Biodiversity Information Standards TDWG

The Proceedings of TDWG

Abstracts of the 2007 Annual Conference of the Taxonomic Databases Working Group

16-22 September 2007
Bratislava, Slovakia
(Hosted by the Faculty of Natural Science, Comenius University)

Edited by Anna Weitzman and Lee Belbin

Published by Biodiversity Information Standards (TDWG)
and the Missouri Botanical Garden

St. Louis, 2007

TDWG 2007 sponsored by:

To be cited as:

Weitzman, A.L., and Belbin, L. (eds.). Proceedings of TDWG (2007), Bratislava, Slovakia.

This book contains abstracts of the papers, posters and computer demonstrations presented at the Annual Conference of the Taxonomic Databases Working Group held 16-22 September 2007 at the SÚZA (Správa účelových zariadení) Conference Centre in Bratislava, Slovakia, hosted by the Faculty of Natural Science of the Comenius University. The meeting attracted more than 141 participants from 25 countries and 88 prestigious scientific research institutions, museums and companies.

The editors acknowledge with thanks the contribution of Arturo Ariño, Reed Beaman, Lee Belbin, Walter Berendsohn, Stan Blum, Alex Chapman, Renato de Giovanni, Stinger Guala, Gregor Hagedorn, Donald Hobern, Roger Hyam, Gail Kampmeier, Jessie Kennedy, Éamonn O Tuama, Ricardo Pereira, Rich Pyle, Adrian Rissoné, Annie Simpson and Neil Thomson towards the editing of this publication, and also the numerous peer reviewers.

Published and distributed as an Adobe® Portable Document Format (PDF) document for free download from the Conference web site at http://www.tdwg.org/conference2007/

ISBN 978-1-930723-71-9

Contents

Proceedings of TDWG
Abstracts of the 2007 Annual Conference of the Biodiversity Information Standards (TDWG)

Session 1. Client's Perspectives: User Needs

1.1. EDIT needs Biodiversity Information Standards

Walter G. Berendsohn, Markus Döring, Malte C. Ebach
Botanic Garden & Botanical Museum Berlin-Dahlem

The European Distributed Institute of Taxonomy (EDIT) is a network of 26 leading natural history institutions and organisations in the European Union, the United States and Russia. EDIT is a "Network of Excellence" project financed by the European Commission, this implies that its main aim is the durable integration of institutional resources to jointly meet the challenges taxonomy faces today. The EDIT project started on the March 1st 2006 and will last until 2011.

An integral part of EDIT is the creation of an "Internet Platform for Cybertaxonomy". This is a distributed computing platform that will allow taxonomists to do taxonomic revisions (including processing the results of field work) more efficiently, expediently and via the web. It consists of interoperable but independent platform components, which can take the form of software applications (desktop or web-based) for human users or (web) services. The envisioned platform will not have a single user interface or website, instead it will be a collection of interacting components which may be combined and assembled according to the task at hand. A central endeavour of EDIT is to establish a Common Data Model that platform components adhere to. In the near future, this will include more or less loose coupling with existing software solutions. More information is available at http://wp5.e-taxonomy.eu/EDIT-Architecture.html.

Present and future EDIT member institutions will join agreements for the maintenance and use of specific parts of the platform (components, standards, data provision or access) once they are considered mature enough for practical use.

For the development of the Platform, EDIT will closely collaborate with projects, organisations and initiatives with overlapping aims, prominently among them TDWG. Members of EDIT staff have been active in TDWG groups and meetings. TDWG offers an indispensable forum for contacts and networking among those active in biodiversity informatics. For EDIT software development, TDWG standards and discussions will be considered and incorporated. One obstacle is the lack of integration among TDWG Standards (*e.g.*, TCS, SDD, ABCD), which on a structural level, are largely incompatible with each other. However, the achievements of TDWG groups on data definition at the atomic level (*i.e.*, definition of the semantics of data elements) are recognised and indispensable for EDIT's planned Common Data Model. This will need further development and well defined content standards (ranging from controlled vocabularies to data services), which requires close involvement of the biological community in TDWG working groups.

Another area EDIT is engaged in is certification of biodiversity informatics software. The new TDWG standards process is a possible model for such an endeavour. One of the criteria for "EDIT certified software" will certainly be compatibility with applicable TDWG standards as well as wider used standards recommended by TDWG.

Support is acknowledged from: European Commission Framework Programme 6

1.2. Biodiversity Heritage Library: Progress & Potential

Chris Freeland

Missouri Botanical Garden

The Biodiversity Heritage Library (BHL) is an international consortium of 10 natural history libraries with a goal to digitize a significant collection of materials across the member libraries. A working prototype for BHL is online at http://www.biodiversitylibrary.org.

Developments in the next two years include enhancing this interface and providing globally unique identifiers and robust services guided by TDWG standards and recommendations. Developments will allow remixing and incorporation of material into complimentary applications.

Support is acknowledged from: Alfred P. Sloan Foundation, John D. and Catherine T. MacArthur Foundation

1.3. One million species in the Catalogue of Life – a triumph for Species 2000 and ITIS, or for TDWG standards?

Frank A. Bisby

Species 2000 Secretariat, School of Biological Sciences, University of Reading

On 29 March 2007 Species 2000 and ITIS held their 'One Million Species Day' celebrating reaching one million species in their Catalogue of Life. This was achieved by federating species checklists from 47 taxonomic databases from around the world. Not only was the Species 2000 programme initiated by TDWG, but from the start in 1996 the programme depended on standards for interoperability within its architecture for federating many taxonomic databases. Then as now, TDWG was considered the community's forum and authority for standards. So how has TDWG served this client community over the eleven years, and how has this client responded? First – the will of TDWG to establish and promote practical standards as different from acting as a forum for innovation in biodiversity informatics has fluctuated over the years. Second – the early cohort of standards were largely content standards, but these, nonetheless can prove valuable to a programme such as ours. Third – the gradual shift to schemas and protocols at the informatics level has done much to widen the generality of solutions and to open opportunities for multiple uses. Fourth – we need to be realistic about the time-lags between design, adoption, implementation and effective adoption in the community, and where possible to manage this life-cycle rather severely. The response from our species checklist database community has been decidedly mixed. Huge variations in the sense of purpose and in perceptions of how it should be done, have led to some exciting innovations, but also to much needless diversity in how simple tasks are done. Some of the disappointing elements in this response relate to the weak uptake of generic software in our community, and the shortage of success stories in this area. Lastly, with participation of more than 50 databases in the Species 2000 programme, we can bring to TDWG incipient standards that have already proved effective within this community. One is the SPICE Protocol for federating species checklists, another is the Species 2000 Data Content Standard for species checklists, and we have started a 'best practice' document that addresses content and management. On behalf of both the Species 2000 and the ITIS programmes it is important to reiterate both the fundamental importance of interoperability standards and the work that TDWG is doing. Nowhere are standards more important than in biodiversity. Our ability to describe,

model and manage global biodiversity depends entirely on our ability to synthesise high level knowledge from the myriad individual observations and syntheses made independently around the world: distributed systems and interoperability are central to this task.

1.4. User Needs - The alpha and omega of system design

Charles J.T. Copp

Charles Copp Environmental Information Management

This presentation will include user needs, the role of interfaces and web services in building systems to serve different types of users and the use of thesauri for providing appropriate user-targeted terms.

Developers jokingly complain that the problem with software is the users: ‘users never read manuals and can be bloody-minded or even downright stupid’. Most potential users do not really understand their data requirements or have a clear idea of what can be delivered. This is especially true in large scale information projects, of which the database software forms only a part, for instance, a local or regional biodiversity network. Is there any consensus on what the potential users want out of a biodiversity network?

The key issues are: who are the users, what are their real needs, what problems can the proposed system solve, how can different levels of user get what they need, will their requirements change over time, and who will pay for it? In the UK at least, a failure to solve these issues contributes to the confusion and demoralisation in library, museum and school services. All too often the debate is of what should they get not what do they need? Is there a danger of this with biodiversity networks?

Establishing user needs is a difficult and under-estimated task. The, now outmoded, Structured Systems Analysis and Design Methodology (SSADM) was particularly good for describing existing systems and establishing user requirements. Data flow diagrams (DFDs) remain one of the most powerful tools for charting the limits of the system and defining what parts affect what users but have little to say about user interfaces. Times move on and the rise of prototyping, extreme programming, object-oriented methodologies and web-related technologies have given us new paradigms for system development but the user definition problems remain much the same. Much of the effort still goes into data capture, data storage and linking or querying distributed databases but not enough effort goes into data re-purposing or repackaging for different types of users. Even less effort goes into what sort of data were needed in the first place. The result is increasingly large, interconnected data systems that solve few real-world problems.

The work to create data models and set terminology, validation and verification standards on an international scale continues to be spectacularly successful and TDWG and related projects can be justifiably proud of their achievements. This is not true for usability of data access applications, which is probably the greatest limiting factor in extending the value of these systems. For instance, it is quite clear that the choice of language and depth of information used in answering questions from children, members of the public or keen local naturalists are very different. Likewise in building applications “one size never fits all”.

