic Web-based approach (RDFS, OWL) then integration with OGC GML applications would not be possible. Similarly, the use of XML Schemas does not currently permit integration with Semantic Web technologies or GML-based technologies, and use of a purely GML-based approach does not permit integration with XML Schemas or Semantic Web technologies (although attempts have been made to express the GML metamodel in OWL). The semantics of the TDWG community need to be mapped into all these technologies.

The same basic semantics must persist across these technologies. A TDWG Specimen needs to be a TDWG Specimen whether it is expressed in GML, OWL or different XML Schemas. This translation process (translation of conceptual schemas) must be explicit, documented and preferably automatic.

It is clear that the constructs used to build the ontology must map as cleanly as possible into not only the existing candidate languages but also technologies that may be developed in the future. The minimum set of constructs that meet the current needs of the community and can be mapped need to be identified. This guarantees applicability in the current situation and minimises the risk of not being able to map into future technologies. The TDWG Ontology must be based on its own metamodel. This is not a radical suggestion as the metamodel will consist of the minimum set of most common constructs available presently – nothing novel – and can be changed in the future if required. The initial metamodel is proposed here and is being implemented in a simple web based application called Tonto.

This short presentation outlines the key constructs in the TDWG Ontology Metamodel:

- Classes;
- Literal properties of classes;
- Properties with ranges of instances of classes. i.e., instance relationships;
- No cardinality of properties;
- Single inheritance of classes and
- Instances of classes within the ontology that are restricted to possessing literal values.

This is an initial model that could be extended in the future should the need be clearly identified.

There are still issues that need to be clarified regarding the data types of literals, ontology governance and the rendering of classes in different technologies but the basic constructs proposed here need to be established before these can be finalised.

Support is acknowledged from: Gordon and Betty Moore Foundation, GBIF, TDWG

4.4. TDWG Data Sharing

Charlie J. Lapham
Southeast Regional Network of Expertise and Collections (SERNEC)

There are serious problems in the sharing of data for curators without IT support. The easiest way to validate ones data is to use authority files, currently TDWG provides definitions but not the related common authority files. There is likely no shortage of local authority files, perhaps not explicitly named as such, and thus data are easily validated at the local level. The lack of common authority files requires schemas to include unrestricted text fields. These fields are loopholes that permit and virtually guarantee invalid data on the portals - including spelling errors and inconsistent nomenclature. The combining of data validated to different authority files has virtually guaranteed inconsistent data on the portals. Additionally, the content of the various authority files is generally unknown to portal users, or the portal itself for that matter, so simple conversion tables are not an option

If curators have IT support these data may be validated and edited, enabling data sharing without the risk of contaminating local data. This is generally not within the scope of the majority of curators without IT support, for a lack of expertise or lack of time. It is a potentially unnecessary expense for all; it seriously complicates the data sharing process and it will continue until the issue is addressed.

If, on the other hand, common authority files were developed, the loopholes could be closed, and the data could be shared without any adjustments, a highly desirable goal. It is proposed TDWG expand its definition efforts to include and maintain common authority tables, a potentially huge job!

Techniques for shrinking the size of the task and gradually approaching the goal are suggested. Published regional authority lists would enable the use of conversion tables within the region. The authority files can be created or derived by curators who are not necessarily entirely comfortable in cyberspace, a much more numerous resource than IT-savvy curators. A new cyber-novice group within TDWG could do the bulk of this work and would also meet some of the outreach initiatives TDWG are attempting to introduce.

Support is acknowledged from: Southeast Regional Network of Expertise and Collections (SERNEC)

4.5. Ontologizing Morphological Terms for Hymenoptera (Insecta) - Implementing and Benefiting from a Controlled Vocabulary

Andrew R Deans, Gregory A Riccardi, Fredrik Ronquist
Florida State University

Hymenoptera is a large group of organisms commonly referred to as sawflies, wasps, ants and bees. This group has historically had numerous communities of researchers untangling its mysteries. Sawfly specialists, parasitoid people, ant taxonomists, bee biologists, and aculeate workers have converged on a common language that describes the morphological characteristics of these insects. Each research group has however also cultivated its own specialized terminology that may not be applicable to other hymenopterans. Complications include: (1) words that apply to structures only found in certain taxa (e.g., “cenchrus” in sawflies); (2) words that are different from terms other workers use (“parapsidal furrow” in some taxa is homologous to the “notaulus” in other taxa); (3) words that are obscure and go unused (“lunule” in microgastrine Braconidae) and (4) words that have the same spelling but are defined by the taxon (e.g., “face” in Symphyta and Aculeata is not the same as “face” in Parasitica).

Ontologies serve multiple functions in a vast array of contexts, from facilitating communication between databases to standardizing the terminology used by a particular field. The hymenopterist community seeks to ontologize the vast array of morphological terminology by: (1) carefully reviewing the literature pertaining Hymenoptera morphology; (2) systematically designating synonyms that are either valid or obsolete; (3) defining relationships between terms and (4) illustrating and verbally defining valid terms.

Ontology development and editing:

Several software packages are available for developing and editing ontologies, for example, OBO-edit (Gene Ontology Consortium) and Protégé-OWL (Stanford), which can export the ontology in several formats. The Hymenoptera community will likely employ a user-friendly, customizable database built using Ruby on Rails/MySQL (mx; http://hymenoptera.tamu.edu/mx/). This strategy allows more Hymenoptera morphologists to participate since there is no new software to learn, and it allows the ontology to be built

remotely, from anywhere with Internet access. After the ontology matures, the database will be assembled by our bioinformatics colleagues into a usable ontology.

The standards and hierarchical nature of an ontology will prove invaluable when applied to resources such as MorphBank (http://www.morphbank.net/):

- Situation 1 - When searching for “mesosoma” images, one could potentially miss images that were tagged as “scutellum”. By incorporating the ontology into the search algorithm one could return images of “mesosoma” and all related terms (e.g., terms that are “part of” the mesosoma, such as the scutellum).
- Situation 2 - Annotating uploaded images. Standard terminology is highlighted and linked and inappropriate synonyms are identified, while non-standard or misspelled terms remain unrecognized. This provides feedback about the quality of the user’s annotation.
- Situation 3 - When scoring phylogenetic characters, users might define characters and states differently. An ontology will assist in homology assignments critical to phylogenetic analyses.

Support is acknowledged from: National Science Foundation

5. New Ideas

5.1. Building Biodiversity Information Education: Next Generation Bioinformaticians

Patrick Bryan Heidorn, Carole Palmer, Dan Wright
Univerity of Illinois

All science is becoming e-science as evidenced by the existence of groups such as TDWG. New scientists being trained in universities and established scientists in their labs must learn to use the information processing tools of the period in order to conduct their work and to publish their results in an efficient manner. Biologists cannot always turn to information or computer scientists to solve the informatics problems for them since few computer professionals are trained in the unique needs, tools and standards of the biological community. The problem is even more acute when it comes to the development of biological information processing tools and standards. TDWG should play a key role in the education of the next generation of bioinformaticians as biologists or information professionals from non-biology fields. People interested in learning about biodiversity informatics should be able to turn to TDWG to identify the resources required.

To meet this objective we would need to take the following steps. 1) Spur interest in education and outreach among TDWG members or potential members. 2) Maintain a list of the knowledge and skills required in biodiversity informatics. 3) Identify individuals with the knowledge and skills. 4) Define educational units which might include key documents, a bibliography, and optional venues for face-to-face and Internet classes. 5) Maintain pointers to relevant educational units outside of TDWG. 6) Create educational materials where they do not exist. 7) Identify the dependencies among required skills and knowledge so that learners can plot a meaningful path through the educational units.

As an example of this process, we will discuss the approach taken by the authors in an NSF-sponsored project http://sci.lis.uiuc.edu/ to develop a Masters of Science Degree program in Biological Informatics at the University of Illinois http://www.lis.uiuc.edu/programs/ms-bioinformatics.html as well as an MS course in biodiversity informatics https://hive.lis.uiuc.edu/display/FA06LIS590EI/Home.

TDWG should expand its education efforts to include a Wiki for education, and by adding support for live and recorded Internet-based training. We could use the Wiki to develop and publish a knowledge and skills list along with educational outlines for educational units. For example, DiGIR is an important technology for our community that is already well organized for education through the GBIF DiGIR Provider classes. The educational index at TDWG could reference the DiGIR portal, the GBIF portal and associated materials. We would also list dependencies of DiGIR on Darwin Core and ABCD, as well as basic database and operating system knowledge. With live and recorded Internet-based training, scientists and informaticians with knowledge of biological informatics will be able to share this knowledge with students and scientists wanting to develop their skills.

Support is acknowledged from: National Science Foundation

6. Observations

6.1. Development of a Provisional Observation Data Standard Capable of Supporting both Species-Based and Ecological Inventory and Monitoring Protocols

Lynn S Kutner, Bruce A Stein, Donna J Reynolds
NatureServe

Observational data on location, condition, and other attributes of species and ecological units is critical for many biological research endeavors, as well as for conservation planning and natural resource management and monitoring. Observational data are however heterogeneous, and this complicates efforts to aggregate and analyze data produced by different inventory and monitoring protocols.

NatureServe coordinated a multi-institutional process to develop a provisional standard for observation data that would apply to a range of data and surveys. This was done to foster interoperability, collaboration, and data sharing among observation-oriented data initiatives. There was a need to accommodate core data elements from species-based inventories, monitoring protocols and ecological or habitat-based protocols. Major entities within the standard are: observation (including identification, location, date, observer, observation methods and evidence, negative observations, and monitoring); survey; search area; species list (e.g., for vegetation plots); documentation; protocol; and project.

This presentation provides an overview of the provisional observation standard. NatureServe is currently using the standard as the basis for Kestrel, a web-enabled prototype observational data management software application.

Version 1.0 of the provisional standard is available at:
http://www.natureserve.org/prodServices/obsStandard.jsp.

Support is acknowledged from: Gordon & Betty Moore Foundation

6.2. Issues of Data Quality in Observational Datasets

Steve Kelling
Cornell Lab of Ornithology

As observational data begin to play a larger role in biodiversity informatics, a general overview is necessary to address issues of data quality. These quality issues fall into a variety of categories, some of which are similar to those of museum specimen data (e.g. taxonomic, spatial, data storage), and some unique to observations (i.e. user and project biases, data gathering and management). While several excellent reviews have been undertaken to define the uses of species occurrence data (Principles and Methods of Data Cleaning, Chapman 2005), or begin to define observational data (Observations on Observational Data, Vieno and Saaksjarvi 2003), none have specifically addressed the measures taken by organizations that gather observational data to ensure data quality. This presentation will identify those measures that ensure data quality in existing and new observational data projects.

Support is acknowledged from: NSF-IIS 0612031, NSF-DBI 0542868, NSF-EF 0409378, GBIF

6.3. The Role of Negative Observation Data in Biodiversity Studies

Kevin Webb, Steve Kelling
Cornell Lab of Ornithology

The usefulness of observation data in biologic, conservation, and environmental sciences is greatly enhanced when supplemented with both the locations where observations of an organism were made and locations where observers searched for the organism but did not encounter it. These latter observations are often called negative observational data.

Visualizations and analyses of combined positive and negative observational data provide a challenge to application developers and data managers. We present an example where we analyzed 800,000 submissions of bird observations. Each submission averaged 10 species reported, resulting in a database of approximately 8,000,000 records. Negative observations are not stored explicitly in the database, but can be inferred from the positive records. We can have a few hundred thousand records for house finch (a common feeder bird in North America), but none for ivory-billed woodpecker (a not-so-common bird anywhere) in the database.

For some analyses, negative observations must be explicitly generated. To build a model for house finch occurrence, we have 800,000 data points. If we want to build a model for ivory-billed woodpecker for North America, we also have 800,000 data points to consider. The only difference is the number and distribution of the positive observations. Thus having negative observations makes a significant difference for building accurate models of species occurrence. Our discussion is focused on logistical considerations and challenges of collection, storage, and presentation of negative observation data.

Support is acknowledged from: NSF-IIS 0612031, NSF-DBI 0542868, NSF-EF 0409378

7. Imaging

7.1. MorphBank's Approach to Determination Annotations of Specimen Images, Including the Results of User Trials

Austin Mast, David Gaitros, Fredrik Ronquist, Peter Jörgensen,
Corinne Jörgensen, Greg Riccardi
Florida State University, Tallahassee

MorphBank (http://www.morphbank.net) is an open web repository for biological images with functionality tailored to disciplines using the resource. New functionality introduced in MorphBank version 2.5 provides a suite of tools for users of specimen images and the biological research collections (BRCs) that contributed them. Users can now create and manage image collections of specimens from multiple BRCs, summarize prior taxonomic determinations of 1-many specimen(s) in tabular form, and measure specimen features in their browsers. Digital determination annotations can be made to single images or groups of images simultaneously. Fifteen taxonomists were asked to use MorphBank images of specimens from the Robert K. Godfrey Herbarium at Florida State University to make digital determination annotations for 50 specimens each. We will present our assessment of their experiences with the system. We will also discuss ways that MorphBank's digital determination annotations can be used to map taxonomic concepts (in a general sense) in the future.

7.2. The Use of Specimen Label Images for Efficient Data Acquisition in Research Collections Cataloguing

Inyigo Granzow de la Cerda[1], Juan Carlos Gómez-Martínez[2], José Luis García-Castillo[2]
[1] University of Michigan Herbarium, [2] SEI, México, D.F.

Digital images of herbarium specimen labels are the core tool for populating catalog data at large scale. A workflow for cataloguing the University of Michigan Herbarium's (MICH) large collection of Mexican land plants was designed to maximize throughput and minimize data acquisition costs. This modular workflow consists of three independent phases that are carried out at MICH and in Mexico, optimizing resources as well as the efficiency and skills of personnel in each location. Specimen images and essential reference data were captured on a reference pre-catalog at MICH. Then both digital images and the reference pre-catalog were sent to collaborators in Mexico where most effort-intensive tasks of geographical data entry and georeferencing of localities, the latter requiring highly skilled personnel, were performed. The complete database is then verified by MICH personnel and made available to the public online. This workflow, when remote data entry onto a single server is feasible, allows for high efficiency databasing at a relatively low cost. The project carried out by MICH has imaged and pre-cataloged ca. 83,000 Mexican and Mesoamerican land plant specimen labels, of which ca. 43,000 Mexican specimens have been fully georeferenced. This project is support by the National Science Foundation BRC Program (Grant #0138621).

Support is acknowledged from: National Science Foundation BRC Program (Grant #0138621)

7.3. Representing and Using Phylogenetic Characters in MorphBank

Greg Riccardi, David Gaitros, Austin Mast, Fredrik Ronquist
Florida State University

The MorphBank project is a repository of biological images and related information that includes extensive user tools in a web site, at http://morphbank.net. Maintaining detailed metadata about the image and its associated specimen increases the value of the images in the repository. The system includes bulk upload operations that make is simple to add image collections.

The MorphBank development team deployed a new version of the web site in summer 2006 adding extensive annotation tools for image annotations and determination annotations. A determination annotation is an object in the system that associates a specimen with an assessment of the correctness of its taxon identification. A scientist can create an annotation in order to register agreement or disagreement with the current determination. In the case of disagreement, an alternative taxon is specified. Each annotation contains image annotations that identify areas of interest related to the determination.