We are still at the rudimentary stage of interface design, ergo the example of the blank text box labelled “Enter a species name”, and hierarchical taxonomic trees are little use to non-specialists. Real progress will only come with interfaces that designed for the level of knowledge of the user. It is especially important to give users the means to explore what is held within a system

according to their level of experience and interest. Users must not be forced to follow a rigid access routine.

1.5. Exploring the Brave New World of eTaxonomy

Chuck Miller

Missouri Botanical Garden

The Missouri Botanical Garden has multiple initiatives in progress that are opening the door to a new world of taxonomic research methods. We are developing new online pathways to taxonomic data, digitizing reference literature, and engaging with other institutions to better integrate plant data world-wide. But to truly fulfill the vision of this new world requires development of standard ways to share and integrate multiple dimensions of data. I will explore the peaks and valleys of this unexplored territory and suggest some priorities for moving forward from the point of view of a taxonomic research center.

Session 2. Client's Perspectives: Examples of TDWG Standards in Use

2.1. TDWG Standards in use within the Global Biodiversity Information Facility (GBIF) Data Portal

Tim Robertson

GBIF

This presentation will include a very high level overview of the Biodiversity Data Portal (http://data.gbif.org) offered by the Global Biodiversity Information Facility (GBIF http://www.gbif.org). The process of harvesting, parsing, and efficiently serving data for graphic user interface (GUI) tools and reporting services will be covered, illustrating the heavy dependency on TDWG standards. An overview of the mechanism employed to normalise the incoming data from various formats will be explained. This will highlight a use for a Universal Biodiversity Data Bus, which is a common set of standards for publishing, discovering and accessing data across the Internet.

From this overview, non technical participants will receive an insight into the data flow involved, some of the limitations faced, and how important TDWG formats are when processing data. It is expected that this will form a good basis for subsequent technical discussions relating to the Universal Biodiversity Data Bus.

The data within the GBIF network is collated using Distributed Generic Information Retrieval (DiGIR) , the Biological Collection Access Service for Europe (BioCASE), and the TDWG Access Protocol for Information Retrieval (TAPIR). These are all protocols encapsulating various versions of DwC (Darwin Core 2) and Access to Biological Collections Data (ABCD), and the data is served to the public through the new GBIF Data Portal in many forms including DwC and the Taxonomic Concept Schema (TCS) and employing Life Science IDentifiers (LSIDs).

Support is acknowledged from: The Global Biodiversity Information Facility

2.2. Assessing the Threat of Invasive Species in South America: an ensemble modeling approach in support of data standards, integration, and dissemination

Miguel Fernandez[1], Wendy Tejeda[2], Guillermo Duran[3], Adriana Rico[2], Christian Arias[2], Maria Laura Quintanilla[2], Alberto Pareja[2], Juan Carlos Chive[4], Monica Rivera[2], Healy Hamilton[5]

[1] University of California, Merced; California Academy of Sciences, San Francisco, [2] Centro de Analisis Espacial, Universidad Mayor de San Andres, La Paz, Bolivia, [3] California Academy of Sciences, San Francisco; San Francisco State University, [4] Museo Noel Kempff Mercado, Bolivia, [5] Center for Biodiversity Research and Information, California Academy of Sciences, San Francisco

Today's global economy moves unprecedented quantities of people and products around the planet, increasing the probability that alien species will be introduced and successfully established beyond their native ranges. Invasive alien species (IAS) are the second most important cause of biodiversity loss, and pose additional threats to agriculture and human health.

Together, IAS, habitat alteration and climate change are dramatically re-shaping biogeographic patterns across the globe. We need accessible data and analysis tools to assess the threats of IAS at multiple stages: to identify at-risk habitats before invasion occurs, to identify potential arrival sites, and to understand potential routes and rates of dispersal. Beyond threat assessment, data and tools are needed to create conservation strategies that mitigate these threats. In Latin America, economic losses from IAS amount to billions of dollars annually, but strategies to minimize the damage of IAS are generally underdeveloped. We describe an international collaboration using novel techniques to predict the potential distributions of IAS in South America.

Researchers from the California Academy of Sciences, The Nature Conservancy (TNC), and the Centro de Analisis Espacial of the Universidad Mayor de San Andres in Bolivia, are using ensemble distribution modeling to generate composite potential distribution maps for 300 of the most threatening IAS in South America. We are using species occurrence data, derived from both museum specimens and observations obtained from the TNC Invasive Species Initiative, the IABIN Invasive Species Information Network (I3N), and the Global Biodiversity Information Facility (GBIF). Global environmental data layers and higher resolution regional layers are being used to predict distributions of IAS. Seven distribution modeling algorithms are being run for each IAS: Bioclim, Minimum distance, Climate space model, Distance to average, Environmental distance, Garp and MaxEnt. The outputs are combined using a consensus method to produce an ensemble model. Composite maps reveal 'hotspots' of IAS susceptibility, depicting which regions of South America are most at risk from the threats of IAS. We are compiling a database of all the biological and spatial data input, as well as all the output models, which will be made publicly accessible.

Our future goals include: 1) creating web access to all project inputs and outputs, including the high-resolution regional environmental data layers we created specifically for this IAS modeling research; 2) building a website to support the collection and distribution of invasive species occurrence data in Bolivia, the only South American country not currently contributing to the I3N effort; and 3) incorporating estimates of future land use and climate change in predicting IAS distributions for South America.

Support is acknowledged from: California Academy of Sciences, The Nature Conservancy, Centro de Analisis Espacial, TDWG Infrastructure Project

2.3. Results of a Needs Assessment Survey of the Global Invasive Species Information Network (GISIN)

Annie Simpson[1], Jim Graham, Michael Browne[2], Hannu Saarenmaa[3], Elizabeth Sellers[4]

[1] US National Biological Information Infrastructure, [2] IUCN Invasive Species Specialist Group, [3] Finnish Museum of Natural History, [4] US Geological Survey

The Global Invasive Species Information Network (GISIN) is developing a system for the exchange of invasive species information over the Internet utilizing TDWG standards. A critical step in the process of creating this system is to determine requirements of its eventual users. The system's users can be divided into four types:

1) data providers: organizations and persons that will provide data;
2) data consumers: intermediary organizations and persons that will use the system's primary data for modeling and other analyses, and then make these value-added products available back through the system;

3) stakeholders: those who support the system without necessarily providing or consuming data; and

4) end users: those who use the system's data and/or analyses, but do not provide products back through the system.

The results of a needs assessment survey to obtain user requirements, which ran from 15 December 2006 through 15 February 2007, had both surprising and expected elements.

With 137 respondents from 41 countries, 80% identify themselves as providers and consumers of invasive species data. As expected, most (77%) offer invasive species spatial/temporal information, profiles/species pages (65%), and checklist information (59%). Although most are data providers, their technical knowledge is surprisingly low: 80% said they do not know what existing protocols are appropriate for invasive species information management; 45% do not know the level of web services their organization provides and/or uses; 75% did not know what schemas/grammars would be acceptable to copy or extend for the GISIN data exchange system. A complete report of survey results is available at http://www.gisinetwork.org/Survey/SurveyResultsFinal.htm.

From the results of this survey, it was determined that standards for the GISIN system will need to be both simple to implement and easy to understand, if the system is to be a success. Because only 23% of respondents said Python is an acceptable programming language for a toolkit, a Py-wrapper application is not being considered at this time. Likewise, SOAP (Service Oriented Architecture Protocol) is not being considered, because it is more complex than is needed and would significantly slow data exchange within the system.

Because the results of the needs assessment survey indicated that a complex solution would not be met with wide acceptance and would be too expensive for current funding levels, the GISIN system operates as a simple HTTP Request/Response protocol. This method is used to serve web pages on the Internet and ensures the best access through firewalls without security problems. This approach also provides the required flexibility with high performance.

The GISIN protocol is a subset of the functionality defined by TAPIR (TDWG Access Protocol for Information Retrieval).Only simple Key-Value Pair (KVP) requests are supported because complex filters encoded as XML (Extensible Markup Language) were not required.

Respondents to the needs assessment survey listed ASP, JSP, and PHP (in that order) as acceptable internet frameworks for a toolkit. Therefore a GISIN data providers' workshop is being planned for 13-16 November with programmers of these three frameworks as instructors. Although 80% of the respondents preferred receiving a software toolkit to install and configure on their server to become a GISIN data provider, at the November meeting programmers and database managers will create their own code to map each of their unique database systems to the GISIN protocol.

Special thanks to Jeremy Kranowitz, who donated his time to configuring, running, and analyzing the survey, and to his organization, The Keystone Center.