The specimens of the Florida State University Herbarium are now included in the MorphBank repository as image and specimen objects. The MorphBank determination annotation tools are being used to evaluate the determinations of the specimen. The tools enable experts to record their assessments of the quality of the herbarium metadata and to correct the determinations as necessary.

The next annotation capability that will be included in the MorphBank tools is character state annotation, the association of a phylogenetic character state with a specific area of interest of an image. Development has begun on tools to create and manipulate phylogenetic characters and their states. These tools support importing and exporting of Nexus files.

Other systems, for example, Mesquite and Morphobank. support the creation of character matrices that associate species with character states. Morphobank provides tools to include and annotate images of the species.

In comparison, character state annotations in MorphBank are focused on identifying areas within images that are show particular character states. Each image may have many annotations. MorphBank image annotations may be used to create character matrices that relate species to states, and may also be used to relate specimens to states.

The search and browse capabilities of the MorphBank tools will provide scientists with capabilities to find and compare character state annotations in ways that are not currently possible. Scientists can - find all images that display particular characters or states, use image annotations to clearly identify and differentiate the states, and easily find appropriate characters to use for their specimens. The character definitions will be shared among all users of MorphBank.

Current development plans call for the character state annotation tools to be fully implemented and deployed in advance of the TDWG 06 meeting.

The presentation will include illustrations of the use of both determination and character state annotations. Details of the data models that support annotations and demonstrations of the tools will be included.

Support is acknowledged from: NSF

8. Biodiversity Heritage Library

8.1. Botanicus.org: Prototyping a Web 2.0 Interface to Digitized Taxonomic Literature

Chris Freeland, Douglas Holland
Missouri Botanical Garden

The Missouri Botanical Garden has been digitizing taxonomic literature since 1995, starting with rare monographic works, and presenting to users as a collection of static HTML pages. Since that time we have changed our selection criteria to include large multi-volume journals and have radically changed how we manage and serve those digitized volumes. The culmination of this work is now available online at Botanicus.org.

Translating the experience of using a physical bound object to an online display of that object is difficult and has been limited by technological gaps in supported browser functionality. MBG is not alone in this effort to digitize literature; many natural history museums and libraries have begun scanning materials individually, but in nearly all cases, including our own, the user interface to these works fails to provide an interactive, multivalent editing system for scientific annotations and taxonomic inquiry. As large-scale scanning efforts like the Biodiversity Heritage Library emerge, a new interface into that digitized literature is required.

Web 2.0 is a term for the paradigm shift in web publishing from individual sites and static content to service-aware web applications that provide robust computing environments. Applications like Google Maps have shown how content can be integrated from disparate sources using open APIs and presented to users in a sophisticated interface within a standard web browser. Further, Wikis and other editing systems have shown promise for how a community of users can edit, annotate, and interlink textual materials. Those users now expect the same rich environment for digitized scientific literature.

MBG is prototyping one such interface for scientific literature at Botanicus.org. Through integration of service-based applications, we are building a system that will allow users to view, edit and annotate scientific texts and interlink nomenclatural databases using taxonomic intelligence. The presentation will cover the proposed system design in full, as well as demonstrate the components already deployed and in use at Botanicus.org.

8.2. Digitizing the Legacy Literature of Biodiversity: An Introduction to the Biodiversity Heritage Library (BHL)

Neil Thomson
Natural History Museum, London UK

The Biodiversity Heritage Library (BHL) is a consortium of eight libraries comprising four museums, three botanic gardens and a university department. The BHL is developing a strategy and operational plan to digitize the published literature of biodiversity held in their respective collections and to make that literature available for open access and responsible use as a part of a global “Biodiversity Commons”.

The combined holdings of these libraries are around two million volumes, assembled over the past 200 years. Many of these holdings are rare or unavailable in the biodiversity-rich countries. The widespread availability of the Internet makes it feasible for researchers to access this material without having funds to travel to the holding library.

The BHL is a focused digitization project in a subject area where the oldest literature can be extremely valuable for current research.

Such a project requires close liaison with several communities to be effective. Discussions are being initiated with rights holders, researchers, publishers, abstracting services and developers of taxonomic intelligence tools. The project is also linking with similar digitization projects and the developers of the GBIF information architecture.

This presentation covers the objectives and current progress of the BHL project. In conjunction with the following presentation by Chris Freeland, we aim to stimulate discussion about additional features that researchers would find most useful in such a resource. Examples might include services for citations and bibliographies, or the use of unique identifiers, such as LSIDs and DOIs for linking the literature to specimen data and taxonomic concepts.

9. Non Symposium Session

9.1. Natural Collections Descriptions: An Introduction to the NCD Data Standard

Neil Thomson
Natural History Museum, London, UK

Natural Collections Descriptions (NCD) is based on the collection-level data standard used in the BioCASE project. NCD has now been extended to cater for library and archive collections, in addition to collections of specimens and observations.

The standard is primarily intended for resource discovery, particularly of collections that have no item-level database. NCD is lightweight as it is pitched between very general resource discovery standards such as Dublin Core (DC) and rich collection description standards such as the Encoded Archival Description (EAD). It is however possible to extract a Dublin Core record from an NCD record or, conversely, to use an NCD record as the basis of an EAD record as and when resources allow.

NCD is one of the emerging standards in TDWG and is currently in testing. The current version of NCD is available as an XML schema, along with the charter and a draft of the User Guide from the TDWG Website (http://www.tdwg.org).

If collection descriptions are new to you, this introduction will outline some of the uses of collection descriptions and the top-level structure of the NCD standard. Progress made since the 2005 TDWG meeting will be outlined. Interested delegates are welcome to attend the workshop session on Tuesday afternoon of TDWG 2006.

9.2. NLBIF Metadatabase: An Implementation Based on NCD Schema

Wouter Addink[1], Ruud Altenburg[1], Cees Hof[2]
[1] ETI BioInformatics, [2] Netherlands Biodiversity Information Facility (NLBIF)

Easy access to metadata on biological collections is essential for optimizing collection use. There are a vast number of collections spread over museums, zoos, botanical gardens, research institutions etc. We need simple methods to find out what collections exist, what they contain, how they can be accessed, and how they are related. Solutions should provide access and search facilities to descriptions, location and contact information, and mechanisms to maintain up-to-date metadata information about collections.

Gathering and updating information about existing natural collections is a time-consuming process and often requires direct contact with collection owners. Inventories are generally done on a regional or national level. The use of a data exchange standard would be beneficial for the exchange of collection metadata, and to aggregate metadata into international overviews. The TDWG Natural Collections Descriptions subgroup creates such a standard called the NCD Schema.

NLBIF (the Netherlands' node of the Global Biodiversity Information Facility), needs an inventory of Dutch natural collections and related organizations for management and resource discovery purposes. NLBIF wants to make this information freely available on the Internet. The National Node Input Tool (NoDIT) database was originally created for the European BioCASE project. A Web interface for the database was built and the information was also made available through the BioCASE network. Based on the experience with the NoDIT database and the emerging NCD standard, a new database schema was developed for NLBIF by ETI. This schema is compatible with NCD. A Web interface for searching and editing the metadata will

become available within the new NLBIF portal (scheduled for the end of 2006). Data will also become available as Web services in NCD compliant XML and other formats.

Although NCD is an emerging standard (0.3 version was used), it was useful to construct the database and to consider possible NCD schema changes. Such issues are communicated with the NCD subgroup and through this TDWG meeting to provide input for further development of NCD into a standard for Natural Collections Descriptions.

Support is acknowledged from: The BioCASE helpdesk

9.3. Best Practice For Updating and Versioning of TDWG Standard XML Schemas

Walter G. Berendsohn[1], Andrea Hahn[2], Anton Güntsch[1], Chuck Miller[3], Javier de la Torre[4], Markus Döring[1], Neil Thomson[5], Patricia Mergen[6], Renato De Giovanni[7], William Ulate[8], Wouter Addink[9]

[1] BGBM Berlin-Dahlem, [2] GBIF Secretariat, Copenhagen, [3] Missouri Botanical Garden, St. Louis, [4] Museo Nacional de Ciencias Naturales, Madrid, [5] NHM London, [6] Royal Museum for Central Africa, Tervuren, [7] CRIA, Campinas, [8] INBio, Heredia, [9] ETI, Amsterdam

The ABCD (Access to Biological Collection Data) Schema version 2.06 was proposed as a TDWG standard by the ABCD content definition subgroup and ratified by the TDWG meeting in St. Petersburg in 2005.

ABCD provides a provisional mechanism for extending the schema, by including extension elements (typed as xs:any) in three locations: at the level of the unit, within site descriptions and within identification results. These elements serve as slots for the inclusion of third-party-schemas (or parts thereof), and can help to avoid duplicating the efforts of other communities in developing data models (e.g. for geographical data). The ABCD extension elements also provide a support mechanism to allow user communities to add missing elements to the current version. Such elements can then be treated as candidates for formal inclusion in future versions.

A new version of the ABCD schema can be released whenever a significant number of necessary additions and changes have accumulated and/or structural changes are urgently needed. The changes can then be integrated using a new namespace for the new version of the ABCD schema.

Several problems were recognised during the implementation of the latest version of the ABCD schema. These show that an additional mechanism for is required for correcting errors between the release of new versions. The ABCD subgroup met in July 2006 and developed a mechanism which can be used to version and update the XML schemas currently in use as TDWG standards.

The group proposes that the example of GML should be followed for these interim corrections, i.e. that the schema should be changed without changing the namespace but in a way that ensures full backward compatibility. Any corrections must not introduce changes which will break applications using previously approved versions of the schema in the same namespace. The root element of the schema should include a version attribute indicating the schema version number. This number should agree with the namespace assigned to the schema. Minor changes are then indicated by letters. For example, the first interim version of ABCD 2.06 should receive be identified as 2.06a.

In practical terms, backward compatibility is maintained when

- no elements or attributes are deleted (although elements may be marked as deprecated)
- no elements or attributes are renamed

- the semantics of all elements are left unchanged
- type changes are restricted to assigning types to previously untyped elements
- new elements are added

The changes between v. 2.06 to 2.06a are documented in detail on the ABCD Wiki (http://ww3.bgbm.org/abcddocs/). The main changes consisted of defining types for several untyped elements that escaped attention during the last version upgrade, eliminating some points of confusion concerning contact information in metadata, adding a new extension slot at the dataset level to allow the metadata to be extended and the addition of several elements for better compatibility to the Darwin Core and the HISPID standard.

Support is acknowledged from: The Gordon & Betty Moore Foundation TDWG Infrastructure Project; CODATA

9.4. The Big Dig

David Vieglais
University of Kansas

DiGIR (Distributed Generic Information Retrieval), a protocol developed with input from the TDWG and receiving significant support from the National Science Foundation, is currently the most widely deployed mechanism for accessing specimen data from natural history collections. DiGIR has approximately 170 registered installations. It forms an integral component of the GBIF data network and a number of domain specific groups such as the Ocean Biogeographic Information System (OBIS) for oceanic observation and specimen data, or in taxonomy oriented systems such as the Mammal Networked Information System (MANIS), HerpNet, ORNIS, and FishNet2.

One significant drawback of current DiGIR deployments is the lack of a unified mechanism indicating the status of the entire network and its constituent components. These components include data provider software, various portals for accessing the data, and data federation definitions implemented as XML schema documents. Most DiGIR deployments are at relatively unsupervised locations with minimal technical expertise available. It is therefore desirable to develop a system for monitoring the installed providers, and optionally alert administrators to possible problems. The "Big Dig" is an ecoforge.net project was started in 2006 and is hosted by the Natural History Museum and Biodiversity Research Center of the University of Kansas. The project aims to provide an automated system for evaluating the status of all known DiGIR data provider installations identified by examining registries of known networks, and reporting the properties of the installed software, referenced data schemas and the performance and reliability of operation.

I will present a summary of network statistics collected during 2006, an analysis of federation schemas in use and some evaluation of the data accessible through the network.

Support is acknowledged from: The National Science Foundation, TDWG, GBIF

10. Building Biodiversity Data Applications

10.1. A Web Based GIS Tool for Exploring the World's Biodiversity: The Global Biodiversity Information Facility Mapping and Analysis Portal Application (GBIF MAPA)

Robert Guralnick[1], Paul Flemons[2], David Neufeld[1], Ajay Ranipeta[2]
[1] University of Colorado, [2] The Australian Museum

Legacy biodiversity data from natural history and survey collections are rapidly becoming available in a common format over the Internet. Almost 100 million records are being served from the Global Biodiversity Information Facility (GBIF). However, our ability to use this information effectively for ecological research, management and conservation lags behind. One solution is a web-based Geographic Information System for visualization and analysis of these biodiversity data. This paper reports on a case study system developed for deployment at distributed database portals. The system, GBIF-MAPA (Mapping and Analysis Portal), allows users to explore worldwide biodiversity data and then perform a set of analyses including summarizing species richness for taxonomic groups of interest. Technical and research challenges included: assuring fast speed of access to the vast amounts of data available through these distributed biodiversity databases; developing open standards based access to suitable environmental data layers for analyzing biodiversity distribution; building suitably flexible and intuitive map interfaces for refining the scope and criteria of an analysis; and building appropriate web-services based analysis tools that are value to the ecological community. The results manifest the value of online biodiversity GBIF data. After discussing how we overcome these challenges, we provide thoughts on the future for web based biodiversity data acquisition and analysis.

10.2. 3I: On-line Virtual Taxonomic Revisions

Dmitry A. Dmitriev
Illinois Natural History Survey

Taxonomic revisions of diverse groups of organisms are challenging because they generally require efficient management and synthesis of large amounts of nomenclatural, morphological, and distributional data. When undertaken using traditional methodologies, such projects often require many years to yield publishable results. Technological advances, including relational databases, digital imaging, and Internet dissemination, provide the means to overcome some of the logistical problems inherent to revisions of highly speciose taxa, and provide systematists with tools to increase both the quality and quantity of such studies.

3I (Internet-accessible Interactive Identification) is a set of software tools intended to facilitate the efficient production of Internet-based virtual taxonomic revisions and published monographs. The package facilitates storage, retrieval and integration of taxonomic nomenclature, specimen-level data on distribution, ecological associations, morphological character, associated illustrations, and bibliographic information. These data are stored in a customized MS Access database. Web interfaces for specialized querying of the database were developed using ASP (Active Server Pages) programming technology. These interfaces include simple and advanced searches on any field in the database, and interactive keys designed to include attributes similar to those of Delta IntKey and Lucid (two popular programs for development of interactive keys). The main features of 3I keys are the following: 1) 3I keys are multi-entrance polytomous keys, with unlimited number of characters, character states, and taxa; 2) after each step of identification the characters not relevant for further identification are removed from the list, not relevant states are marked; 3) 3I keys support numeric characters; 4) the characters in key can be sorted by morphology, by rank (assigned by the author), or by separating power recalculated after each step of identification; 5) characters can have hyperlinks to explanatory images; 6) a key can handle taxa of different hierarchical levels, and

the software can also generate keys for higher hierarchical level taxa, based on data matrix scored for taxa of lower hierarchical level; 7) uncertainties and user-specified error tolerance are allowed during identification; 8) phenetic trees are generated from the morphological data, or the data matrices can be exported in a format suitable for phylogenetic analysis and 9) 3I has an utility to convert interactive keys into conventional ones.

Clicking on a taxon name in the search or key interface opens another browser window that displays a complete taxonomic treatment of the taxon generated on the fly from the underlying database. The listing is organized as it might appear in a published monograph, including synonymy, description, distribution map, list of specimens examined and table of associations (for species only) and bibliography. For higher taxa, the taxon treatment includes a link to an interactive key to the included subordinate taxa. Because 3I is web-integrated with database queries performed on the server side, the web interfaces are stable and are compatible with virtually any computer operating system/browser combination.

More information and examples of the interactive keys ant taxonomic databases developed using 3I are available from http://ctap.inhs.uiuc.edu/dmitriev/.

10.3. TAXI: A Framework for Synchronizing Taxonomic Change Across a Distributed Network

Maggie Woo, Leah Oliver
NatureServe

NatureServe needs to synchronize and reconcile variant taxonomies in use across the 76 nodes of its distributed database network. Synchronization of taxonomic concepts across this network requires managing and implementing taxonomic changes to provide consistent data products for end users, and to facilitate future distributed online querying capabilities.

TAXI is a framework for communicating taxonomic change among organizations or software applications, allowing local nodes to review and apply decisions (adopt or reject) prior to implementing changes to their local systems. NatureServe's reference implementation of this taxonomic “shuttle service” will be specific to the application of taxonomic change management within the NatureServe Network and its supported software applications (e.g., Biotics 4). The initial implementation will include a TAXI Registry, which publishes Taxonomic Change Capsules (XML document) that may be consumed by any interested application to examine or report changes implemented or endorsed centrally by NatureServe. NatureServe Network nodes will behave like outside organizations by consuming capsules directly from the TAXI Registry in order to implement taxonomic change to their internal Biotics Systems. Conversely, in a future implementation NatureServe may consume capsules sent from local nodes in order to track taxonomic changes being adopted in the field, as a means to research and consider changes for endorsement and network-wide adoption.

The TAXI Framework's principles are generic enough to be of interest to any organization that generates taxonomic change or that wishes to communicate the relationship between two taxonomic views of a particular set of taxonomic concepts. It may have future application for harvesting information about taxonomic change from organizations that implement a TAXI Registry. Although designed with species taxonomies in mind, the framework is being designed to support management of similar needs for Ecological Classification Concepts.

Support is acknowledged from: National Science Foundation, U.S. Environmental Protection Agency

10.4. The Importance of Standardization of the Data Format: A Case Study from the National Herbarium of the Netherlands

Luc P.M. Willemse, Johan B. Mols, Peter C Welzen, Erik F Smets
National Herbarium of the Netherlands

The National Herbarium of the Netherlands (NHN) has been storing label data of its botanical collections for over a decade in digital format. So far about 850,000 collections have been digitized. During this period various choices had to be made to improve and maintain high levels of data quality. Implementing internationally accepted standards, the use of search lists and the built-in functionality in collection registration software were helpful in reaching acceptable levels of data quality. The most important factor in improving data quality and consistency in data entry within and between institutes was a protocol with data-entry guidelines. We demonstrate that such a protocol is crucial in addressing the exact representation (syntax and/or otherwise) of specific pieces of information (the data format). Using format-related difficulties for particular data elements as examples, we suggest that data accessibility and data exchange are often better served by improving the consistency of the data format used rather than the data structure: adjusting data structures is often relatively easy when compared with adjusting the data format.

Based on the experience gained in database management and botanical collection digitisation at the NHN over the past decade, we suggest the need for standards for the format of particular data elements like collector names, geographical names, dates and numbers.

10.5. Tracking Our Progress: Improving the Search for Biological Information Online

Rebecca Shapley
Google

Many challenges exist on the road to providing an excellent experience while searching for information about taxa online, including:

- Collocating all information about taxa, when large parts are in databases (deep web), off-line (taxonomic literature as yet undigitized and/or OCR'd) and in non-text formats (DNA and other molecules, images/video, specimens);
- locating information about the same taxon filed under different names;
- missing information, best provided by extrapolating from related known taxa;
- browsing to information about related taxa and other biological entities;
- visualizing relevant geographic locations;
- distinguishing current information from out-of-date information and
- communicating to non-taxonomist audiences

E.O. Wilson's vision of a centralized resource with authoritative information on every species known to science (an Encyclopedia of Life) has shaped biodiversity informatics for decades. Though the Encyclopedia entries have become database records, the metaphor evokes a static collection bound between two covers and published as of a given moment in time. Even database records have practical limitations-

- they assume all records have a certain equality, and
- fields are often blank or have multiple resources competing to fill them.

This limits our vision for what it means to organize biological information from this planet.

Instead, we should seek to have a search engine tap the dynamic flow of information from all authoritative sources. A search on a namestring taps the flow and returns all the relevant information available for the taxon. Information gets richer as we develop the capacity to

automatically process it into useful representations, including distribution maps and tentative synonym lists, even species in the news. Experts' updates should be quickly and accurately reflected in the content. Providing a quality experience of searching for biological information requires a shift of focus from edited content to information retrieval, and collaboration among people and institutions with diverse skills.

I propose that we measure our progress by how well we are serving three key information use cases, for both professional and popular audiences-

- What's in my backyard?
- What did I find?
- Tell me more about it.

Good answers vary by audience. We should serve real-time, relevant, quality information to both professional entomologists using DNA barcodes to identify one insect specimen among thousands of possibilities and school kids who want to turn a few simple observations into a name of a bird and find out what it eats.

10.6. Experiences on the Application of Services Oriented Approaches to the Federation of Heterogeneous Geologic Data Resources

Douglas R. Fils, Cinzia Cervato
CHRONOS, Iowa State University

The federation of databases is not a new endeavor. Great strides have been made in health, astrophysics, and other communities. Reviews of those successes indicate that they have leveraged key cross-community core concepts. In its simplest implementation, a federation of databases with identical base schemas that can be extended to address individual efforts is relatively easy to accomplish. A review of CHRONOS's experience (http://www.chronos.org/) with federation of very diverse databases shows that the wide variety of encoding options for items like locality, time scale, taxon ID and other key indexes makes it difficult to effectively join data across databases. However, the response to this is not to develop a large monolithic database, which will suffer growth pains due to social, national, and operational issues, but rather to systematically develop the architecture that will enable cross-resource (database, repository, tool and interface) interaction. Using an ontology to resolve schema relations will be vital to this effort, as will useful metadata on the attributes of the data providers and data quality.

This presentation reports on CHRONOS's experience with services, semantics and syndication with various partners, and the approaches and lessons learned from them. A collaboration between CHRONOS and the Geological and Nuclear Sciences Institute in New Zealand (GNS) has begun development and implementation of a Taxonomic Synonymy Definition Framework (TSDF). Based on the Resource Description Format (RDF), the TSDF defines a mechanism to codify and exchange synonymy concepts in RDF using principles and concepts from the Simple Knowledge Organization System (SKOS). In addition, the implementation of the GeoSciML schema of CSIRO Australia, specifically the Geologic Time schema, into CHRONOS's structure has revealed many issues related to the exchange of data synchronized to different time scales.

CHRONOS has several working relations with various data providers (e.g., groups of taxonomists) and this creates a unique matrix of requirements. To address this, we have implemented methods that conform to open standards and formats in a services-based approach. This has facilitated the development of the architecture that will be illustrated in this presentation.

10.7. Non-Functional Requirements for Invasive Species Data Exchange

Robert A. Morris[1], Michael T. Browne[2]
[1] UMASS-Boston, [2] IUCN Invasive Species Specialist Group

Non-functional requirements for a software system constrain the design, but neither expand nor constrict its functionality. The XML Schema under design by the Global Invasive Species Information Network (GISIN) will serve a community with widely varying social and political requirements and technical resources. Addressing this variation gives rise to non-functional requirements leading to a number of technical compromises that reduce the utility of validating XML parsers, with the attendant need to provide for external validation tools. Some are familiar from other TDWG endeavors, such as requirements to support multiple expressions of the same semantic concepts, or different legal, political, or regulatory requirements to use particular terms for the same thing

An important example is the requirement by many jurisdictions to use specific values for certain enumerations, such as geographic feature types (e.g. water body types). To address this we are exploring the use of a GUID-based mechanism, by which resolution provides permitted values at run-time. An external validity check can thus detect whether the resolved permitted values contains the offered one. This can improve the utility of external ontologies which try to map between values offered by different providers. In the current draft, such external “Defined Schemas” are always accompanied by an enumerated preferred element that providers are strongly encouraged to use in preference or in addition to a Defined Schema. When resolution is to data in the instance document, the Defined Schema corresponds functionally to the architecture of the TDWG SDD Schema, in which descriptive data are constrained to values provided elsewhere in the document. The Defined Schema mechanism is technically simpler, but requires external validation and processing.

Sometimes data of great importance to one provider simply cannot be offered by another. This may be because the data are not gathered, the provider has insufficient technical expertise to synthesize them from existing data, or policies prohibit sharing them. However, data sharing is critical to the management of invasive species, and encouraging openness is the prime non-functional requirement of the GISIN Schema. Promoting acceptance presents a social tension between providing for useful data and reducing mandatory elements. To address this, a number of optional elements carry the attribute recommended = “true”. This is meant to urge writers of import software to support this element, at least to the extent of signalling the absence of a recommended datum.

Some non-functional requirements can be addressed by provision of external invertible transforms, some of which may be required in practice in any case. For example, some providers organize checklists, one of the supported data types, first by species then by location, and some by location and then species. Strong typing permits our current schema to support both robustly, with the result that all import and export software must be able to transform between them, e.g. by a pair of XSLT transforms. Hence, arguably, the schema could be simplified by supporting only one, and providing the transforms as part of a standard.

Support is acknowledged from: U.S. National Science Foundation; Global Biodiversity Information Facility; Convention on Biological Diversity

10.8. The New GBIF Data Portal – Web Services and Tools

Donald Hobern
Global Biodiversity Information Facility

Since 2004, the Global Biodiversity Information Facility (GBIF, http://www.gbif.org/) has been operating a prototype data index and web portal for biodiversity data from around the world. These tools have provided basic mechanisms for users to discover relevant data on the occurrence of individual species from anywhere in the network, whether the source data are shared using DiGIR and Darwin Core or using BioCASe and the ABCD schema.

This prototype data portal has allowed GBIF to learn a great deal about the characteristics of available data and about integrating such data. During 2006, GBIF is completely redeveloping the portal infrastructure based on lessons learned, and plans to release a new data portal early in 2007.

Key innovations in the new architecture will include:

- Greater flexibility in linking to data resources using new data formats;
- Dynamic evaluation of potential inconsistencies and overall characteristics of each data resource (during indexing);
- Improved management of descriptive metadata for each data resource;
- A wide range of web service interfaces;
- Improved and more flexible search options through the HTML user interface;
- User interface components enabled for inclusion within other web sites (accessing GBIF data through web services) and
- Open interfaces for developing visualisations and analyses of GBIF data.

10.9. DNA Barcoding: Bane or Boon (or Both) For Taxonomy?

Mehrdad Hajibabaei[1], Gregory Singer[2], Donal Hickey[3]
[1] University of Guelph, [2] Ohio State University, [3] Concordia University

DNA barcoding has been proposed as a method for quickly assigning biological specimens to known species. Barcoding has also been used to identify putative cryptic species and to assess phylogenetic relationships.

We have developed a method of DNA barcoding that focuses on the primary application of DNA barcoding only, i.e., the assignment of unidentified individuals to known species. Our analysis draws upon both phylogenetic and taxonomic information, but this information is not derived from the barcode data themselves. We use a simple text-searching algorithm to compare a DNA barcode sequence from an unidentified specimen to a database of sequences from taxonomically verified voucher specimens. The assignment of voucher specimens to species is done independently of their barcode sequence information. The placement of these species on a phylogeny is also based on independent phylogenetic information.

We show that this simple approach, which is free of both phylogenetic and taxonomic assumptions, can quickly identify matches in the database. It can also flag sequences with varying degrees of mismatch for future analysis by taxonomists and evolutionary biologists. This method should be especially useful for rapid screening of large numbers of field-collected samples.

Support is acknowledged from: Genome Canada, NSERC Canada

10.10. Tips for Natural History Institutions: Using Google to Improve the Flow of Biological Information

Rebecca Shapley
Google

These suggestions are offered from a win-win perspective. We share a goal of improving Internet users' experiences while searching for biological information. When Google's indices know about information being provided by museums and other natural history institutions, Google can provide better search results when a biological category name is typed in, and send searchers to the relevant websites. Google offers natural history institutions the opportunity to stick with their core competencies of generating and curating biological information, and not spend valuable funds on hosting or scanning infrastructure. With this in mind, suggestions include-

- Host videos on video.google.com
- Museums who publish or have published journals and any copyrighted materials can participate in books.google.com, making the scanned material accessible at a level the publisher feels comfortable with. With the upcoming Online Access program, researchers around the world can pay to view a journal they need for their taxonomic work. Google provides the digitization. Collectively, museum by museum, this can help bring taxonomic literature online.
- Share the existence of structured information datasets through Google Base. Using the bulk upload feature, create a custom data type for your data set, and provide either the data records you wish to share, or enough of the record to create a Base listing that will point to your institution's website where the rest of the data can be found.
- Use Google Co-op to make your web data applications available to your membership directly from their Google.com searches. For example, the IUCN and the Consortium for Barcode of Life have used Co-op's Subscribed Links feature to provide data and links to results from their own datasets at the top of search results when subscribers search on Google.
- Publish geographically enabled data in KML file format, so that people can use the free Google Earth and Google Maps to view it. Pasting the URL of a KML file into the Google Maps search box will display it on the map without any downloading required. This means museums don't need to support a map server to make data viewable.

10.11. WASABI: Web Application for the Semantic Architecture of Biodiversity Informatics

Steven Perry, Dave Vieglais
University of Kansas Biodiversity Research Center

WASABI is a framework for constructing data-sharing networks built on semantic web technologies and standards. It consists of three primary components: a highly configurable server; a customizable portal; and a client library.

WASABI represents data objects internally as RDF resources named by globally unique identifiers using the LSID (Life Sciences Identifiers) scheme. Because of this, WASABI can serve, query, and display complex data models described by multiple RDF-Schemata or OWL ontologies. The use of semantic web technologies makes it easier for WASABI to support the type of modular-but-related schemata currently in development under TDWG. This strategy helps to simplify the task of data integration.

The WASABI server is extensible in that it allows data access protocols to be implemented as plug-ins. WASABI natively supports several standard data access protocols including the W3C standard SPARQL protocol and query language for RDF and OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). When designing WASABI, the implementation of

existing protocols like SPARQL and OAI were favored over the development of a custom data transport and access protocol. This strategy increases the opportunities for interoperability between WASABI services and existing and emerging systems designed outside of the Biodiversity community such as SWED (Semantic Web Environmental Directory), projects from the W3C Semantic Web for Life Sciences group, from ecology (the Ecosystem Location Visualization and Information System), and geography (the OGC's Geospatial Semantic Web Interoperability Experiment).