Support is acknowledged from: US National Biological Information Infrastructure; GBIF; IUCN-Invasive Species Specialist Group; US National Institute of Invasive Species Science; The Keystone Center

2.4. When Taxonomies Meet Observations: An Examination of Taxonomic Concepts used by the Observation Systems eBird and the Avian Knowledge Network

Paul Edward Allen

Cornell Lab of Ornithology

Ideally, observations of organisms are identified by the observer with a taxonomic concept, consisting of the taxonomic name and the reference defining that name. However, systems that manage observational data must be able to accommodate imprecision or uncertainty in concepts since observers are not always able to classify an organism as a single, well-established species (or subspecies) taxonomic concept. There are several instances in which indefinite concepts are required. First, an observer may identify an organism as a hybrid of two species. Second, imperfect observation conditions (*e.g.*, limited visibility), limited experience, or other factors might limit an observer to classifying an organism only as a member of some subset of concepts, where the subset has meaning to field observers, but may not be circumscribed by a academically established taxonomic concept. Finally, similar factors might lead an observer to identify an organism only to a genus or higher taxonomic rank. The first two cases may lead managers of observation systems to informally become taxonomists, since they must create concepts to accommodate the observations they hold and which do not fall into a well-established taxonomic concept. This presentation shows how uncertainty and imprecision in taxonomic identity are handled by the Bird Monitoring Data Exchange standard used by the Avian Knowledge Network (AKN, www.avianknowledge.net) and the TDWG Taxonomic Concept Transfer Schema standard.

Analysis of 29 million avian observation records from eBird (www.ebird.org) and the Avian Knowledge Network shows that uncertain and imprecise taxonomic concepts represent 5% (eBird) 18% (AKN) of the concepts in these systems. However, observations labeled with uncertain or imprecise concepts represent only 0.05% (eBird) and 1% (AKN) of the observations held in those systems.

2.5. Taxonomists at work: relationships of process and data

Anna Weitzman[1], Christopher Lyal[2]

[1] Smithsonian Institution, [2] Natural History Museum, London

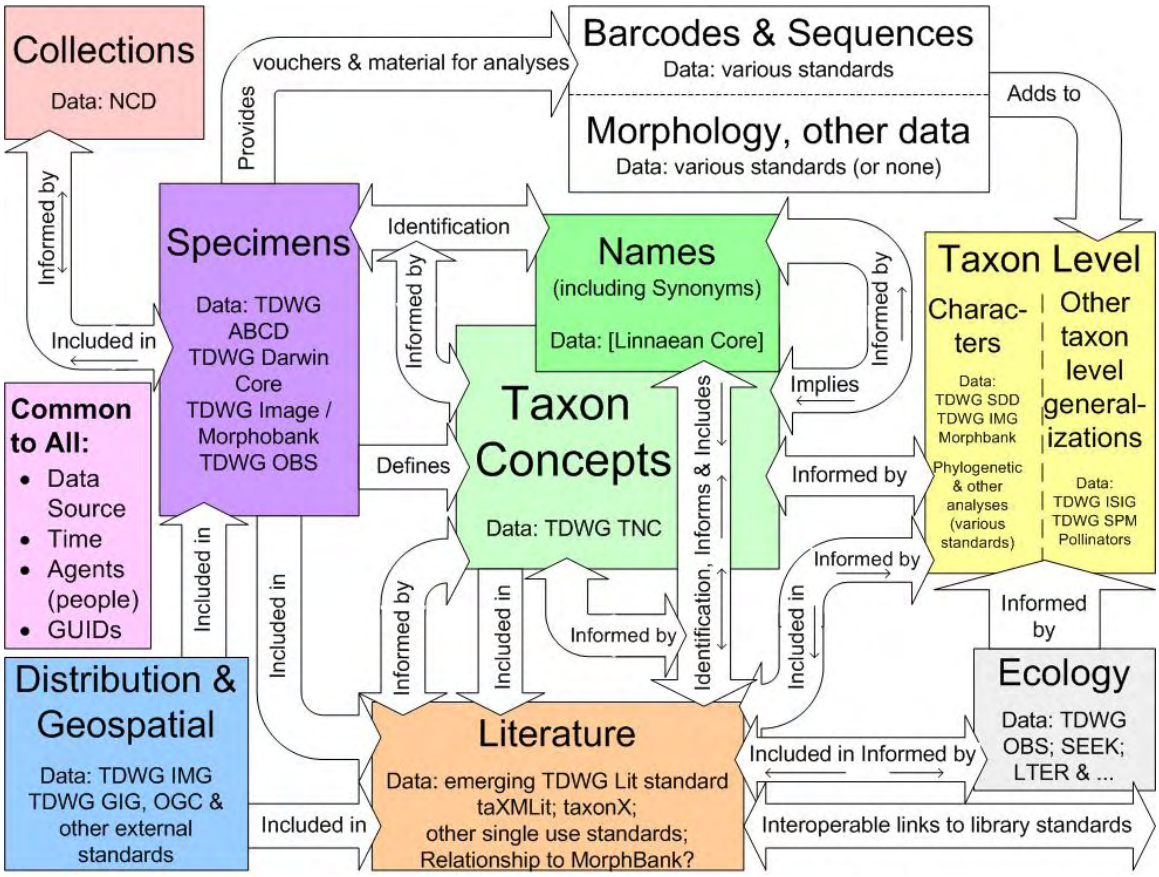
Taxonomy has developed in practice over hundreds (or thousands) of years. Humans have always been interested in the world around them and using names to communicate what they know about the organisms that they see. From simple beginnings lost in the origins of human culture, this process has developed into taxonomy as we know it. Though it has become formalized, it is still mainly about learning about the organisms that we share the planet with and using names to communicate about them.

In order to do this, we have developed systems of nomenclature for applying names to organisms; collections of preserved organisms which serve to help us understand, document, and apply names to what we observe to be taxa (taxon concepts); and ways to document the information in publications. After 300+ years of generating these systems and collections, there is a vast body of existing knowledge that is used routinely in current taxonomic work. Additional sources of data have been added recently and been incorporated into workflow.

Understanding the information flow between different data and information sources as employed by taxonomists and others is important to model how interoperable data systems should connect.

The results of an analysis of the data flow and working practices can be depicted in the following diagram. Standards and schemas employed for the different elements are identified. The diagram also indicates where interoperability between particular schemas must be developed.

We will present and explain the diagram, especially as it relates to the user needs presented at the opening of TDWG 2007. At the close of TDWG 2007, we will present it again in the context of the entire meeting's discussions and presentations, with any amendments that have been shown to be needed.



Support is acknowledged from: The Atherton Seidell Fund of the Smithsonian Institution

Session 3. Needed Technologies: Introductions and Demos

3.1. TDWG Standards Architecture - What and Why

Roger Hyam

TDWG Infrastructure Project

In 2005, the TDWG Infrastructure Project (TIP) was given the remit of devising an umbrella architecture for TDWG standards. A meeting (TAG1) in April 2006 led to the establishment of the basic principles for underlying the standards architecture. The TIP has been promoting adoption of this common architecture over the last 18 months. But why have a standards architecture at all?

There is no need for a standards architecture when exchanging data within the federation of similar applications such as natural history collections. The federation is a closed system where a single exchange format can be agreed on. The federation can grow by adding new members whose needs are met by the format. This model has worked well in the past but it does not meet the primary use case that is emerging. Biodiversity research is typically carried out by combining data of different kinds from multiple sources. The providers of data do not know who will use their data or how it will be combined with data from other sources. The consumer needs some level of commonality across all the data received so that it can be combined for analysis without the need to write computer software for every new combination. This commonality needs to seamlessly extend to new types of data as they are made available. An architecture is required to provided this commonality.

What form should the architecture take? A degree of commonality could be achieved simply by specifying how the data should be serialised. If all suppliers passed data as well-formed XML for example, it would provide a degree of interoperability. Clients would however, still not know how the elements within one XML document relate to those in another, or how the items described in those documents relate. At the other extreme, the architecture could provide a detailed data type library which describes the way in which each kind of data should be serialised at a fine level of granularity. In other words, which XML elements must be present and what they should contain? It is however unlikely that a single set of serialisations would meet all needs any more than a single federation schema would. Some thematic networks require that they have well defined data types to ensure that the data passed is valid and fit for purpose.

The architecture has to meet two needs. It has to allow generic interoperability but also restricted validation of data for some networks. It does this using three interlinked components. 1) An ontology is used to express the shared semantics of the data but not to define the validity of those data. Concepts within the ontology are represented as URIs (Universal Resource Identifiers). 2) Exchange protocols use formats defined in XML Schemas (or other technologies) that exploit the URIs from the ontology concepts. 3) Objects about which data are exchanged are identified using Globally Unique Identifiers. This means that, although exchanges between data producers and clients may make use of different XML formats, the items the data is about and the meaning of the data elements is common across all formats.