The WASABI portal harvests data from one or more WASABI servers to build an index. This index is presented to the user for both browsing and searching. Whenever the portal displays a data object to the user, it can notify this usage to the WASABI server that hosts the data. The use of an index to back search and browse operations, combined with a usage-tracking system, balances the needs of data consumers who want fast searches, and data producers who want to track where and how often their data objects are used.

The WASABI client library can be used to create custom applications with the ability to consume WASABI services. The library provides a Java implementation of each of the supported protocols as well as a fast multithreaded HTTP client that is capable of querying many servers simultaneously. Because WASABI uses standard protocols, the client library can also be used to interact with any SPARQL or OAI web service.

WASABI is being developed at the University of Kansas Biodiversity Research Center. It is funded by a U.S. National Science Foundation grant and will be used to build the FishNet2 and PlantCollections networks in 2006 and 2007.

Support is acknowledged from: US National Science Foundation

10.12. An Internet Platform for Cybertaxonomy

Walter G. Berendsohn, Malte C. Ebach
Botanic Garden and Botanical Museum Berlin-Dahlem

One of the focal points in the establishment of the EU-supported European Distributed Institute of Taxonomy (EDIT) is the creation of an “Internet Platform for Cybertaxonomy” This is an effort directed at the practical application of methods developed in information science and biodiversity informatics for use in revisionary taxonomy and taxonomic field work. At the same time, an integration of the participant institutions’ biodiversity informatics and IT resources is to be achieved.

During the first 18 months period of the project we will set out to analyse the pre-requisites for the establishment of co-operative processes, model the information domain and particularly the work processes, and provide first practical applications in a rapid prototyping approach addressing identified bottlenecks. On the organisational side, points that need to be addressed include forming inter-institutional coordination structures to identify parallel activities, inventory techniques and procedures used in these activities, investigate possibilities for harmonised procedures and/or techniques, and pursue the integration of the data holdings of the participating institutions. With respect to the necessary analysis of the taxonomic work process, an in-depth modeling effort of the revisionary work process and of taxonomic field work in the context of inventory and monitoring projects is carried out. In parallel, we will reach out to identify existing electronic tools for taxonomists, test their usability and, where necessary, provide developer time to improve their interoperability. The aim is to build a distributed computing platform that assist taxonomists by providing certified and tested existing labour and time saving tools in order to do taxonomy expediently and via the Web. Although the approach in the short and medium term will be strictly pragmatic and product oriented, room will be left to think about long term integration of taxonomy into a broader eScience environment.

The character of EDIT as a joint effort of large taxonomic institutions with high level support in the institutional hierarchy provides a unique opportunity to build sustainable structures – if acceptance of the tools by the taxonomic researcher can be assured. In order to achieve this, EDIT has already begun to enlist taxonomists to test, document software and provide feedback on various existing taxonomic tools.

EDIT project: http://www.e-taxonomy.eu/
EDIT Work Package 5 (Internet Platform for Cybertaxonomy): http://www.cybertaxonomy.org

Support is acknowledged from: European Commission

10.13. ZooBank - The Open-Access Animal Name Registry

Andrew Polaszek
International Commission on Zoological Nomenclature

ZooBank, an open-access registry for the scientific names of animal species, genera and higher taxonomic categories, was released as prototype in August 2006. It currently comprises 1.5 million names that have been compiled by the periodical Zoological Record over the last 150 years. Releasing zoological data on such a large scale freely to the public is without precedent and is in keeping with the current trend in science for open-access publishing. ZooBank will allow the official names of animals to be recorded more quickly and effectively in future, and will ease communication between scientists.

The scientific names of animal species are crucial to effective global communication about them, and hence their use and conservation. If you can't agree on the name of a disease-bearing microbe, vital food species, or threatened animal, you can't even begin to combat, exploit or conserve them.

The universal acceptance and adoption of a system for naming animals is an incredible achievement for mankind, and started in 1758 with the publication of the 10th edition of Systema Naturae by the Swedish biologist Carolus Linnaeus. Almost exactly 250 years later we are on the verge of achieving something even greater, the universal availability, for the first time in history, of a complete list of all the scientific names of the 1½ million known animal species, free to anyone at the click of a mouse.

The 1½ million original scientific descriptions of animals are located in hundreds of thousands of different journals and monographs, and consequently often very difficult to access or retrieve. ZooBank will bring all these animal species names together in a single database, something that has never been done before. By introducing a mandatory registration system for new species, the ZooBank database will by definition always be complete, enabling access to every known animal species name for the first time.

ZooBank currently is a joint venture between the International Commission on Zoological Nomenclature (ICZN) and Thomson Zoological Ltd, producers of the periodical Zoological Record. We are presently exploring closer links to GBIF, and also GenBank and MorphBank, in order to be part of the provision of a comprehensive and complete service of data on authoritative animal nomenclature, taxonomy and images.

We are also developing full compatibility with the United States Government's Integrated Taxonomic Information System Database (ITIS), and other comparable international initiatives that affect legislation for quarantine, conservation, human and animal health and biodiversity studies.

Support is acknowledged from: Wellcome Trust, Taylor & Francis Ltd

10.14. Taxonomic Literature - Standards and Synergies

Anna L. Weitzman[1], Christopher H.C. Lyal[2]
[1] National Museum of Natural History, Smithsonian Institution, [2] The Natural History Museum, London

A standard is needed for taxonomic literature, especially given the increasing number of books and papers that are and will be appearing on the Web, in particular as the Biodiversity Heritage Library gears up. Such a standard should be available in a range of formats.

As a first step we need a standard for citations of published works. Without this there will be problems in using digitized text with other applications. Libraries have been using several different standards, and our objective should be to develop a standard that meets the specialised needs of the taxonomist user but which also provides easy cross-links to library standards and is therefore interoperable with them. We also need a simpler format to use for citations within our taxonomic treatments.

Finally, and most extensively, we need a standard to allow us to place taxonomic literature itself into an interoperable form on the Web. Such a schema should enable interoperability with congruent information including specimen data, nomenclatural data and taxon concept data. Several schemas have been proposed and we compare these for their functions and applicability. A TDWG working group has made progress on the simpler citation standards, and will soon be starting to focus intensely on the more detailed full schema.

10.15. Developing Uncertainty Measures Related to Taxonomic Determinations

Larry Speers[1], Arthur David Chapman[2]
[1] GBIF, [2] Australian Biodiversity Information Services

The value of georeferenced biodiversity data to end users is greatly increased if these data include appropriate measures of the level of uncertainty for each georeference. (See: Chapman, A.D. & J. Wieczorek, eds., Guide to Best Practices for Georeferencing, GBIF, Copenhagen, 2006, http://www.gbif.org/prog/digit/Georeferencing). Such measures of uncertainty help users to determine how fit the data are for particular uses and hence serve as a measure of data quality.

Users have also requested similar documentation of the level of uncertainty surrounding taxonomic determinations for both specimen and observational records. Appropriate factors in such documentation could include:

- how current the identification is;
- how experienced the individual was who performed the identification;
- how suitable the material was for carrying out an identification;
- what the conditions were under which the identification was made and
- what the basis is for the cited determination?

The presentation will discuss various approaches to these developing such measures with the aim of stimulating further discussion.

10.16. The Growth of PLANTS

Gerald Guala
USDA NRCS National Plant Data Center, Baton Rouge, Louisiana

The USDA NRCS PLANTS database is evolving with new functionality, data and data management paradigms. The database sees a minimum of a million unique user sessions every month from a global and diverse user community, so any changes have a large and immediate impact.

New capabilities at PLANTS will be discussed with emphasis on the move to a concept-based and a more distributed data management paradigm, the first releases of data in the construction of a large scale interactive identification environment, and new web services. An interactive key to all wetland monocots in the US will be available.

Feedback from the community on potential web services is expressly requested. See http://plants.usda.gov for further details.

Support is acknowledged from: United States Department of Agriculture

10.17. Aligning Biodiversity Software with User Needs: An Industry and Market Analysis

Bruce A. Stein[1], Larry Sugarbaker[1], Keith Carr[1], Christopher Lenhardt[2]
[1] NatureServe, [2] Columbia University

To identify opportunities for developing software that is broadly applicable to the needs of the biodiversity community we carried out a user needs assessment coupled with an industry and market analysis. User needs were identified through a combination of in-depth interviews and web-based surveys, while the industry and market analysis was based on a review of more than 635 software offerings. The most notable feature of the biodiversity software development "industry" is its high degree of fragmentation and the large number of locally developed applications with small user bases. The top ten providers, based on software licenses, account for just 35% of the market's estimated $15 million in addressable revenue. With an estimated 50,000 users split among government, NGO, academic, and for-profit segments, the biodiversity software community is very small compared with the 2 million users for GIS software products. Among the most frequent requests in the user needs assessment were applications for gathering and managing observational data, including hand-held field data input devices. Responding to the results of this survey NatureServe has begun developing a web-hosted application for observational data management—known as Kestrel—based on a newly developed provisional observation data standard. With support from the National Science Foundation, we are launching development of a handheld field data logger (or "BioPDA") designed to apply contemporary geospatial data management concepts in support of digital field data capture.

Support is acknowledged from: Gordon and Betty Moore Foundation

10.18. EDIT and the European Taxonomic Information Services

Yde de Jong[1], Eduard Stloukal[2]
[1] Zoological Museum Amsterdam, [2] Department of Zoology, Comenius University Bratislava

EDIT (European Distributed Institute of Taxonomy), the European Network of Excellence on Taxonomy of 27 prominent taxonomic institutions situated in European Union and some external countries, launched its activities on March 2006 and its diverse activities are planned for period of five years. Fulfilling Europe's contribution to worldwide species list initiatives requires establishing a secure organisation and management for European biodiversity information databases and repositories. EDIT specifies work towards this taxonomic information infrastructure network in workpackage 3.2, including the founding of a Pan European checklist.

EDIT workpackage 3.2b will focus on methods to maintain, update, integrate and improve its repository contents by:

- organising involved experts into a network;

- achieving a common management approach for European electronic biodiversity data that arranges European taxonomic experts into a partnership structure, deals with ownership and copyright issues proceeding from the efforts of the Society for the Management of European Biodiversity Data (SMEBD), and organises expert and national focal point meetings;
- establishing plans and protocols to generate the checklist's required content and updates;
- setting up mechanisms for hosting, validation and extension of these species repositories to ensure their continuation;
- locating financial support for the European species lists for technical labour, data management and system maintenance;
- the integrated publication of European species list data and initial steps towards the production of an annual edition of the Pan-European Species Checklist;
- contributing to global biodiversity informatics standardisation efforts, for example setting up a common management hierarchy for the European databases and participating in the preparation of a consensus management higher hierarchy to be adopted by the global Catalogue of Lifeand GBIF-ECAT programs;
- exploring more sophisticated classification methods, for example, non-ranking but sequential hierarchies that are more efficient and more stable to internal changes;
- securing long-term delivery of data from the Pan-European species lists projects Fauna Europaea, ERMS, and Euro+Med PlantBase into GBIF, which currently occurs via the Catalogue of Life via Species2000 Europe;
- adding and extending data types as appropriate, for example to support more detailed geographical units, common names and the original descriptions of species (or original reference) and
- extending the geographic scope of the current Pan-European databases to cover the Caucasus, the African-Mediterranean, and Arabic areas as well as the Russian eastern Palaearctic.

EDIT workpackage 3.2 is coordinated by the Zoological Museum Amsterdam. Partnerships will include the Pan-European species lists projects (Fauna Europaea, ERMS, and Euro+Med PlantBase) and their associated institutes and organisations, as well as partners from the EDIT Expert and Expertise basis workpackage (EDIT workpackage 2) and selected others. For partnership and other details see the respective EDIT websites: www.mnhn.fr/edit and www.e-taxonomy.eu.

10.19. Wetland Information Network

Santosh Shantaram Gaikwad
Salim Ali Centre For Ornithology & Natural History

There is great need to collect, collate and disseminate wetland data, as it is now globally recognized that fresh water biodiversity is among the most threatened in the world. According to recent studies (Vijayan et al 2004 and Prasad et al 2004), the situation of the wetlands in India is no different. The main goal of the Wetland Information Network is to promote online access to wetland-related information. This paper presents our experience in this area. The Wetland Information Network is part of the Environmental Information System (ENVIS) created by the Indian Ministry of Environment and Forests.

Wetlands of India (www.wetlandsofindia.org) provides a vast collection of spatial data about wetlands available at the Sàlim Ali Centre for Ornithology and Natural History (SACON) through static and dynamic mapping tools. Until recently, access to digital wetlands data was limited to simple JPEG or PDF maps. The map portal makes use of DjVu from Lizard tech (http://www.lizardtech.com) for simple maps. ALOV (http://alov.org) map, a free Java Geographic Information System (GIS) technology, is used to provide interactive maps.

DjVu/Document Express is a suite of applications for creating and viewing highly compressed documents. It can be used for geographic image data and document management. DjVu produces small files even for high resolution maps, so sharing data using this format can be very effective. It offers a free, lightweight plug-in / viewer that requires minimal memory from the client. DjVu technology not only helped delivering maps easily on Internet but also helped publishing mangrove atlases and wetlands reports in the same way.

ALOV Map/TM Java is a free, portable Java application for publishing vector and raster maps on the Internet. It supports a complex rendering architecture, unlimited navigation and allows working with multiple layers, thematic maps, hyperlinked features and attribute data. All spatial and non-spatial datasets being used are stored in a MySQL database. Providing GIS functionality and simple map viewing capabilities over the Internet allows an organization to share geospatial data that have been collected over many years.

WordPress (wordpress.org) blogging software has been used to set up the news section on the Wetlands of India website. Web interfaces to a bibliographic database and a wetland species database were also developed as part of the Wetland Information Network.

Wetland Informatics (WI) is the beginning of a framework of Web based tools. WI provides the general public, administrators, mangers and other stakeholders with a better geographic perspective of the wetlands thus allowing wiser use of this information. It is hoped that this initiative will become an effective tool in providing detailed environmental data about wetlands, important landscape features, and other information that can be helpful with conservation issues.

10.20. Untangling Names: Lessons Learned from the Linking of IPNI and TROPICOS

Julius Welby[1], Robert Magill[2], Sally Hinchcliffe[1]
[1] RBG, Kew, [2] Missouri Botanical Garden

In 2005 IPNI received a small grant from the Moore foundation for data standardisation work on IPNI, including the automated linking of names from IPNI and TROPICOS. Data in IPNI contain many irregularities including (but not confined to) scanning errors, orthographic variations in the names themselves, non-standardised author and publication abbreviations, differences in recording collations, duplication of records and parsing errors. Homonymy rates in botanical names are estimated at about 3%, further complicating the problem with the likelihood of false positives in any automated match routine that is sufficiently lax to overcome the problems noted above. Linking IPNI names with their better standardised TROPICOS equivalents would speed the process of standardisation and help identify duplication and fill in omissions in the data (particularly of basionym authors in earlier records). However, without standardising, reliable matching is difficult. Without reliable matching, the standardisation task is made more prolonged. Good, robust matching routines that use fuzzy-matching techniques and other intelligent approaches to non-standardised data will be invaluable not just to this immediate project but to other IPNI users, and to anyone with similar legacy data to standardise. With the adoption of GUIDs for IPNI and other systems, routines to cross link names with an IPNI ID will be an important step towards the disambiguation of Biodiversity data.

Initial findings from our matching investigation have been:

- Fuzzy matching across large datasets on multiple fields is computationally expensive, making performance (speed) of matching a significant factor;
- A sequential approach to field matching, using 'must match' parameters can significantly reduce the computational overhead of data matching and

- A Python framework has been created utilising this approach, and this has produced encouraging results when run against real plant name citation data from IPNI and TROPICOS

Support is acknowledged from: Gordon & Betty Moore Foundation

10.21. Providing Itinerary Related Datasets and Tools for Integration, Visualisation and Quality Check

Patricia Mergen[1], Bart Meganck[1], Danny Meirte[1], Javier de la Torre[2], Michel Louette[1]
[1] Royal Museum for Central Africa, [2] Museo Nacional de Ciencias Naturales

Several reviews and end-user needs assessments have shown there is a great interest in using the now-accessible geo-referenced natural sciences primary data for various purposes. In order to use these data in an efficient way, end-users need additional information and tools to assess the "fitness for use" of the available information.

Many of the geo-referenced specimen and observation data available have been collected during expeditions and surveys. Some itineraries which followed the collecting campaigns are well known for historical reasons (e.g., famous expeditions in the Belgian Congo or in Polar regions), but many others are only poorly known or documented. Such information is often hidden in field notebooks, which need to be digitised and analysed.

The purpose of task NA-D 3.7 of the SYNTHESYS project is to provide itinerary-related services (http://www.biocase.org/products/geo_services/itineraries/). The objective is to detect itinerary patterns in geo-referenced primary data presumably collected during a collecting event. A first validation approach is to use geo-referenced primary information from well-known itineraries, and to evaluate whether itineraries obtained from coordinates and collecting date correspond to what is known from the literature. This is done with several specially-selected datasets considered as complete and reliable.

In a second step, the defined algorithms are tested and applied to geo-referenced primary data available in the GBIF and BioCASE network in ABCD and DarwinCore formats, where the expedition routes are less documented or even completely unknown.

It is likely, depending on the accuracy of the available data, that several possible alternative expedition routes will be extrapolated. These routes and the related collecting points will be shown to the end-users on online maps using GIS services. These latter tasks will be done in close collaboration with SYNTHESYS NA 3.6 (Core GIS services). OGC Open Standards like WMS, WFS, WCS and GML and Open Source GIS software like the Deegree framework have been used in the implementation.

Support is acknowledged from: EU project SYNTHESYS (http://www.SYNTHESYS.info).

10.22. Using TAPIR in Biodiversity Networks

Markus Döring
Botanical Garden & Botanical Museum Berlin

TAPIR 1.0 is ready to be deployed. With at least one implementation (http://pywrapper.org) and others coming, projects building biodiversity information networks can now use TAPIR to set up their basic infrastructure. Existing networks, like BioCASE, or the Generation Challenge Program have begun to deploy TAPIR.

This presentation will discuss strategies for using TAPIR and will explain the different architecture components needed to build efficient networks. The presentation will focus on

TAPIR models and their role in creating specialized networks on top of widely agreed conceptual schemas. An updated roadmap of implementations will be presented to better help people organize and target their projects.

Support is acknowledged from: IPGRI; GBIF; SYNTHESYS

10.23. The Global Invasive Species Information Network / Socio-Technical issues in Invasive Species Data Exchange

Annie Simpson
National Biological Information Infrastructure

The Global Invasive Species Information Network (GISIN) requires building an information network for sharing and exchange of invasive species data, information, knowledge, and related metadata, for all organism types. The distributed network will use common standards, protocols and services, as many existing and new invasive species information systems as possible throughout the world. The Global Invasive Species Information Network (GISIN) was formed in April 2004 after several years of related meetings. Four goals of the 2004 meeting included:

- creation of an online community to support global collaboration,
- identification and agreement on common data elements for global database cross-searching and interoperability,
- creation of a proposal funding toolkit containing such things as example proposals, proposal-writing guidance, suggested funding sources & other related information, and
- a review and listing of existing online invasive and alien species (IAS) databases (http://www.gisinetwork.org/Documents/DRAFTIASDB.html)

The interim Steering Committee (http://www.gisinetwork.org/contact.html) formed at the Baltimore meeting worked with the Secretariat of the Convention on Biological Diversity (CBD) to commission the creation of an Invasive Alien Species Profile Schema (IAS-PS). Jerry Cooper (LandCare New Zealand) and Michael Browne (Invasive Species Specialist Group, Auckland, New Zealand) created the draft schema (http://invasivespecies.nbii.gov/ documents/CBB-report-to-CBD.pdf), which was posted for public comment in August 2005.

In February 2006, the CBD Secretariat, the Global Biodiversity Information Facility (GBIF), the government of Morocco and other organizations collaborated to convene an experts meeting to examine and refine the IAS-PS in detail and to consider reviewer comments. A progress report (http://www.biodiv.org/doc/meetings/cop/cop-08/information/cop-08-inf-35-en.pdf) was made to the CBD COP8. This meeting also built on recommendations made in the CBD information document UNEP/CBD/ COP/6/INF/18 (http://www.biodiv.org/doc/meetings/cop/cop-06/information/cop-06-inf-18-en.pdf), "Report of the joint convention on biological diversity/global invasive species programme informal meeting on formats, protocols and standards for improved exchange of biodiversity-related information," for the establishment of the GISIN as a pilot initiative.

This GISIN Symposium will discuss sharing invasive species information, including:

- "Defined Schemas"
- Simple thesaurus representations
- Overlapping concerns with SDD and Observations Interest Groups and
- Geopolitical concerns inhibiting data sharing.

Support is acknowledged from: US Geological Survey, National Biological Information Infrastructure

10.24. Improving Performance and Access to DiGIR Based Data for Applications Including Forecasting for Invasive Species Ranges

Jim Graham, Greg Newman, Catherine Jarnevich, Thomas Stohlgren
Colorado State University

Museums and herbaria have made progress in providing their data online by using the DiGIR and BioCASe protocols and the GBIF portal. These databases represent almost 100 million records from an estimated pool of over 2 billion. Other types of organizations are now looking to access this valuable dataset for analysis and modeling. Due to sequentially searching multiple servers, inconsistencies in Internet performance, and variances in the performance of the provider servers may take hours to extract all available data for a single species. As the number of servers and records increases, the time to search will continue to increase. Through performance analysis we have found that there is an exponential increase in the amount of time to harvest data from DiGIR providers as the number of records requested increases. This problem can only be addressed by changes to the provider software and the provider's databases. Once addressed, DiGIR harvesters will be able to harvest billions of records per month just as Google harvests billions of web pages each month. Harvesters will then allow users to search billions of biological data records in seconds and still link to the original provider database for more detailed information. The existing DiGIR protocol can be used for harvesting while control of the original data remains with the original provider. We have created a Global Organism Detection and Monitoring System (GODM) that provides land managers and scientists with access to data on invasive species. Data are uploaded directly to GODM by users in the field and will be harvested from other databases on the Internet. These data are available for users to download and to create predicted ranges for invasive species online.

Support is acknowledged from: United States Geological Survey (USGS), National Aeronautics and Space Administration (NASA)

10.25. Specify Software Project: Requirements, Design, Components and Support

Rod Spears, James Beach, Andrew Bentley, Jean Burgess, Kathy Coggins, C.J. Grady, Glenn Garneau, Meg Kumin, Tim Noble, Joshua Stewart
Biodiversity Research Center, Univ of Kansas

In a business sense, community-oriented, grant-funded, application software initiatives are often perceived to be neither fish nor fowl. The incertae sedis comes from the apparent internal contradiction of projects that offer software licenses or services at no cost to the user, but which are organized to provide high value in a manner characteristic of commercial, for-profit, software vendors.

Cyberinfrastructure projects in this hybrid space, which are expressly focused to meet the data processing requirements of a community of research users, are excellent conduits for the widespread realization of novel, standards-based, computational and communication capabilities.

We will describe our Specify Software Project engineering process for requirements analysis, user interface design, open source component choices, and our new platform architecture for Specify 6. The presentation will be within a context of integrating useful, standards-based, capabilities for the benefit of biodiversity collection database researchers.

Support is acknowledged from: US National Science Foundation, University of Kansas

10.26. Development of Information Technologies for Botanical Gardens of Russia

Alexei Prokhorov
Petrozavodsk State University Botanic Gardens

Information technology developments address both internal and external requirements of botanical gardens. Knowledge of collections enables us to effectively manage garden collections and to increase the value of special collections to increase public interest. Governments and society have requirements for biodiversity conservation through botanical gardens. Gardens provide access to genetic resources, systems for estimation, inventory and monitoring of ex situ (genetic) resources, and analyses of these resources.

This presentation reviews a project of estimation, inventory, monitoring and biosafety of genetic resources of vascular plants ex situ in Russia. This project was based on the plant database management system "Calypso" using the ITF-standard of TDWG, and the information-searching system "Botanical collections of Russia and adjacent states". The information-analytical system "Botanical collections of Russia" has been created for the comparative analysis of botanical collections.

We examined new methods for the analysis of a wide set of botanical collections to estimate the role of ecological factors in mobilization and conservation of the biodiversity of plants in botanical gardens. The analysis of collections of Russian botanical gardens includes: an estimation of a taxonomic diversity of the collections in relation to the world biodiversity of plants; an estimation of the influence of the key climatic factors on spatial distribution of genetic resources of vascular plants, and the development of strategies for the formation of a national collection of rare and endangered plants of Russia.

Information technologies used for databases, searching and analysis are creating new opportunities for coordinating botanical information between botanical gardens. New systems are providing a unique opportunity for the comparative analysis of each garden’s collections and helping to form individual collection policies that increase their uniqueness. For example, access to these data is improving the conservation of biodiversity by increasing the number of taxa to be kept ex situ based on the uniqueness of each collection.

The results of the analyses will be used by the Council of botanical gardens of Russia to help botanical gardens coordinate a) research on plant introductions under different environmental conditions, and b) the conservation and mobilization of the genetic resources of plants.

Support is acknowledged from: Russian educational agency

10.27. A Generic Data Import Layer for the Berlin Taxonomic Information Model

Anton Güntsch, Walter G. Berendsohn, Andreas Müller
BGBM Berlin-Dahlem

The Berlin Taxonomic Information Model is a relational information model based on the potential taxon concept (Berendsohn, 1995). The model incorporates nomenclatural rules and traditional taxonomic relationships (synonymies, taxonomic inclusions) and the capability of representing taxonomic concepts as name-reference pairs (Berendsohn & al., 2003). The additional inclusion of non-traditional set-theoretical concept-relations provides the means for accurate and transparent storage of concept graphs (Geoffroy & Güntsch, 2003). The model has been implemented as a Microsoft SQL-Server database together with a suite of application programs such as a taxonomic web-editor, WWW publication software, and various parser programs. Berlin Model users range from taxonomists writing monographs to international checklist projects.

Experience from the existing Berlin model application projects suggests that data imports consume a substantial share of project resources. This is mainly due to the heterogeneous structure of available taxonomic data and the complexity of the target model.

A generic data import method using two XML schema layers and three phases of transformation flow between a data source and the target Berlin model database aids importation. In the first phase, importers transform the source data into data valid against a "soft schema" that best fits the semantics of elements in their source. Users may choose from a comprehensive Java library of transformation tools. If an appropriate soft schema does not exist, it is possible to use a new one (e.g. a new version of TCS).

"Soft schema" data are then transformed by defined rules (including atomizing and restructuring) to the final "strict schema" representing a fixed definition of elements and structures for taxonomic data sets. Like the Berlin model, this schema is capable of representing concepts and arbitrary relations but it hides the complexity of the database model from the user. Malformed source data are highlighted and may be corrected during the semi-automatic transformation from the "soft schema" to the "strict schema" (phase 2).

An automated phase 3 consists of duplicate detection and an object-relational data transformation.

The method has been used successfully in the course of the Med-Checklist project which imported Vol. I, III, and IV into a Berlin model database from heterogeneous sources (http://ww2.bgbm.org/mcl/home.asp). Further importing tasks for the EU project EDIT, for the IOPI Species Plantarum initiative, and for the Euro+Med project will be used to refine the scheme.

Support is acknowledged from: European Union, German Federal Agency for Nature Conservation

10.28. System Architecture of the Avian Knowledge Network

Tim Levatich, Steve Kelling
Cornell Lab of Ornithology

Providing informative biodiversity resources to a broad spectrum of users is more than simply providing access to raw data. Interpretations of these data via data visualization or analysis are arguably far more useful for land managers, policy makers, or educators to make informative use of these data.

This presentation will be an overview of the system architecture of the Avian Knowledge Network (AKN), a data federation, archiving, data source, and analysis system for observational data gathered on bird populations. The presentation will show how we: 1) manage the federation of these data utilizing the existing biodiversity informatics infrastructure; 2) ensure a persistent archive; 3) implement access control measures, including how they are expressed in data usage; 4) provide data access to a network of organizations; and 5) briefly describe methods of data integration to develop a suite of data visualizations (including maps, graphs, tables), and interactive exploratory analysis via new techniques in data mining and hierarchical statistics.

Support is acknowledged from: NSF-IIS 0612031, NSF-DBI 0542868, NSF-EF 0409378

10.29. The New Norwegian National Thesaurus of Species Names

Stein Alexander Olsen, Christian-Emil Ore
University of Oslo

The Norwegian Species Information Centre (http://www.artsdatabanken.no) was founded in 2004 by the Norwegian government to serve as the focal point for information on threatened species and general biodiversity in Norway. The Centre has close contacts with the natural history museums in Norway and operates in close connection with the Norwegian GBIF node.

The pivot in the Centre's information system is the species/taxon name thesaurus database, currently under development. The thesaurus is meant to be a tool to help the general public to find information about species in Norway, a name authority register for natural history collections and a tool for the (governmental) management of the environment of Norway. The thesaurus will contain all names used in natural history collections, even non-valid or unpublished names. The thesaurus will be connected to international authority registers like Fauna Europaea (http://www.faunaeur.org), and the content will be maintained by expert groups.

The Museum Project, 1998-2006, (http://www.muspro.uio.no/engelsk-omM.shtml) is a co-operative digitisation and database project of the Norwegian University Museums (natural and cultural history). The Museum Project participated in the specification and modeling of the new national taxon name thesaurus. The core ICT-group of the project has been participating in the development of the ICOM-CIDOC's Conceptual Reference Model (ISO 21127, http://cidoc.ics.forth.gr), an event oriented model/ontology covering natural and cultural history. This model was presented at TDWG 2005 in St.Petersburg (see http://www.edd.uio.no/nedlasting/foredrag/tdwg2005_Lampe_Ore_final.ppt).

The specification and modeling work was done during the first half of 2006 by an expert group comprising biologists and information specialists. The specification process was an interesting meeting between the working systematic biologists with a more common-sense understanding of taxon and taxon names and the information scientists with an understanding of the same concepts based on the pure ontological model CIDOC-CRM.