Support is acknowledged from: The Gordon and Betty Moore Foundation

3.2. Life Sciences Identifiers (LSID) and the Biodiversity Information Standards (TDWG)

Ricardo Scachetti Pereira

TDWG Infrastructure Project

Over the last few decades, the biodiversity information community has made primary data available for environmental analyses and decision making. Information on a million scientific names is now available through data providers such as the Integrated Taxonomic Information Service (ITIS), Species2000 and the Catalogue of Life (CoL). Almost one hundred million specimen records are provided by Herbaria and Natural History Museums around the world.

To use these data more effectively, clients need mechanisms to: a) refer to authoritative information resources, b) facilitate data integration and c) detect duplicates of the same resource. To achieve these goals, a system of globally unique identifiers (GUIDs) is needed.

The TDWG Infrastructure Project (TIP) established a TDWG Globally Unique Identifiers Task Group (TDWG-GUID) to provide recommendations for use of GUIDs in our domain. The GUID members concluded that the Life Sciences Identifiers (LSIDs) were the most appropriate technology to address current problems.

LSIDs are unique, persistent, location-independent, resource identifiers for biologically significant resources such as species names, concepts, occurrences, genes or proteins. LSIDs identify and locate biological objects via the web and overcome limitations of current naming schemes.

I will provide an overview of Life Science Identifiers and how they solve current problems. I will report on the work performed by the GUID group over the last two years and provide recommendations and a plan on the use of LSIDs in the biodiversity information domain.

Support is acknowledged from: The Gordon and Betty Moore Foundation

3.3. Nala: A Semantic Data Capture Extension for Mozilla Firefox

Ben Szekely[1], Ricardo Scachetti Pereira[2]

[1] Cambridge Semantics Inc., [2] TDWG Infrastructure Project

Collecting and integrating biodiversity informatics data from diverse websites and transforming these data into the formats accepted by the analysis tools takes considerable resources.

Semantic Web tools such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL) make it easier for computers to interpret the meaning of data items. Life Sciences Identifiers (LSIDs) are another Semantic Web product that allows information resources to be uniquely named and easily located.

Nala is a Semantic Web data capture tool that we have developed to demonstrate how Semantic Web technologies, in particular, RDF, OWL and LSIDs, may be used to improve the process of data capture and integration.

Nala is a Mozilla Firefox web browser extension, similar to Piggy Bank, which allows users to capture and integrate data while browsing the Web. Nala looks for data that may be acquired and transformed into RDF from web pages that are browsed. When such data are detected, the user is given the option to acquire, transform it into RDF format and store it in a repository called an RDF triple store. Data in the repository may then be integrated using OWL vocabularies such as

Dublin Core or the TDWG Ontology and LSID Vocabularies and exported in CSV and MS Excel formats.

Support is acknowledged from: Gordon and Betty Moore Foundation

3.4. Key Enabling Technologies: Transfer Protocols

Donald Hobern

Global Biodiversity Information Facility

The new TDWG data architecture relies on three core abilities:

1. Constructing data objects representing objects and concepts in biodiversity informatics. This is the purpose of the TDWG data standards.
2. Referring reliably to data objects. This is why TDWG has adopted Life Science Identifiers (LSIDs) as a globally unique identifier technology.
3. Discovering and accessing data objects. This why TDWG develops its own data access protocols and explores other protocol standards.

TDWG's work in this area has led to the family of protocols beginning with DiGIR and BioCASe and leading to TAPIR (the TDWG Access Protocol for Information Retrieval) today.

The DiGIR protocol has been used extensively by a range of major projects to support exchange of specimen and observation data using Darwin Core. DiGIR provides a flexible XML language for making remote search requests against a web-connected database. More importantly DiGIR provides a tool for organisations to map their databases into a common set of concepts such as Darwin Core.

BioCASe introduced support for records with a significant nested structure such as the ABCD schema. BioCASe simplified the use of the protocol with external data models developed without knowledge of DiGIR or BioCASe.

The TAPIR protocol learns from DiGIR and BioCASe and adds new features of its own. Two implementations of the protocol are currently available, pyWrapper (written in Python) and TapirLink (written in PHP).

To use a protocol such as TAPIR, a data administrator maps a local database to a set of concepts recognised by the community (*e.g.*, ScientificName, Locality and CatalogNumber are Darwin Core concepts recognised by a wide range of projects). TAPIR software then offers the following operations:

- Metadata – retrieve descriptive information about a dataset;
- Capabilities – retrieve the technical capabilities of the TAPIR server and the concepts mapped by the data administrator;
- Ping – check that the TAPIR server is active;
- Inventory – retrieve a list of distinct values within the dataset for one or more concepts, with counts of matching records; and
- Search – retrieve records matching a set of filter conditions.

TAPIR can handle requests encoded as XML documents or as a set of parameters supplied within a URL. TAPIR supports common request and response templates to format results for different tools. For example, TAPIR can issue requests based on Darwin Core concepts and receive results

as a Google Earth KML document or an RSS feed. Installing TAPIR software may therefore be an efficient way to expose data for a range of other client tools.

TDWG's re-engineering of its data standards as reusable vocabularies enables the use the same terms and definitions in different contexts. TDWG could use its own standards with many general purpose data access protocols. Examples include:

- OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) – standard access to metadata for a wide range of online resources.
- WFS (Open GIS Web Feature Service) – a standard that could be used to map locations of species occurrences.
- SPARQL Query Language for RDF – a standard allowing complex queries across different data sets.

Session 4. Ontologies and Vocabularies: Atomizing biodiversity information

4.1. The Role of Ontologies in the TDWG Architecture

Roger Hyam
TDWG Infrastructure Project

The TDWG Standards Architecture is based on three pillars.

1) An ontology or ontologies,
2) A series of exchange protocols and associated message formats, and
3) The use of Globally Unique Identifiers for primary data objects.

The nature of the ontology and how it integrates with the exchange protocols and GUIDs is discussed here. The justification of the overall structure of the architecture was given in an earlier talk "TDWG Standards Architecture - What and Why". A specific example of the application of the ontology is given is a later talk "RDF over TAPIR".

Prior to the TDWG standards architecture, data exchange was based solely on passing XML documents. This is good for federation networks but it is not as suitable for sharing different types of data across generic bus-type architecture – which is emerging as the primary use case. Combining documents is difficult because the meaning of the elements within the documents depends on their context. If we initially model the shared data as an ontology of linked classes of objects rather than documents it becomes possible to construct documents, from the perspective of different base classes that map directly to the ontology. Clients can then combine documents from different perspectives (and of different formats) because the documents are composed of serializations of objects that are typed in the ontology and the clients understand the ontology.

Applying the principle of separation of concerns, it is possible for the definition of the validity of the documents exchanged to be defined outside the ontology. The ontology can be used to specify the meaning of the namespaces whilst XML Schemas (or some other technology) can be used to specify valid document structures for any particular exchange application. An ontology is therefore central to unifying disparate application schemas.

Last year a team lead by Jessie Kennedy and including representatives from across TDWG interest groups developed an initial high level ontology of the biodiversity domain. This ontology is available through the TAG Wiki. Creating the ontology was a valuable exercise but everyone involved recognised it needed more work before it could be put to production use. At the same time a programme was actively rolling out LSID (Life Science Identifiers) authorities. The meta-data returned by LSIDs are in RDF format and, to be useful, requires an RDF vocabulary or ontology that at least defines the object types. The TDWG ontology was not going to be developed to a sufficient level of detail in the allotted time. The decision was therefore taken to develop a series of smaller ontologies that could serve as an application layer within the larger TDWG ontology and to only loosely link them into the more general or higher classes in the ontology. These two ontologies are referred to as the "LSID Vocabularies" and the "Current TDWG Ontology".

The LSID Vocabularies are now entering production use and, in due course, there will be a requirement for them to be linked to a higher level ontology so as to permit inference. Here, I propose that this link is not made but that a separation of concerns is again followed. There are

multiple ways in which the basic classes of exchanged data could be related. No one set of these relationships is suitable for all applications. It is therefore important not to impose a top-down interpretation of the data but to allow for the possibility of multiple higher level classifications of which the Current TDWG Ontology may only be one.

Support is acknowledged from: The Gordon and Betty Moore Foundation

4.2. Integrating TDWG standards with EDIT's Common Data Model

Markus Döring[1], Andreas Müller[1], Ben Clark[2], Marc Geoffroy[1]

[1] Botanic Garden Botanical Museum Berlin, [2] Imperial College, London

The European Distributed Institute of Taxonomy (EDIT) is building a distributed computing platform that assists taxonomists to do taxonomy efficiently, expediently, and via the web. At the heart of this network is a shared domain model, EDIT's Common Data Model (CDM). The UML has been chosen to define the object oriented domain model which is still under active development. From this platform independent model a Java specific UML model is derived which in turn is translated into Java source code using Enterprise Architect. The Java classes, together with a persistency and a thin service layer, are then released to be used by several applications within EDIT and also as part of the CATE project.