The resulting model is not as complete as originally suggested by the Museum Project but is an event oriented model compliant with CIDOC-CRM. Although the original intention of the model is to specify a database for Latin scientific species names, the use of abstract concepts like type and taxon makes it possible to differentiate between nomenclatural and taxonomic synonymies. The central role of events in the model makes it easy to express the history of names and which taxon they denote. The standard operations in systematic botany and zoology like splitting, joining, (re)defining and (re)naming of taxa are all expressed as events with corresponding actors (author), dates, abstract definitions and written descriptions. The model makes it easy to see what a certain name denotes at a certain time.

10.30. Federating Taxonomic Databases: Progress with the Catalogue of Life Dynamic Checklist

Richard J. White[1], Andrew C. Jones[1], Frank A. Bisby[2]
[1] Cardiff University, [2] University of Reading

The Spice system federates species databases in order to aggregate taxonomic coverage. Species 2000 and its Catalogue of Life partner are creating a global Dynamic Checklist for all organisms on the Internet. This checklist presently has more than half a million species from 37 interlinked databases. The process of federating species databases to broaden taxonomic coverage and present a holistic perspective is a key element in many other large-scale information systems. For example, almost every continent has a programme or plan to aggregate databases to assemble a regional-scale species checklist. Other programmes are

working to assemble coverage for marine and freshwater organisms.

The Spice aggregation software implements a hub or Common Access System to address a distributed array of species databases. The system uses the Spice Protocol to manage the exchange of data complying with the Species 2000 Common Data Model and the Spice XML Schema (documented at http://www.sp2000.org/tech/). Depending on the settings, Spice can query the databases directly or use data harvested into a central cache to provide users with faster and more reliable responses. Spice dynamically threads together a combined taxonomic hierarchy using hierarchy branches from each supplier. Each provider database is linked to Spice over the Internet using a wrapper program to receive Spice protocol requests, query the database and return responses using the XML Schema. Spice provides a managers' test interface, and a variety of user interfaces. Other software and information systems can address the system through the system's Web Services.

Species 2000, built on Spice version 5, assembles and delivers the Catalogue of Life Dynamic Checklist. This is a dual implementation in which two Spice hubs are working together: one for the Global Checklist using global species databases, and the other for the Pan-European Species Checklist composed of the regional databases Fauna Europaea, ERMS, and Euro+Med PlantBase. The user interface (http://spice.sp2000.org) provides access to the individual hubs, and combined access to both hubs. The checklist data for each species is found by searching on common or scientific names, including synonyms, or by browsing the taxonomic tree.

Depending on which databases and hubs are connected, a Spice system may encounter multiple representations of the same species and alternative taxonomies. Species 2000 has experimented with the Litchi version 2 software to investigate taxonomically intelligent linkage between hubs using different taxonomies. Automatically created cross-maps may for example, connect a broad-concept species to two corresponding narrow-concept species, where at least one has a different name to that used in the original search.

Spice is open-source software released by the Spice Software Consortium, and both the Consortium and Species 2000 are open programmes. We welcome further collaboration with the taxonomic and biodiversity information communities.

Support is acknowledged from: BBSRC (UK), European Commission, etc.

10.31. The Transition to Taxon Concepts in a World of Legacy Data

Robert K. Peet[1], Alan S Weakley[1], Xianhua Liu[2], Nico Franz[3]
[1] University of North Carolina, [2] NESCent, [3] University of Puerto Rico

Application of taxon concepts has the potential to greatly improve integration of biological data collected at different times and places by different investigators following different taxonomic treatments. However, transition to concept-based data integration presents many challenges, including populating databases with relationships among concepts. Another challenge is integrating datasets where some components refer to taxon concepts and others do not. We have developed protocols for mapping relationships among concepts from multiple treatments. These mappings may be employed in cases where only some datasets have taxa documented using concepts while others must have their taxa treated only as nominal concepts. We use the flora of the Southeastern United States as a case study to demonstrate our approach. We have documented relationships among concepts for some 6300 taxa treated in 11 major floras and multiple narrow treatments. The resulting database of taxon relationships contains approximately 100,000 entries. Our approach to data integration is demonstrated with a new floristic atlas for the Southeastern US flora that integrates records based on concepts (e.g., local floras with range maps) with records documented only with a name (e.g. specimen databases).

Support is acknowledged from: NSF ITR-0225635 to KU, NSF DBI-0213794 UNC-CH

10.32. Invasive Alien Species (IAS): Terminology

Michael Thomas Browne
IUCN Invasive Species Specialist Group

Sharing information about invasive alien species (IAS) is challenging because no generally accepted terminologies and associated definitions are available. Policy related definitions of the word 'invasive' tend to emphasise the harm caused by introduced organisms to biodiversity (and sometimes to economies and human health), while more explicit scientific definitions focus on the process of establishment and spread.

The terminology working group of the Global Invasive Species Information Network has proposed a matrix of atomised terms to manage different uses of the word 'invasive'. Experience gained in compiling information for the Global Invasive Species Database (http://www.issg.org/database) since 2000 has been applied to the development of preliminary high level standard terminologies for habitats, introduction pathways and vectors, impacts and management. The Global Invasive Species Database will provide an initial example of mapping between local terminology and the standardised terminology.

Support is acknowledged from: GBIF, IUCN Invasive Species Specialist Group, NBII/USGS, Centre for Biological Information Technology at the University of Queensland, Colorado State University, UMass-Boston, Manaaki Whenua-Landcare Research New Zealand

10.33. PlantCollections

Boyce Tankersley
Chicago Botanic Garden

PlantCollections™ - A Community Solution, is an Institute of Museum and Library Services National Leadership grant in the Building Digital Resources category. PlantCollections accesses the data found in plant records databases of botanic gardens and arboreta through distributed queries. The collaborative is led by the Chicago Botanic Garden, the University of Kansas Biodiversity Research Center and Natural History Museum and the North American Plant Collections Consortium of the American Public Gardens Association. Sixteen botanic gardens have agreed to participate in 2 phases to develop, test and implement an open source software application utilizing WASABI for the portal, possibly GoogleEarth for maps and MorphBank for images. The intended data users were surveyed and feedback from curators, taxonomists, educators, horticulturists, ecologists, weed scientists, conservation scientists and gardeners defined the 161 fields found in the federated schema. Deliverables of the project are a federated schema, improved Website for the APGA, development of software applications, servers for each institution and training of staff at each institution.

Success will benefit collaborative research into complex biodiversity phenomena, educate the next generation of plant scientists, improve collections and advance technology.

11. Posters

11.1. Georeferencing Specimens by Combining Expedition Maps with Landsat 7, JERS-1 SAR and SRTM Satellite Imagery

Niels Raes, Johan B Mols, Luc Willemse, Erik Smets
National Herbarium of the Netherlands

In the last decade, many herbaria and Natural History Museums collections have been digitized and are available on the Internet. These data are used for understanding ecological and evolutionary determinants of spatial patterns of biodiversity, conservation planning, identification of 'hotspots' of biodiversity, forecast of the effect of habitat change and global warming, establishing potential locations for species reintroduction, and to predict the likelihood of invasion of exotic species. Specimens need to be georeferenced if they are to be useful to these applications. Many old collections only have collection site descriptions, many of which cannot be found in online digital gazetteers. Detailed expedition maps of specimen collection locations are however available.

At the National Herbarium of the Netherlands, we digitized and georegistered these maps by matching rivers, coastlines and mountains with high resolution Landsat 7, JERS-1 SAR (radar) satellite imagery and SRTM Space Shuttle Digital Elevation Data using the Manifold GIS package. We allowed Manifold to transform the expedition maps during the georegistration process to overcome navigational errors made during the expeditions. Once the expedition maps were georegistered, they were overlaid with the satellite images.

This procedure allowed us to identify and georeference most collection localities and substantially increase the number of georeferenced collections. Most of the time we used Landsat 7 images, in case of cloud cover we switched to JERS-1 SAR radar images, and for mountains and hilltops, we used SRTM Space Shuttle Digital Elevation Data.

11.2. Benefits of OGC Compliant Standards and Tools for Biogeography Related Information Sharing

Patricia Mergen, Bart Meganck, Danny Meirte, Franck Theeten, An Tombeur, Michel Louette
Royal Museum for Central Africa

End-user and stakeholders surveys have identified easy access to geographic information and distribution maps in re-usable formats as a major need in biodiversity conservation. According to GBIF, around 75% of the current (July 2006) 97 million records connected to the system are provided with geographic coordinates. In order to use these data efficiently, users need additional information and tools to assess the "fitness for use" of the available information in form of primary data. The Royal Museum of Central Africa has as goal to provide tools and services for integration, visualization and quality checking of biodiversity data. Care is taken to keep these developments compliant with both TDWG and OGC standards.

RMCA is involved in the GBIF Seed Money awarded project HerpNET (http://www.herpnet.org), in which around 200,000 georeferenced and checked amphibian records from Sub-Saharan Africa will be made available through GBIF.

This poster will also illustrate how the OGC compliant Deegree Java Framework (suitable for a complete Spatial Data Infrastructure - http://www.deegree.org), has been used to display the itineraries followed during scientific sampling expeditions (SYNTHESYS Network Activity D project, http://www.biocase.org/products/geo_services/itineraries).

It is envisaged that the developed services will be made available to the Cybertaxonomy

Platform (EU project EDIT, http://www.e-taxonomy.eu) as additional tools for taxonomists wishing to use species and specimen distribution maps and related geographic information for online taxonomic revisions.

Support is acknowledged from: The authors wish to thank all the colleagues from SYNTHESYS, EDIT and Herpnet in charge with or actively collaborating to these various projects as well as also the developers of the Deegree and associated tools for their enthusiastic support.

11.3. The Global Invasive Species Information Network

Elizabeth Sellers, Annie Simpson
National Biological Information Infrastructure

The Global Invasive Species Information Network (GISIN) requires the building of an information network for the sharing and exchange of invasive species data, information, knowledge, and related metadata, for all organism types. The network aims to use a distributed approach to connect as many existing and new invasive species information systems as possible. GISIN will promote and use common standards, protocols, and services designed to achieve connectivity.

The poster will address current aspects of the development of the GISIN http://www.gisinetwork.org and the Invasive Alien Species Profile Schema http://wiki.cs.umb.edu/twiki/bin/view/IASPS/IASPSSchemaAlphabetical.

Support is acknowledged from: US Geological Survey, National Biological Information Infrastructure

11.4. A New Model for Descriptive Knowledge

Antoine Chalubert, Régine Vignes Lebbe
Université Paris VI - France

The description of biological entities, such as species and other taxonomic groups is the base of most biological knowledge. To be able to compare taxonomic descriptions is essential to analyze, classify and identify. A formal representation of descriptive knowledge is needed in order to justify the relevance of a particular algorithm. In general, this justification is provided by the implemented method itself, for example, a matrix of taxa by characters for phylogenetic analysis. However, these methods do not provide an explicit and complete knowledge representation that can express all the meanings of "character" that can be found in systematic literature. Another consequence of these partial representations is the impossibility of integrating or combining various methods using the same knowledge base (e.g. identification and phylogenetic analysis).

Our aim is the development of an extensive data and knowledge-processing platform for systematics, integrating taxonomy as well as identification and phylogenetics. Our proposal is the continuation of previous works on knowledge base editor and computer-aided identification (KB-CAI) computer software like XPER, NEMISYS, DELTA, IKBS and the proposals of the working group SDD (Structure of Descriptive Data). All recent results offer a limited formalism, restricted to treatments of descriptive data: they deal with the descriptions of the properties of objects but cannot handle knowledge related to organisms such as their structural and anatomical description (see Pullan et al., The Prometheus Description Model: an examination of the taxonomic description-building process and its representation). They can manage the polymorphism of objects but do not allow any estimation of the reliability of the data or its traceability. They have reduced extension possibilities into different character types than initially considered and do not allow the descriptions to be modified for example, because

a new type of data is available. They do not provide any complementary contextual information (such as the required proficiency, the conditions of observation). Or their methods do not allow the conversion from one type of information to another (e.g. numerical to qualitative).

We propose a new model for descriptive data that tries to address these failings. We show that our model allows innovative representation of concepts and treatments, an extensive pool of state character types, the representation of complex anatomical descriptions, phylogenetic analysis and control of the reliability of data and user proficiency level. Our proposal is partially implemented in the computer program KB-CAI. This software is being improved to build a complete framework for computer-aided systematics.

11.5. TDWG and the European Distributed Institute of Taxonomy

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

The European Distributed Institute for Taxonomy (EDIT) is a Network of Excellence (NoE) project in the 6th Framework Programme of the European Commission. The project is led by the Muséum National d'Histoire Naturelle in Paris and unites 23 major Natural History institutions and herbaria in Europe with 2 USA institutions in an attempt to integrate their resources to make taxonomic research more efficient.

EDIT's primary goal over its five year project period is to transform taxonomy into an integrated science by strengthening the technological and human resources of the participating institutions. Various collaborative efforts will be initiated at a technical and administrative level and research. One of the focal points of the project is the integration of IT departments and the creation of an "Internet Platform for Cybertaxonomy", for which strong links with TDWG's standardisation efforts are needed. The poster will inform participants in the TDWG meeting about the structure, membership and contact points for EDIT, a project which intends to create a durable and extendible infrastructure for taxonomic research.

Support is acknowledged from: European Commission

11.6. DarwinCoPE, a Proposed Paleontological Extension to DarwinCore 2

Jessica Theodor
University of Calgary

DarwinCoPE (DarwinCore Paleontology Extension, http://darwincope.museum.state.il.us/) is a proposed draft extension of the DarwinCore 2 XML schema (http://darwincore.calacademy.org), to include the specialized data for geologic time and rock units needed to search fossil collections using distributed databases.

DarwinCoPE was developed at an NSF-sponsored Paleontology Collections Databases meeting held at the Illinois State Museum in May 2005 as a draft for a community standard. It is very similar to the schema already in use by the PaleoPortal project and is compatible with the more detailed proposed European schema, ABCDEFG, used in GeoCASE (http://projects.naturkundemuseum-berlin.de/synthesys_activity_d/). DarwinCoPE has been proposed to the Taxonomic Databases Working Group (http://www.tdwg.org) as a draft standard extension to DarwinCore 2. As a standard extension to DarwinCore 2 DarwinCoPE would allow collections database developers a standard interface that would allow collections managers to make their collections data more easily and widely available over the Web.

DarwinCoPE includes basic fields for geologic time units, biostratigraphic zonations, and lithostratigraphic units, which, when combined with fields from DarwinCore 2 and the Geospatial and Curatorial extensions, should allow more widespread adoption of the TAPIR protocol for creating distributed databases among paleontological collections. A working

demonstration using a version of the PaleoPortal provider software adapted to use the DarwinCoPE schema is available at http://darwincope.museum.state.il.us/portal/index.php, using data from the Illinois State Museum, the University of California Museum of Paleontology, and the Sam Noble Oklahoma Museum of Natural History.