The CDM is being modeled mainly with the TDWG schemas and ontology in mind, but also incorporating ideas from the Berlin Model, CATE, BibTex and others. For serialization XML together with XML schemas has been selected. For the purpose of a tightly integrated domain model the TDWG ontology did not seem appropriate and unfortunately the existing TDWG XML schema standards did not integrate well with each other. The upcoming CDM will therefore be yet another data model, but an integrated one, with a single XML schema which will translate to current TDWG standards nicely.

http://dev.e-taxonomy.eu/wiki/CommonDataModel
http://dev.e-taxonomy.eu/trac/wiki/CdmLibrary
http://www.cate-project.org/

Support is acknowledged from: EDIT Network of Excellence, 6th FP EU

4.3. ALTER-Net: A Data Ontology for LTER Observations and Measurements

Kathi Schleidt

umweltbundesamt

ALTER-Net is a "Network of Excellence" (NoE) funded by the EU's 6th Framework Programme with the goal of creating a European long-term interdisciplinary facility for research on the complex relationship between ecosystems, biodiversity and society. Within this NoE, Work Package I6 has the task of constructing a framework within which can be built a system to manage biodiversity data, information and knowledge from the NoE.

Based on the user requirements, we have formulated our vision of an ideal architecture for such a network, based on an object oriented Data Structure, also referred to as an ontology. This means that there are classes, instances (of classes) and relations between them (classes as well as instances). Additionally, it should be possible to derive or inherit new data types from existing ones.

In such an ontology, metadata and data are all represented in the same object oriented data structure, thus removing the artificial divide between metadata and data. "One man's metadata is another man's data (and vice versa)."

One important concept in the creation of ontologies is that, while instances may be created independently by the net-work partners, the creation of classes requires a commitment by the community members. Once the necessary classes have been defined, partial ontologies dealing with individual topic areas can be created and maintained independently. Some topics of partial ontologies are:

- Taxonomic lists (species, vegetation types, soil types…)
- Lists of political administration unites (as references)
- Topologic (geographic) data
- Literature citations
- Socio economic data. (Crop yields…)

Metadata mark-up then consists not of entering the actual information but rather linking to the relevant partial ontologies. We will be presenting the basic core ontology required for the representation of ecological and socio-economic data.

Support is acknowledged from: European Commission FP6

4.4. An ontological approach to describing and synthesizing ecological data, using a generalized model for "scientific observations"

Mark Schildhauer[1], Matthew Jones[1], Joshua Madin[1], Shawn Bowers[2]

[1] National Center for Ecological Analysis and Synthesis, [2] UC Davis Genome Center

Research in the ecological and environmental sciences increasingly relies on the integration of traditionally small, focused studies to form larger datasets for synthetic analyses. A broad range of data types, structures, and semantic subtleties occur in ecological data. This extreme heterogeneity makes discovery and integration of environmental data a difficult and time-consuming task. By formally defining the notion of "scientific observation", we have developed an ontology that captures the basic semantics of ecological data required for synthesis. Observations are distinguished at the level of entities (*e.g.*, location, time, thing, concept); and the characteristics of those entities (*e.g.*, area, height, color) are measured (quantified, named, or classified) as data.

Our framework permits observations to be inter-related via context (such as spatial or temporal containment), further enhancing the possibilities for comparison and alignment (*e.g.*, merging) of heterogeneous data. Advanced forms of data discovery and integration are made possible through the use of a semantic annotation language that links observational constructs within the ecological data, to concepts that can be drawn from different domain ontologies (*e.g.*, a biodiversity ontology, or ecosystems ontology). Our current framework accomplishes this by enabling a researcher to annotate metadata descriptions (such as in Ecological Metadata Language, EML) of individual datasets with appropriate ontological terms.

The generalized approach to modelling "scientific observations" using ontologies that link to metadata descriptions of the raw data provides a powerful, extensible mechanism for enhancing data discovery and integration, allowing scientists to address questions that were previously intractable. Prototype demonstrations of these capabilities are operational within the Science

Environment for Ecological Knowledge (SEEK) research project, and are based on open-source technologies and standards for metadata and ontology construction, that are compatible with recommendations from the World Wide Web consortium.

Support is acknowledged from: The National Science Foundation, USA

Session 5. LSIDs: Gluing it together to meet users' needs

5.1. LSIDs for Taxon Names: The ZooBank Experience

Richard Pyle
Bishop Museum

The International Commission on Zoological Nomenclature (ICZN) has, for the past 112 years, set the rules by which scientific names for animals are established, as described in the ICZN Code of Nomenclature. In 2005, the ICZN Secretariat and Commissioners announced "ZooBank", a proposed registry of zoological names and nomenclatural acts. The intention of ZooBank is to serve as a mechanism for making information about new and historical scientific animal names more available and accessible than by traditional means of information dissemination through paper-based publications. The complete implementation details of the ZooBank registry are currently being discussed, developed, and tested. The first step of the implementation process involves the creation of a prototype web site that will eventually mature into the full-blown ZooBank registration service.

The Bishop Museum in Honolulu has agreed to host the initial implementation of the ZooBank prototype. With financial support from TDWG/GBIF through ICZN, in partnership with Landcare Research (New Zealand), Bishop was able to establish a functioning resolver for Life Science Identifiers (LSIDs) and a content provider following the TDWG Access Protocol for Information Retrieval (TAPIR). LSIDs were assigned to a sample dataset of verified taxon names and literature citations from the Catalog of Fishes database. An LSID resolver service was set up on a Windows/IIS server using VB.NET code developed by Kevin Richards of Landcare Research. A TAPIR provider service was also implemented to return metadata associated with these LSIDs. LSIDs assigned to taxon names return metadata in accordance with the TDWG Taxon Name LSID Ontology, and LSIDs assigned to publication citations return metadata in accordance with the TDWG Publication Citation LSID Ontology.

A discussion of the implementation of these services, including alternate strategies for defining "taxon name objects" and associated implications, and the role of nomenclators for providing taxonomic services, will be provided.

Support is acknowledged from: TDWG Infrastructure Project; Pacific Basin Information Node (PBIN) of the U.S. National Biological Information Infrastructure (NBII); Landcare Research (New Zealand); Bishop Museum, Honolulu (BPBM)

5.2. LSID and TCS deployment in the Catalogue of Life

Richard John White, Andrew C Jones, Ewen R Orme
Cardiff University

This paper describes a project to add support for Life Sciences Identifiers (LSIDs) and the Taxon Concept Schema (TCS) to the Annual and Dynamic Checklists assembled and delivered by the Catalogue of Life (CoL) partners, Species 2000 and ITIS. We plan to improve the compatibility of the protocols and public software interfaces used by Species 2000 with TDWG standards. We wish to increase the usefulness of the CoL to users, including GBIF, by improving the CoL's compatibility with other biodiversity tools, by supplying its information to clients expressed as

taxon concepts, and by enhancing interoperability between data providers and consumers by means of LSIDs referring to these concepts. It is hoped this will increase the use of TDWG standards, accelerate LSID deployment and the uptake of TCS, assist providers and users to ascribe data unambiguously to specified taxon concepts, and speed the growth of shared biodiversity data resources.

At Cardiff University we are investigating approaches for adding LSID and TCS support to the CoL and implementing them in evaluation versions of its systems. We have implemented a new prototype of the Annual Checklist which issues LSIDs for taxon concepts and established a resolution service to support the use of these LSIDs by giving provisional RDF/TCS responses generated from the Annual Checklist.

We are developing modified Spice protocols and a new Spice software prototype to provide LSIDs and TCS data in response to Web Service requests and to receive any name or taxon concept LSIDs from data providers. A new version of one of the data providers is being implemented for this purpose. We will develop a validation tool to check that the data and responses are valid, correctly structured and internally consistent. We plan to complete the project by the end of December 2007.

The Species 2000 Secretariat in Reading is assisting in this project. Its responsibilities are to survey the needs, capabilities and preferences of data providers and users in the light of these demonstration systems; to deploy the enhanced Spice software in the CoL global and European regional hubs; to use the validation tool and other means to perform testing and quality assurance of the data served; and to assist the CoL partners to agree a plan for the introduction of LSIDs.

The updated Spice protocol, documentation and enhanced Spice software will be available for use by other projects to build species information systems for their own purposes and to create regional hubs which can be linked to the CoL, both to enhance its usefulness in those regions and to help set up new global data providers.

Planning and carrying out this project has raised a number of interesting questions, some to be resolved during the project, others for wider consideration and future research. They include the choice of which kinds of entity will be identified by LSIDs (including names as well as taxon concepts), how users (human or software) will obtain LSIDs for entities of interest, how any GUIDs (not necessarily LSIDs) that data providers supply will be propagated through the CoL, users' expectations concerning tasks that LSIDs might assist, including navigating the taxonomic hierarchy and linking data to taxa, and the role of CoL LSIDs in building the biodiversity information systems of the future.

Further information about this project and its progress, updated periodically, is at http://spice.cs.cf.ac.uk/lsid/

Support is acknowledged from: TDWG Infrastructure Project

5.3. An LSID authority for specimens and an LSID browsing client

Kevin James Richards

Landcare Research

The requirements and use-cases for globally unique identifiers (GUIDs) have been developed by the TDWG community over the last 18 months. Use-cases include:

- • unique and persistent identification of taxon name and specimen data;
- • linking specific specimen records to accepted taxonomic names; and

- • detection of duplicate records.

After careful investigation of several identification schemes, the TDWG Globally Unique Identifiers group (TDWG GUID) endorsed the use of Life Science IDentifiers (LSIDs) for use in biodiversity information applications.

LSID resolvers are Internet services that return the data and metadata associated LSID to a requester. Resolvers have been set up mainly to process taxonomic name data. Important data types such as specimens have been neglected. It is therefore important to examine and test the use of LSIDs and related technologies within the context of specimen data.

'Herb IMI' is a collection of over 300,000 fungal specimens from the International Mycological Institute (IMI). The Herb IMI database contain fungus/host identification data and these records also have corresponding LSIDs for the fungal names in Index Fungorum and plant names in the International Plant Names Index (IPNI). Both these global nomenclators have Taxonomic Names LSID resolvers in place. This combination made Herb IMI a candidate for testing LSIDs and related technologies.

We developed an LSID resolver for the Herb IMI collection and a tool for demonstrating LSID related technologies such as the use of Resource Description Framework (RDF). RDF is a language in which entities are modelled with subject-predicate-statements known as "triples". Associated protocols enable a user to query sets of these triples and to infer relationships between entities. The tool that we have developed works with the specimen LSID resolver and TDWG's LSID vocabularies to display, browse, store and query the RDF associated with LSIDs.

This talk presents the processes, problems and outcomes of implementing LSIDs and RDF with the Herb IMI specimen database. The browsing tool developed for this project will also be demonstrated.

Support is acknowledged from: TDWG Infrastructure Project, Landcare Research

5.4. LSID policy and implementation in Australia

Greg Whitbread, Alex R. Chapman, Ben Richardson

Australian National Botanic Gardens

In April 2007 a 2-day workshop of representative of Australian museums and herbaria was held in Canberra, with TDWG assistance, to develop recommendations for a policy to apply to adoption of Life Science Identifiers (LSIDs) within the Australasian biodiversity federation. This meeting established the business case for LSIDs and guidelines and a roadmap for LSID implementation by and for local data providers and biodiversity informatics networks (http://www.tdwg.org/fileadmin/subgroups/guid/LSID_policy_workshop_Report_Canberra.pdf).

The workshop generated recommendations for the delegation of responsibility for allocation, persistence and resolution of LSIDs within the Australian biodiversity federation and drafted a work plan for our implementation of LSID technology.

Progress against these recommendations however has not been good. Resources are limited and the integration of LSID technology into an existing biodiversity information network is not without issues. There are elements in our LSID implementation plan that require more careful consideration: ambiguity within the classes of information identified for LSID assignment; the role of LSIDs in version control and the discovery of duplication; resolution; and metadata standards and access to data in appropriate formats. A more detailed specification, beyond best practice, for the form and function of LSIDs within the biodiversity informatics context is still required.

Support is acknowledged from: TDWG Infrastructure Project, Australian National Botanic Gardens, Council of Heads of Australian Herbaria (CHAH)

5.5. LSID Mashup

Daniel Miranker

University of Texas at Austin

Morphster* is a productivity tool for annotating specimen images and organizing the features into character state matrices suitable for phylogenetic reconstruction. Central to the architecture is distributed data integration where the data are tagged with global unique identifies (GUID); usually a life-science identifier (LSID). Source data for Morphster includes certain image databases, records from the uBio Taxonomic Name Server and Nomina Anatomica in the form of OBO ontologies. Each of these data sources associates a GUID with each record.

Persistent data records created by Morphster are tagged with LSIDs and made available per the protocol. For example, character definitions, character state definitions and the assignment of states to specimens are all separate records that may need to be archived and/or reused and are made uniquely identifiable. These records themselves reference the source images, the taxon and, usually, a field from the Nomina Anatomica. Thus, when resolving a Morphster LSID, the data returned will include a number of additional LSIDs. It is anticipated that Treebase II will store LSIDs in addition to encoded character states. The result will be a distributed data structure enabling on-line access to the complete provenance of a morphological phylogenetic study.

*The project (see http://www.morphster.org) is a collaboration with Julian Humphries and Timothy Rowe, Jackson School of Geology, University of Texas at Austin.

Support is acknowledged from: NSF, IIS:0531767

Session 6. Enabling Technologies: Protocols

6.1. TapirLink: Facilitating the transition to TAPIR

Renato De Giovanni

TapirLink is a free and open source data provider tool which implements the TAPIR protocol. It is based on the earlier DiGIR PHP provider, which is used by many institutions around the world to serve a total of more than 100 million specimen records. TapirLink has been designed to be as simple to use as the DiGIR PHP provider and to enable rapid and seamless migration of existing DiGIR providers to the TAPIR protocol.

TapirLink is a general-purpose tool and can be used to serve other classes of data as well as biological collections data. It supports most of the advanced features of the TAPIR protocol, including all TAPIR operations, searches using complex filters and flexible output models (for example KML, RSS2, DarwinCore 1.4 application schema, ABCD 2.06 and TDWG Ontology RDF).

Additional features include the ability to import configuration files from the DiGIR PHP provider, a user interface for UDDI registration, a configurable LSID resolver and the option to associate XSLT stylesheets with the XML responses to present the data in a human-readable form in Web browsers.

TapirLink allows data providers to participate in TAPIR networks or simply to offer a Web Service interface to their data. This presentation will describe the TapirLink software, showing the installation requirements, configuration details and main features of the tool.

Support is acknowledged from: TDWG Infrastructure Project

6.2. RDF over TAPIR

Roger Hyam

TDWG Infrastructure Project

The TDWG standards architecture relies on the melding together of two technologies that are often thought to be antagonistic: the Resource Definition Framework (RDF) and XML Schema.

RDF is based on a modelling language that describes everything in terms of subject-predicate-object statements (known as triples). This may be familiar from formal logic. RDF can be serialised in many ways. One of those ways is as XML.

XML Schema is a language for defining XML document structures. It is possible to define an XML document structure using XML Schema so that the resulting documents are valid serialisations of RDF.

TAPIR is a data exchange protocol designed to pass XML messages. The output from a TAPIR data provider is described using XML Schema. TAPIR knows nothing about RDF but by using XML Schemas that define RDF instance documents it is possible for a TAPIR data provider to behave as an RDF data source. This is demonstrated using the TapirLink provider software.

One of the strengths of the TAPIR protocol is that it allows the definition of custom response types (output models). This can act as a mapping point between conceptual schemas. It should therefore be possible to map other TAPIR concepts into RDF that uses the TDWG ontology.

This is demonstrated using data sources mapped to DarwinCore. It should also be possible to map any TAPIR data source to generic RDF.

There are a series of limitations to these approaches. Defining RDF instance data using XML Schema is not ideal because it is not possible to control the use of attributes of elements according to whether the element has content and thus prevent the simultaneous occurrence of an rdf:resource attribute and embedded content, which would be illegal. This is largely overcome in the demonstrations because it is known a priori whether a value is a literal or resource link. XML Schema is awkward to use when there are many namespaces in the instance document. Current examples use around ten separate XML Schema documents. This could become a performance issue in the future and imposes an implementation burden on TAPIR wrapper software. The current examples make use of the TapirLink provider software which does not implement complex internal data structures, only 'flat' tables. The PyWrapper TAPIR provider has been shown to support RDF in initial tests but not tested with the current examples.

Support is acknowledged from: The Gordon and Betty Moore Foundation

6.3. TAPIR networks in Australia's Virtual Herbarium and the Atlas of Living Australia

Greg Whitbread[1], Shunde Zhang[2], Paul Coddington[2]

[1] Australian National Botanic Gardens, [2] University of Adelaide

The first, and currently the major, iteration of Australia's Virtual Herbarium (AVH) uses a very simple protocol designed for a single task, to assemble partial HISPID documents from a number of providers and display species occurrence on a map. It is web-based, easy to implement, fully distributed, and praised and lamented by the community it serves.

AVH2.0 will accommodate full data interchange between Australian Herbaria and enable development of products to meet increased local expectations and support provider participation in global markets for biodiversity information. The story is: a network based on TDWG standards.

Development of the AVH2.0 portal has been completed, using Java. The full AVH1.0 functionality has been enhanced to interrogate and deliver HISPID, ABCD and Darwin Core, and to offer full indexing of distributed BioCASE and AVH1.0 providers, interfaces supporting pluggable services, and instance replication. However, schema support for AVH functionality and full provider compliance is yet to be achieved. In practice, deployment has proven difficult with technical, financial and social issues all presenting barriers to successful implementation. TAPIR integration using TAPIRUS and PyWrapper may be a solution to these problems.