Support is acknowledged from: National Science Foundation

11.7. Introducing 'mx', a Sharable Digital Workbench for Systematic Biologists

Matthew Yoder[1], Krishna Dole, Andrew R Deans[2],
[1] Dept. of Entomology, TAMU, [2] School of Computational Science, Florida State University

'Mx' (short for "matrix") is a new, web-based, open-source application for use by systematic biologists. Mx is built using the object-relational framework provided by Ruby on Rails, and uses MySQL and AJAX technologies. Mx manages a wide range of data including taxonomic names, descriptions, images, morphological characters and matrices, biological associations, ontologies, references, specimens and collecting events, multiple entry and traditional keys and more. Multiple projects may be created each with multiple users enabling long distance collaborations (e.g. compiling taxonomic catalogs or scoring morphological matrices). One of the primary goals of mx is to promote data capture during the research process. By centralizing the workbench to a web application, dissemination of a revision or monograph (the end product) to both print and web formats will be greatly simplified.

The central object in mx's database schema is the Operational Taxonomic Unit (OTU). Most data are tied to OTUs rather than taxonomic names. This abstraction allows for data to be captured during the research process while taxonomic entities are not formally named (or even circumscribed). This feature is particularly useful for morphospecies-based biodiversity studies. A flexible content system allows for any number of labeled categories (text fields) to be defined and tied to a given OTU. This system allows for traditional text-based descriptions to be created, edited and compared within the application. The use of OTUs further allows for different concepts of the same taxa (i.e., multiple incompatible usages of a single taxonomic name) to be recorded and related. Objects (i.e., any record with a unique ID) in mx may all be tagged with keywords, references, and figures. The combination of the OTU as a central object, definable content types, and tags provides a flexible system that may be adapted to the specific needs of the user or team of users.

The database schema for mx is sufficiently parsed such that export to a number of standards should be easily accomplished. The Ruby on Rails framework allows for data to be readily provided in a variety of formats (e.g. HTML, XML) with very little modification to existing code. Various output types have already been developed including Nexus, tnt, and the ITIS taxonomic names format. A single instance of mx is presently serving data to 3 separate public websites (two taxonomic catalogs and a hosts/parasites database) and is being used by 4 labs in three countries. These projects will be highlighted in the poster.

11.8. The National Biodiversity Information System of Korea

Sangyong Kim, Seung Sun Jung
Korea National Arboretum, Pocheon

The objectives of Korea National Biodiversity Information System (NaBIS, http://www.nature.go.kr/) are to provide a service to the public and to support research and industry by providing online linked text, image and specimen data. In August 2006, the database contained information about 4,309 plant taxa, 531,313 specimens from 27 institutes, and 46,274 plant records from 15 arboreta. There are images and text of 5,700 insect taxa, 333,470 specimen records from 22 institutes. And there are images and text data of 1,002 fungi taxa. Biology for children is included and organized by illustrative information.

Scientific names and Korean names in the system are linked to the Korean Plant Names Index (KPNI, http://www.koreaplants.go.kr:9090/english/); a database of the names and bibliographical details of all Korean vascular plants. KPNI includes new plant names through careful discussions and a committee confirmation. KPNI is the product of collaboration between Korea National Arboretum (KNA), the Plant Taxonomic Society of Korea, the National Plant List Committee and the Cultivated Plant List Committee. KPNI is a tool for botanists who work with current and prior plant names. KPNI includes information on names in current use, plant name changes, previous names for a renamed plant, the journal or place where the name was formally published, the author of the plant name, additional references and relevant comments and notes on the naming process.

The major function of NaBIS is to link information from plant and insect resources. The system shows text and image data in one window and is extended by linking images and text to specimen data. Morphological characters of each taxon are built into the system enabling advanced search by shape and color (flowers, leaf, stem, fruits, etc.) without knowledge of taxa names.

The web platform is composed of IBM AIX 5 (OS), oracle oc4j (web server application), JAVA / JSP (jdk 1.4.2; language) and Oracle 9i DBMS. For global data sharing, NaBIS is being improved according to a GBIF schema based on DarwinCore and the DiGIR Provider will be applied as a sharing tool.

Support is acknowledged from: Korea National Arboretum

11.9. Prototyping a Generic Slice Generation System for the GBIF Index

Jörg Holetschek, Anton Güntsch, Cristian Oancea, Markus Döring, Walter G. Berendsohn
BGBM Berlin-Dahlem

The GBIF index is maintained by the GBIF secretariat in Copenhagen. It contains a list of all specimens and observations registered within the GBIF network together with some data items considered most relevant for searches and output, such as taxon name, gathering/observation date and site geography. These data currently (Aug 30th, 2006) derive from 804 collections around the world and are harvested by the GBIF indexer using the Darwin Core and ABCD data schemas. In the process, the data are decomposed and stored in the highly normalized data model of the index.

The EU-funded SYNTHESYS project and the development of the German GBIF Node have included efforts to set up specialized search portals for biodiversity data. As a first step, a prototype system has been set up in association with one of the mirrors of the GBIF index. This system creates subsets of the GBIF index that could be used as the base for the search portals of special interest networks or regional organizations. This offers an opportunity to these groups to draw on their resources to enhance the usability of the data in the GBIF system, for example, by adding additional information provided by other data sources such as regional or group-specific taxonomic thesauri, local geographic services, or translation mechanisms.

The process of the geographic slice generation comprises three stages:

1. Filtering data from the GBIF index using different criteria (taxa, country codes, geographic coordinates, regional place or area names, collection metadata);
2. Transforming the data into a query-optimized data model and
3. Processing data in order to enhance data quality (optional) and/or augmenting data with additional information (optional)

At the moment slices are updated regularly during the night (01:00 GMT).

As an example, SYNTHESYS is implementing the BioCASE search portal for European biodiversity data. This will ultimately be integrated with European taxonomic backbone systems (Fauna Europaea and Euro+Med PlantBase) as well as with the evolving European geographic data infrastructure. In parallel, GBIF-D Botany is prototyping a search portal for botanical data that will be linked to the standard lists of plants available for the German flora. For these two projects, slicing can be performed by applying filters on geography and taxon information, respectively. We suggest that basic slice generation based on geographic criteria (e.g. for countries) could be among the services offered by the new GBIF index system. At the prototyping stage, rules must be specified as SQL statements, which permits slicing rules based on all fields contained in the index database.

The system is still in a prototype stage, but a slice system is being tested with the SYNTHESYS user interface currently under construction (http://search.biocase.org).

The major challenges with such a system are not technical, but relate to GBIF's obligations to the data providers. Before such a system can be deployed more widely, it is essential to ensure that sliced data are kept current as providers make modifications and corrections, all data providers are fully and appropriately acknowledged for their contributions, and data providers are kept informed on uses to which their data are put. We will be working with the GBIF Secretariat to address these issues and also to accommodate changes arising from the current redesign of the GBIF index system.

11.10. Collaborative Georeferencing Using WASABI and GEOLocate

Nelson Rios, Henry L. Bart
Tulane University Museum of Natural History

The number of biological specimens in museums and herbaria worldwide is estimated to exceed 2.5 billion. Revived interest in these collections and biodiversity informatics has brought about many efforts to digitize museum records. Sadly, much of this data lack geographic coordinates, so vital to our utilization of this vast information resource in large-scale studies. Recent developments in automated georeferencing tools have greatly facilitated the task of generating geographic coordinates from textual locality descriptions, yet a bottleneck still exists whereby users must manually verify each record. Using DiGIR/WASABI and GEOLocate, we are developing a framework for collaborative georeferencing that will reduce verification effort. Users will be able to import data from DiGIR providers as well as return the corrected data back to the provider. Similarity relationships among network-wide records will be used to identify records that describe the same place, but only need to be corrected once. The distributed fish collection database network (FishNet) is being used as a test case for implementation. Support for this project is provided by the U.S. National Science Foundation.

Support is acknowledged from: U.S. National Science Foundation

11.11. NatureServe Vista: A GIS-Based Decision Support System for Conservation Planning

Kristin Barker, Bruce A. Stein
NatureServe

NatureServe Vista is a decision support system (DSS) that integrates conservation information with land use patterns and policies. It provides planners, resource managers, and communities with tools to help conserve and manage natural resources.

Version 1.3 of the system, released in March 2006, is built as an extension to desktop ArcView

9.1 in C#. The application allows users to assemble a geospatial database of conservation targets, such as species, habitats, vegetation types, or other spatially defined features (e.g., historic sites, viewsheds). The system facilitates mapping of each element's distribution over the planning region, allowing users to depict the condition (viability) and locational certainty of each occurrence. Biodiversity elements can then be assigned explicit and quantitative conservation goals that reflect stakeholder values. Users can calculate and visualize an aggregate "conservation value" across the planning region as a function of both scientific attributes and stakeholder values. Stakeholder goals can then be evaluated against existing conditions and proposed plans, with detailed maps providing users with a visualization of potential biodiversity/land-use conflicts. Interactive tools allow users to understand the biodiversity composition, land use compatibility, and stakeholder valuation of individual parcels.

Vista also integrates with MARXAN, a popular conservation analysis and site portfolio optimization tool. Details about NatureServe Vista 1.3 can be found at: http://www.natureserve.org/prodServices/vista/overview.jsp. Development of NatureServe Vista version 2.0 is now underway, and will feature integration with NatureServe's standardized XML Web Services (http://services.natureserve.org/).

Support is acknowledged from: Doris Duke Charitable Foundation

11.12. Variable-Level Nomenclators

Arturo H. Ariño
University of Navarra

Current taxonomic databases can belong to roughly three categories: first, maintained catalogues (e.g., Sp2K, ITIS), which are intended to be taxonomically consistent and authoritative, ideally with taxon concepts adequately linked to the relevant source, and having a relevant hierarchy. Second, species lists, which can be merely listings of taxonomic names found in publications, surveys or museum collections, and which may or may not be taxonomically accurate. And third, there are nomenclators. These have been called to fulfil an intermediate role, with taxon names (but not necessarily concepts) assigned to some high-level taxonomic group that allows for placement, spell checking, disambiguation, or validation.

Nomenclators are very valuable adjuncts to the digitisation of museum collections, as a reference source for checking names found in such collections. For such a role, nomenclators should include alternate (mis)spellings and synonyms. Thus, nomenclators should bridge taxonomic catalogues and names lists, allowing curators to adequately place specimens within a taxonomy. Recent prototypes allowing for multiple taxonomies (e.g., Prometheus) are of particular value for this task. However, nomenclators, or taxonomic databases acting as nomenclators, frequently lack some of the features that would enhance their usability for management or digitisation of collections. A simplified taxonomic code system, flexible enough to be started from simple, unqualified species lists, but able to be progressively converted into a full taxonomic tree (including synonyms), could add much taxonomic functionality to nomenclators, without the burden (or necessity) of being completely accurate and up-to-date. Such a system can be of use when, in the process of digitisation, there is lack of readily available specialists, and tentative taxonomies (or taxonomies based on somewhat outdated literature) are needed for organisation purposes.

We have been dealing with such problems when digitising a series of zoological collections at the Museum of Zoology of the University of Navarra. For the past 25 years we have been building a nomenclator that is populated from literature records, sampling survey results and museum specimens. Taxon names are entered verbatim and later assigned to a taxonomy through the use of taxonomic codes in a self-referencing manner. This vintage system, now slated for replacement continues to be useful by allowing specialists to directly specify to which

level each taxon can be assigned, according to the information available, thus creating a variable-level taxonomic nomenclator. Full, consistent taxonomic trees can be constructed from the table once the information is complete for a taxonomic group, while the same table can be used as nomenclator or species lists in the meantime. Specimen occurrences are related directly to the nomenclator. In this poster I describe this variable-level nomenclator, common shortcomings, and identify potential issues that may arise when moving to a newer system.

11.13. CATE - Creating a Taxonomic e-Science

Benjamin Clark[1], Malcolm Scoble[2], C. Godfray[3], Ian Kitching[2], S. Mayo[4]
[1] Imperial College London, [2] Natural History Museum, London,
[3] Imperial College, London, [4] Royal Botanic Gardens, Kew

CATE is an acronym for Creating a Taxonomic e-Science and is a project funded by the United Kingdom's Natural Environment Research Council (NERC) under its e-science initiative. The particular goal of CATE is to test the feasibility of creating a web-based, consensus taxonomy using two model groups, one from the plant and the other from the animal kingdom. The wider aim is to explore practically the idea of unitary taxonomy and promote web-based revisions as a source of authoritative information about groups of organisms for specialist and non-specialist users.

11.14. Fonoteca Zoologica (www.FonoZoo.com): The Web-Based Animal Sound Library of the Museo Nacional de Ciencias Naturales

Rafael Marquez[1], Gema Solís[1], Xavier Eekhout[2], Laura González[1], Mercedes Pérez[1]
[1] Museo Nacional de Ciencias Naturales, Madrid, [2] Museo Nacional de Ciencias Naturales. Madrid

We describe the functional structure of Fonoteca Zoológica (FZ) and introduce its website FonoZoo.com. The animal sounds from FZ are separated into two different collections depending on their origin: the FZ Sound Collection and the Published Sound Collection. The FZ Sound Collection includes recordings made by researchers in the museum and other collaborators and the Published Sound Collection includes commercially available animal sound guides published all over the world. Here we present statistical data on the number of species of anurans (frogs and toads) included in both collections and emphasize the usefulness of the FZ in the study of anurans.

11.15. Content Management System for Biodiversity Data Application – Experience in Taiwan

Hsin-Hui Wu, Kun-Chi Lai, Eric Yen, Alan Yong, Hsin-Yu Chen,
Kwang-Tsao Shao, Ching-I Peng
Academia Sinica, Taipei

Biodiversity Informatics is an emerging field providing integrated services of distributed multi-model, multi-type and multi-disciplinary content resources, and fostering new research paradigms and new knowledge from them. A new infrastructure supporting better data collection, analysis, query and access, management, resource discovery, and dissemination is necessary to meet the requirements of biodiversity researchers, data curators and museums. The Taiwan biodiversity information facility (TaiBIF) is an organization which collects and integrates biodiversity data of different institutes in Taiwan. Based on the experiences of TaiBIF and related works, a web-based content management system (CMS) would be the most viable solution to reach the goals stated above. TaiBIF fabrics will be integrated into the Taiwan e-Science infrastructure in the near future and long-term preservation services will also be deployed. Content metadata, collection metadata, and resource directory are the entirely indispensable data framework for the CMS. Federated search services and unified access interfaces are the major system services in terms of users. By integrating AJAX technologies and Google Maps APIs, user experiences show that it can lower the efforts of data management and exploration tremendously. Moreover, internationally recognised metadata schemas, such as

Dublin Core, are adopted to share data with other biodiversity applications. This paper describes the structure of this CMS and the essence of the related technology. In the future, we will further improve and enhance functions of this CMS and develop biodiversity research process services under a Grid Computing environment.

Support is acknowledged from: Academia Sinica, Taiwan

11.16. UNIBIO: Integrating Biodiversity Information Using Public and Institutional Archives

Joaquin Gimenez-Heau
Universidad Nacional Autonoma de Mexico

The Biodiversity Informatics Unit (UNIBIO) is a unit of the Biology Institute of the National Autonomous University of Mexico (UNAM), founded in 2005. UNBIO's goal is to create a system to digitize, analyze and share two centuries of the biological information that the University has stored principally in the Mexican National Biological Collections.