TAPIRUS, the TAPIR Unit Seeker, is a Java library providing a programming interface to the underlying XML protocols for entering queries and for parsing, aggregating and post-processing result sets. TAPIRUS replaces the BioCASE UnitLoader. It supports both the BioCASE and TAPIR protocols, simplifies indexing, supports protocol extension, provides better performance and improved memory management.

The AVH is also a component of the Atlas of Living Australia (ALA). The ALA is a significant new initiative modelled on the AVH and related biodiversity informatics projects. It is designed to provide a national information infrastructure supporting biodiversity science and to establish a sustainable architecture for biodiversity informatics in Australia. Without the foundation of the pioneering work and standards of TDWG, the ALA would not be possible. The first breaths of the ALA will undoubtedly arise from a network of TAPIR providers and an open evolution will contribute to the future for TDWG process, protocols and standards.

Support is acknowledged from: Australian National Botanic Gardens, Centre for Plant Biodiversity Research, Council of Heads of Australian Herbaria (CHAH), TDWG

6.4. Checklist Provider Tool: a GBIF Application for Sharing Taxonomic Checklists Using TAPIR and TCS

Wouter Addink, Jorrit van Hertum

ETI BioInformatics

There is currently no simple way to connect nomenclatural and taxonomic resources to the Global Biodiversity Information Network (GBIF) network. The Taxon Concept Schema (TCS) standard was developed within Biodiversity Information Standards (TDWG) to make the exchange of taxonomic data possible. Providers need easy-to-use tools to connect such data to the GBIF network and to other networks such as the Catalogue of Life using TCS.

GBIF commissioned the development of a Checklist Provider Tool for sharing datasets held in tab-delimited files or Excel spreadsheets. This tool is scheduled to be made available at the GBIF GB14 meeting in October 2007 in Amsterdam.

The Checklist Provider Tool uses a MySQL relational database to store data in a TCS-compliant format. Data can be imported into the database through web forms written in PHP. The tool includes a pre-configured TAPIR-compliant access point which connects directly to the database and will facilitate connection to the GBIF network and to other networks. The access point will be based on the TapirLink PHP implementation of the TAPIR protocol (the same implementation is in use for sharing occurrence data using Darwin Core). This TAPIR access point will expose the taxonomic checklist data in TCS format.

The Checklist Provider Tool will be available for download as open source software at the GBIF website. In addition GBIF is considering using the tool to host small data sets directly at GBIF without data providers needing to install the software locally.

Support is acknowledged from: GBIF

6.5. Shibboleth, a potential security framework for the TDWG architecture

Lutz Suhrbier[1], Andreas Kohlbecker[2]

[1] Institut für Informatik - AG Netzbasierte Informationssysteme, Freie Universität Berlin, [2] Biodiversity Informatics, Botanic Garden & Botanical Museum Berlin-Dahlem

Shibboleth is a project of the Internet2 Middleware Initiative (http://middleware.internet2.edu/). It provides an architecture and an open-source implementation for a federated, identity-based authentication and authorization infrastructure. Groups of organisations or projects may develop a federation by agreeing on common security policies and practices. They can use SAML/Shibboleth protocols to manage single sign-on across domains. This removes the need for content providers to maintain usernames and passwords.

Authorisation is instead based on trusted user attributes supplied by trusted Identity providers (IdPs) and consumed by service providers (SPs) which then gate access to secure content.

We will introduce the main concepts of the Shibboleth architecture. In addition, we outline potential benefits for the entire TDWG architecture and present the current approach to federation within the EDIT project.

Session 7. Models for Integrating TDWG: Species Profile Model

7.1. Main aspects of the Species Profile Model and the TDWG architecture

Andreas Kohlbecker, Markus Döring, Andreas Müller

Botanic Garden & Botanical Museum Berlin-Dahlem

The Species Profile Model (SPM) was created at the GBIF Species Model Workshop in Copenhagen, on April 16-18, 2007, to support the retrieval and integration of species data.

The SPM describes a root element referring to a TCS (Taxon Concept Schema) taxon concept, under which one or more facts (InfoItems) about the species can be listed. Each fact belongs to a category drawn from a set of controlled terms.

The aim of the SPM was to provide a flexible data model which allows for naive implementations, in which InfoItems can be freely tagged with terms from controlled vocabularies independent of one another.

However, the free tagging approach leads to problems when the SPM is used for data exchange, *e.g.*, when using the SPM with TAPIR. The document resulting from a TAPIR request would consist solely of InfoItem elements having no child elements. To get the category for an InfoItem, TAPIR would need to send additional requests to retrieve the subelement category for each of these. Therefore, free tagging was replaced by a subclassing concept, in which a set of specialised InfoItem classes, each inheriting from an InfoItem super class, are created for each fact category. Documents now contain instances of these InfoItem subclasses, each named according to the category of data it contains.

At this time during the evolution of the SPM, it is crucial to shed light on how and whether the current version of SPM fits the needs of the use cases in scope. Identifying possible problems in a timely fashion will help us decide where to go from here.

7.2. Species Profile Model: Data integration lessons from GBIF

Donald Hobern

Global Biodiversity Information Facility

The Species Profile Model (SPM) is being developed as a standard to simplify integration of species information from multiple sources and to maximise its usefulness and reusability.

The Global Biodiversity Information Facility has spent the last five years working to integrate biodiversity data from a wide range of different resources and has learned several lessons which are likely to be relevant to those developing the Species Profile Model and to those planning to build species information networks.

1. Any data exchange model should be optimised for the key use cases. If the intended applications for the data are well understood, it is relatively easy to determine which elements in a model should be mandatory and which elements will require tightly controlled vocabularies. This allows attention to be focused on the most important areas and perhaps for less critical aspects to be deferred to future versions of the model.

2. Despite the difficulties in gaining consensus, some elements in a data model require strictly controlled vocabularies to ensure that applications can discover which data records are genuinely of interest for a given purpose. Tools are required to help data providers to map their existing data to such vocabularies.
3. Metadata require as much planning and design as the fields regarded as data. Associating appropriate metadata with a data resource makes it much easier for applications to select requested data records and to interpret them correctly.
4. Data models may atomise the same fundamental information to different degrees. The key driver has usually been to ensure that different data providers are all able to map their existing data into the model. For many purposes the existence of alternative ways to represent the same information will not matter. However it is important to consider the effect of the variation on applications that seek to consume and analyse these data. In the long run less effort may be involved in modifying the source data than in building client applications which can handle this extra complexity.

7.3. SPM from an SDD perspective: Generality and extensibility

Gregor Hagedorn

Federal Biological Research Center for Agriculture and Forestry (BBA)

The current Species Profile Model (SPM) is urgently required because of the need to integrate descriptive data with the new TDWG models based on semantic web technologies.

The Structure of Descriptive Data (SDD) model (standardized by TDWG in 2005) is not, so far, based on semantic web technologies, but has been developed explicitly to integrate a wide range of data types, from categorical ("character states") to quantitative (including statistical measures like average, variance, sample size, etc.). It offers methods to deal with thousands of characters and other terms, has been designed to support federated systems, and contains an extensible data type system.

Most importantly, SDD succeeds in integrating: (a) natural language markup (ranging from breakdown into “subject fields”, as considered by SPM, to detailed markup of concepts, characters and values in legacy literature); (b) coded descriptions (matrices as known from the DELTA or NEXUS encoding systems for taxonomic descriptions); and (c) original sample data.

SPM, while aiming for simplicity, has already started to become complicated, expanding from natural language text into structured data such as categorical and quantitative measurements. With respect to the sequence of data type and semantics, the approach taken by SPM seems to be the opposite of SDD. It is difficult to judge how complex the model will become as it develops further, and how well it will scale if dealing with thousands of descriptive concepts.

Important topics for investigation include:

(a) how SPM can be extended to support multiple descriptions of different scope per taxon (*e.g.*, geographic, seasonal, stage-specific);

(b) how an extensibility of categorical values through modifiers (“perhaps”, “frequently”, “at tip”, “in winter”) or free-form text comments can be added;

(c) how a rich vocabulary of statistical measures can best be introduced; and

(d) how issues of free form text markup (especially if arranged differently than the current SPM categories) and sequence versus set data can be addressed.

It is hoped that a merger of the requirements of SDD and SPM will be possible to maintain a common platform for future development.

7.4. Coming to Terms with SPM

Robert A. Morris
UMASS-Boston

The Species Profile Model (SPM) is a proposed ontology, defined in terms of the emerging new TDWG ontology, centered architecture based on the Resource Description Framework (RDF) language of the World Wide Web Consortium. Although RDF has a representation, one of several, in XML, it and related languages attempt to capture semantics, not only syntax, of data. By contrast, XML-Schema, the principal language for constraining XML for data validation purposes, only constrains syntax. RDF does so by making it easier to express in standard ways the relationships among concepts meant to be addressed by particular data and allow such data to be compared and integrated without requiring transformation into a particular syntactic form. Many intuitions arising from syntax-only constraint languages like XML-Schema give little or no insight into semantic questions (*e.g.*, when is an attribute value, or even its applicability, inherited from objects to subobjects, when are two attributes mutually exclusive, etc.). Consequently, the learning curve for RDF technology can be steep, all the more so without good tools for producing, editing, and visualizing information expressed in RDF and related technologies. This presentation will describe and demonstrate two such tools, Protege and Altova SemanticWorks.