UNIBIO has developed tools for the curators and taxonomist to standardize, capture and digitize all data from their collections. The tools migrate and share their information using the Darwin Core standard and the Distributed Generic Information Retrieval (DiGIR) protocol. We also have created an Institutional Repository using the Dublin Core Standard and the Open Archives Initiative Metadata Harvesting Protocol (OAI-MPH) to manage the entire digital objects associated with the specimens of biological collections. This Institutional Repository is connected to our DiGIR Portal.

The biological data bases can be searched and the results can be used to query the UNAM Institutional Repository to find images or articles related to the same specimen or species genus and family. The digital objects in this Institutional Repository can be also consulted from any OAI portal, such as OAIster or DSpace.

We have developed separate Internet interfaces using Java ServerPages to capture data from collections not already in a database. These data go to a PostgresSQL database and eventually into the DiGIR Data Provider using the Darwin Core Standard. We ensure that all data are supervised by taxonomists and other biological specialists.

We have developed tools in collaboration with CONABIO (National Biodiversity Council of Mexico) and the University of Kansas to predict species niche distributions and extinction rates using GARP through Web Services. We also have developed data mining techniques to predict species distributions from biological and environmental factors.

UNIBIO is a prototype node of a network of 17 institutions of the UNAM called the Informatics System for Biodiversity and Ambient (SIBA). The objective of this network is to store and share primary information related with biodiversity and environment data, and to promote multidisciplinary connectivity and research among these institutions.

Support is acknowledged from: National Autonomous University of Mexico (UNAM)

11.17. Species Checklist Database and Capacity Building Training in Bangladesh

Badrul Amin Bhuiya[1], Mohammad Shawkat Hossain[2]
[1] Chittagong University, [2] Biodiversity Research Group of Bangladesh (BRGB)

Although Bangladesh is known to be very rich in biodiversity, this wealth is rapidly dwindling for many reasons. Taxonomic impediments have been identified as the main hindrances for

conservation and sustainable use of Bangladesh's biodiversity. Lack of trained taxonomists, lack of awareness among users of biodiversity data, and unavailability of complete information are also considered impediments. With a view to coordinating taxonomic research and improving biological collection infrastructure so that reliable information on biological diversity is available to all branches of science in Bangladesh, a project was initiated by the Biodiversity Research Group of Bangladesh (BRGB) in 2001. Checklists for fauna and flora under 8 different groups (viz.: Protozoa, Fungi, Algae, Microbes, Non-vascular Plants, Vascular Plants, Invertebrates and Vertebrates), are being compiled. User-friendly software has been developed by a BRGB taxonomist for data entry by BRGB members from published information. As the National Coordinating Institute (NACI) within the South Asian Network for Taxonomy Capacity Building (SACNET) Bangladesh, BRGB will make biological information in this database available to all via the web. Training courses are being organized for taxonomic capacity building.

Support is acknowledged from: Ministry of Science & ICT

12. Computer Demonstration

12.1. The Flora of California: Demonstration of Digital Innovations at the Jepson Flora Project

Christopher A. Meacham, Bruce G. Baldwin, Jeffrey Greenhouse, Staci Markos, Richard L. Moe, Thomas J. Rosatti, Margriet Wetherwax
University of California, Berkeley

Providing comprehensive, current, and rigorously supported data on the extraordinarily diverse and complex California flora requires worldwide scientific collaboration and extensive institutional cooperation. The Jepson Flora Project (JFP) was initiated in recognition of those needs, which have gained increasing urgency as direct and indirect human impacts on California's biodiversity and ecosystems have accelerated. The JFP relies on a large body of scientific talent (>200 contributing authors) and on botanical institutions that curate major collections of Californian tracheophytes. Connecting JFP authors with herbarium specimens from diverse sources has been a significant logistical challenge, especially in light of the importance of small, regional herbaria in a state where much of the flora is narrowly endemic to particular geographic areas. The rise of the Consortium of California Herbaria, a fully communal body of statewide herbaria, during the last year has been of tremendous benefit to JFP authors. The consortium provides a common database of Californian vascular-plant specimens from most major public and private collections within the state. Data entry and geo-referencing efforts, facilitated by the Consortium, and the continuing expansion of data presentation and query capabilities for the Consortium web interface have provided JFP authors with a greatly enhanced means of accessing and integrating current, collection-based information.

The most significant and lasting product of the JFP is the Jepson Manual on the vascular plants of California. The Jepson Manual will be made available both in a print edition and a web-based publication. The size of the California flora makes it difficult to produce a book that is suitable for field use. The JFP is developing a digital prototype of the Jepson Manual for Windows Mobile devices that will be more useful for field botanists. The final digital product will include descriptions of all taxa native or naturalized in California with color illustrations and digital keys. This demonstration will show our recent progress. The JFP content formatted for mobile devices is available at http://ucjeps.berkeley.edu/mobile/. The online presentation includes links to taxonomic treatments from the first edition of the Jepson Manual (1993) formatted for a PDA web browser.

Support is acknowledged from: California Digital Library

12.2. TaxonX: A Lightweight and Flexible XML Schema for Mark-up of Taxonomic Treatments

Terry Catapano[1], Donat Agosti[2], Guido Sautter[3], Drew Koning[2], Klemens Boehm[3], Norman F. Johnson[4], P. Bryan Heidorn[5], Thomas D. Moritz[6], Indra Neil Sarkar[2], Christie Stephenson[2]
[1] panix.com, [2] American Museum of Natural History, New York, [3] University of Karlsruhe, Germany, [4] Ohio State University, Columbus, [5] University of Illinois, Urbana-Champaign, [6] The Getty Research Institute, Los Angeles

An increasing number of legacy documents are being made available, through the Biodiversity Heritage Library, the American Museum of Natural History Novitates digital repository and antbase.org. The representation of these documents in machine-readable form will be essential for subsequent organization and research. We have designed Taxonx (http://wiki.cs.umb.edu/twiki/bin/view/Ants/TaxonXDocumentation) as an XML schema to encode legacy taxonomic literature in order to:

- Create open, stable, persistent, full text digital surrogates of taxonomic treatments;
- Identify taxonomic treatments and their major structural components to enable networked reference and citation;
- Identify lower level textual data such scientific names, localities, morphological characters, and bibliographic citations to facilitate their extraction by, and integration with external applications and resources; and,
- Study and describe the structure of systematics publications by creating few typical corpora of literature such as an entire journal (e.g., AMNH Novitates), across taxa (e.g., all ant systematics papers post 1995), or faunistic (e.g., all ant systematics paper covering Madagascar from 1758 to 2006).

We have applied Taxonx to a corpus of scanned AMNH Novitates documents. It is a lightweight and flexible schema that can be quickly learned and subsequently applied to a wide variety of formatting present in legacy documents. Taxonx permits, and sometimes relies on the use of external schemata (e.g., MODS for file-level bibliographical metadata). For large scale retrospective digitization initiatives, Taxonx‘s loose content requirements permits progressive markup of instances over time and at many levels of granularity, yet maintains validity through iterations (which can be done through automated methods). Taxonx’s flexible design and loose requirements enables instances to be readily converted to the NLM/NCBI Journal Archiving and Interchange DTD and schemas utilized by publishers. This allows for easier retrieval of descriptions of new taxa contained in publications for initiatives like ZOOBANK. To enable greater interoperability across applications, Taxonx also contains mechanisms for semantic normalization (e.g. inclusion of LSIDs) of the data contained in treatments (see, for instance, the "mashup" iSpecies.org which gathers together on the fly information and resources, including Taxonx encoded treatments related to ant species).

Support is acknowledged from: National Science Foundation

12.3. Demonstration Proposal: Using Google for Biodiversity Search Features

Rebecca Shapley
Google

The demonstration will showcase various biodiversity information search features that have been developed using Google’s tools. For example, the IUCN and the Consortium for Barcode of Life have worked with Google’s new Co-op feature. Data and links to their own datasets are presented at the top of search results when subscribers search on Google. Google Earth’s AntWeb community layer will also be available to view. The demonstration will include details about how other biodiversity institutions can set up similar features.

12.4. GOLDENGATE, Automation Support for XML Mark-up of Legacy Literature

Guido Sautter, Donat Agosti, Klemens Böhm, Terry Catapano
Universität Karlsruhe (TH)

Digitization of legacy literature is currently a big issue, e.g., Biodiversity Heritage Library, AMNH digital library and antbase.org. In order to preserve the structure of the documents, e.g., paragraphs, they can be marked up with XML. Additional XML mark-up may encode the logical structure of systematics publications such as the description of taxa. TaxonX and TaXMLit are two candidate schemas for this purpose (see related abstracts). These schemas build upon the fact that taxonomic publications are highly structured and standardized, each of their elements related to a particular taxon. The main structural elements include the description of the species, nomenclature, distribution, materials examined, tools for identification, phylogenies, illustrations, and bibliographic references. Such marked-up documents allow more

detailed search and text mining than provided by other digital library projects for taxonomic literature. This approach does however call for specific tools to automatically create this fine-grained mark-up.

GoldenGATE is a dedicated XML editor to encode taxonomic publications using taxonomy-specific schemas. Manually creating such detailed markup, which can reach down to the sentence level and below is cumbersome, time-consuming and therefore expensive. Automation is desirable wherever possible. Bio-NLP has lately developed algorithms like TaxonGrab (Koning et al, TaxonGrab: Extracting Taxonomic Names from Text, Biodiversity Informatics, 2005) and FAT (Sautter et al, A Combined Approach to Find all Taxon Names (FAT) in Legacy Biosystematics Literature, submitted), which identify taxonomic names in texts. NLP tools for the recognition of (collecting) locations have existed since the late 1990s, even though they are not specifically intended for bioinformatics. Both types of tools can significantly reduce the manual effort of markup. However, they do not achieve 100% accuracy, implying the need for manual corrections. Manual steps usually rely on XML editors like XMLSpy or Oxygen. But the purpose of these editors is handling existing XML data rather than creating XML documents from plain text. With these editors, marking up a document means applying NLP tools first, and then doing the rest of the markup in the editor manually, including the correction of NLP errors. The requirement to go back and forth between NLP tools and an XML editor induces further effort. Our GoldenGATE editor is designed to tightly integrate the NLP application and provide as much automation as possible to the manual markup. GoldenGATE integrates existing NLP tools through a slim programming interface. Implementations of this interface currently exist for FAT and a location extractor. Manually inserting XML tags works by selecting the tag content (the selection automatically extends to word boundaries) and creating the tag by simply selecting the XML element name. Further features of GoldenGATE comprise sequencing of automated editing steps to Pipelines, automatically processing a set of files, basic NLP (gazetteers, regular expressions), and basic mark-up transformation and filtering. Preliminary experiments show that GoldenGATE incurs significant performance gains over conventional XML editors.

GoldenGATE is available at http://idaho.ipd.uni-karlsruhe.de/GoldenGATE/.

Support is acknowledged from: DFG, NSF

12.5. A Demonstration of the Atrium Biodiversity Information System

John P Janovec[1], Amanda K Neill[1], Mathias A Tobler[2], Jason Best[1], Anton Webber[1]
[1] Botanical Research Institute of Texas, [2] Texas A&M University

Atrium is an online biodiversity information system designed to support the research activities of the Andes to Amazon Botany Program at the Botanical Research Institute of Texas (BRIT) and to make data available to collaborators and the general public. Atrium facilitates the collection, organization, and sharing of organismal and ecological information generated by the biologists, ecologists, geographers, students, and local field assistants working in this area. Development of the requirements and design of Atrium is funded by a grant from the Gordon and Betty Moore Foundation. Atrium is part of a larger More-funded project to increase knowledge of the Andes-Amazon area and to develop and test new tools and technology to document, describe, and disseminate information about the species and ecosystems in the area. Atrium provides researchers with tools to collaborate worldwide in real-time. Atrium features a digital herbarium of over 11,000 plant collections representing approximately 4,000 species, with many tools for entry, organization, and analysis of collection data. Collaborators can view complete collection data and high-resolution images (over 22,000 images are now available), print labels, annotate collections and produce annotation labels remotely. Atrium hosts geospatial data connecting to the botanical dataset using Google for desktop and online mapping of collections. Atrium also hosts extensive bibliographic records pertaining to the biodiversity and conservation of southeastern Peru. The most recent development is the tool

developed to produce color field guides and floras that can be designed and printed on-demand. Atrium: atrium.andesamazon.org.

Support is acknowledged from: Gordon and Betty Moore Foundation, Beneficia Foundation, World Wildlife Fund, Stanley Smith Horticulture Trust

12.6. Specify Software Project: Demonstration of Specify 5

J. Beach, A. Bentley, J. Burgess, K. Coggins, C.J. Grady, G Garneau,
M. Kumin, T. Noble, R. Spears, CJ Grady, J. Stewart
Biodiversity Research Center, University of Kansas

Since 1987, the Specify Software Project and its predecessor have supported biological collections institutions with database management software and related specimen data management services. Specify is licensed as an open source application for Windows and is available at no cost, thanks to ongoing US National Science Foundation grant support. Specify is used as the primary catalog database by 160 collections world-wide. Our latest production release, Specify 5 has HTML and DiGIR interfaces, full-text database searching and intuitive navigation and collection work task support.

Specify 6, in development for a late 2007 release, will be a wholly new re-implementation. Specify 6 will have the same look and feel, same capabilities and same code base across Mac OS, Linux and Windows platforms. Specify 6 will use open source components: Java for cross-platform applications; Hibernate for object relational mapping; JGoodies for user interface layout; Jasper Reports and JFreeChart for complete integration of screen, printer and exported reports; and Apache Lucene for full-text search. Specify 6 will run on the following databases: MySQL, PostgreSQL, Oracle, or MS SQL Server. Specify 6 is designed to accept plug-ins to extend its user interface, business rules and database schema. This will allow Specify to support collection data processing needs which require access to specialized network services or projects which have requirements on the boundary of traditional biological collection information processing. Specify 5.x users will be able to update their existing databases to Specify 6.

We will demonstrate the user interface and data management capabilities of our production version, Specify 5, and answer questions about our support services and upcoming releases (5.2 and 6.0) during the session.

Support is acknowledged from: US National Science Foundation grant support (BIO/DBI 0446544)

13. Workshop

13.1. LSID and RDF Hands-on Tutorial

Kevin Richards[1], Ricardo Scachetti Pereira[2], Roger Hyam[2], Lee Belbin[2], Donald Hobern[3]
[1] Landcare Research New Zealand, [2] TDWG Infrastructure Team, [3] Global Biodiversity Information Facility (GBIF)

The successful development of interoperable networks depends on the ability of clients to uniquely identify and locate data items provided by multiple sources. To address this need, TDWG has adopted Life Science Identifiers (LSID) for use as stable identifiers for data items in Biodiversity Informatics data networks.

This tutorial will take a look at LSID, the Resource Description Framework (RDF) and the Semantic Web and how they fit together and can be used within our domain. The tutorial is intended for people who have had limited exposure to these technologies, and will run through how to set up an LSID resolver and how to write RDF to describe taxonomic data that will be returned by this resolver. The tutorial will be run as a hands-on style workshop, running through working examples of RDF, LSID and the Semantic Web.

Support is acknowledged from: Landcare Research, The Gordon and Betty Moore Foundation, The Global Biodiversity Information Facility.