Although the ultimate point is to generate SPM programmatically, the use of tools like Protege and Altova SemanticWorks can ease the RDF learning burden, and of course serve as a debugging aid, if properly configured. Both tools let users interactively tinker with SPM instances in ways that help one better understand the SPM design and clarify the use of it.

No species profile will live in isolation, so one question these tools help you consider is whether to attempt to design an ontology (described in the RDF-based Web Ontology Language, OWL) for descriptions of your group of interest, or whether to be content simply with RDF consistent with the TDWG architecture. To a certain extent, this speaks to the question of whether your goal is to support machine reasoning about the species described by your SPM, or simply to support mashups as a means of integrating data for human analysis. Time permitting, I will discuss how such RDF editors can be used as visualization and debugging aids for RDF produced by screen scraping tools generated by the MIT SIMILE suite (http://simile.mit.edu/wiki/Main_Page).

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

Session 8. Models for Integrating TDWG: Literature Model

8.1. Linking Bibliographic Data to Library Content

Julius Welby

Natural History Museum

The European Distributed Institute of Taxonomy (EDIT) Work package 5.3 will provide bibliographic and literature discovery tools which will help reduce literature related bottlenecks which can hinder the progress of day-to-day taxonomic research.

In response to discussions and a broader requirements gathering exercise we are planning a website supporting federated searching of taxonomically relevant data sources accessible via standard protocols (*e.g.*, Z39.50). Data sources include library catalogues of EDIT partners and others, the Biodiversity Heritage Library (BHL), and electronic journals. Users will be able to browse through the results of their searches to see links to useful resources and metadata, and there will be a link to the original content where this is available, via an OpenURL service.

Where full text content is not available, registered users will be able to nominate non-copyright texts which they would like to use, and these nominations will be made available to the various digitisation projects currently in progress at institutions around the world.

The Virtual Taxonomic Library (ViTaL) will also provide a place for taxonomists to view and search aggregated bibliographic references harvested from a number of reference management services and web sites. References will benefit from the same linking technology used for the search results.

The team will share their overall vision for the site, outline some of the practical and technical challenges, and give a brief overview of the technology used to provide linking from search results.

Support is acknowledged from: EDIT, Natural History Museum

8.2. Use cases from taxonomists, conservationists, and others

Cynthia Sims Parr[1], Christopher Lyal[2]

[1] Information International Associates, [2] The Natural History Museum, London

Digital versions of taxonomic literature are increasingly readily available, but significant work remains to make this literature fully accessible to users and delivered in a manner best fitting their needs. As part of the INOTAXA project, we conducted interviews with a wide range of potential users and developed a collection of use cases that may guide the development of systems to provide access to the taxonomic literature. The use cases span the range from those closest to the originators of taxonomic literature (systematists gathering material for taxonomic revision, often including phylogenetic analysis) to those that will extend the impact of the literature (*e.g.*, ecologists harvesting species associations for modelling of interactions). They fall into eight broad categories:

1. general exploration of source material;
2. taxonomy (*e.g.*, preparing revisions and checklists, conducting phylogenetic analyses);

3. specialized taxonomy-related (*e.g.*, preparing author catalogues, itineraries or histories of expeditions);
4. identification (*e.g.*, identifying specimens for taxonomic, pest control, surveys, or other purposes);
5. extra-taxonomy (*e.g.*, harvesting data for ecological, morphological, or character evolution studies);
6. policy decision-making;
7. data maintenance (*e.g.*, correcting information in databases); and
8. web services.

While there are many common functions required by these use cases, for example, searching and browsing by taxonomic name or geographic location, sequences of tasks and desired results often differ. Some use cases, particularly those beyond more traditional taxonomy, involve users who are interested only in specific parts of the literature. As they are likely to be less familiar with taxonomic literature and how to search for their exact needs, they may require more support for browsing or in assessing data completeness and fitness for use. It will be a challenge to design schemas and interfaces which support multiple use cases well without overwhelming users. The practices of many users are currently constrained by print formats. Future systems can be freed from these constraints and support database-oriented rather than document-oriented uses of literature. Such a perspective will foster closer integration of the literature with other kinds of biodiversity information such as specimen and nomenclatural databases. This will not only allow published biodiversity data to be used much more extensively and in novel ways, it will open the door to more flexible 'publication' and delivery of taxonomic information and data in the future.

Support is acknowledged from: The Atherton Seidell Fund of the Smithsonian Institution

8.3. Progress in making literature easily accessible: schemas and marking up

Terry Catapano[1], Anna Weitzman[2]

[1] Columbia University, [2] Smithsonian Institution

An important component of making biodiversity content available is the vast quantity of taxonomic information in printed form. Even 300+ year-old works remain relevant to taxonomy. Taxonomists have traditionally accessed this information by reading and taking notes, which are later incorporated into subsequent treatments. Similar, though more widespread, access exists for images of pages on the Web (*i.e.*, the user still needs to know for what and where to look). Another step forward is to reproduce the printed information as machine-readable text. Even this still leaves the task of distinguishing relevant information in the potentially vast quantities of data. In order to make data in literature fully accessible, it must be encoded, have proper metadata added, and be made available for searching, linking and processing. Two projects, taxonX/GoldenGate (GG) and taXMLit/INOTAXA are attempting to tackle this task.

The aim of the taxonX schema is to provide a minimally sufficient XML tagset to identify and delineate taxonomic treatments and their significant components, particularly scientific names, geographic names, bibliographic citations, and descriptions. Once encoded in taxonX, the treatment and its associated data can be more readily extracted and incorporated into other databases as well as accessed and integrated into external resources. Owing to the diverse heterogenous forms of taxonomic treatments, the schema design is loose and flexible. Similarly,

the content of the data itself requires normalization in order to be useful within existing and future digital infrastructures.

Developed independently, but alongside taxonX, GG contains tools for the semi-automatic markup of scientific names and treatment boundaries, and work proceeds on similar tools for bibliographic citations and geographic names. Tools to assist in identification of normalization of descriptive data are possible, but more difficult. GG can input a cleaned OCR (optical character recognition) file in xml, html, or text format and export a taxonX instance.

taXMLit is another schema for tagging taxonomic literature. Unlike taxonX, it is deliberately a fairly complete representation of data within the literature and thus is a complex schema. Taxonomic literature has a limited number of 'kinds' of information. These may be recognized in several ways, including using GG. Using xml text with those designated, *e.g.*, a taxonX instance, another set of tools is underway to further parse and normalize data from kinds of paragraphs most likely to be needed by taxonomists (*e.g.*, taxon heading, synonymy, specimen citations). As different formats of these kinds of paragraphs are identified, a library of tools will be built. Artificial intelligence should be able to select which tool is needed for each paragraph.

We believe experiences gained in the development of taxonX and taXMLit can inform future efforts to establish TDWG standard(s) for taxonomic literature. Two approaches to this task might be considered. First, the development be of a Standard, not necessarily a Schema. A core Vocabulary could be developed, with a number of different expressions, each ontologically harmonic, but in forms optimal for particular processes and uses. Secondly, the NLM/NCBI Journal Archiving DTD (a Document Type Definition defines the allowed building blocks of an XML document) should be investigated as one of the forms for expression of a TDWG Literature Standard. The NLM DTD enjoys strong and committed maintenance and has been adopted widely. It is deigned to be modular, with domain specific elements added to the base generic markup elements.

Support is acknowledged from: US National Science Foundation; Atherton Seidell Fund of the Smithsonian Institution

8.4. Literature & interoperability: a working example using Ants

Donat Agosti[1], Terry Catapano, Guido Sautter[2]

[1] plazi.org /American Museum of Natural History, [2] University of Karlsruhe

Print is still the main medium to communicate taxonomic results. Traditionally printed taxonomic publications may include all the information (data, analysis, conclusions) needed to understand new research results. This system has been very successful, surviving almost a quarter of a millennium. Even today with widespread technologies for electronic distribution, the basic means of taxonomic communication has not altered, not yet taking full advantage of these technologies.

An understanding of the successful print model of systematics should orient efforts in the shift to a new digital knowledge infrastructure. In essence, a taxonomic treatment is the amalgamation in a single record of information we consider relevant to describe our taxa, including often not just the inferred hypotheses but also the underlying data. If sufficiently detailed, the latter can be identified, extracted, and populate dedicated databases on specimens, nomenclature or bibliographic citations.

Our German DFG / US NSF funded digital library project has been built upon this premise. In order to digitally represent the significant components of systematics literature, the XML schema TaxonX (http://taxonx.org) has been developed. The prospect of encoding the tens of millions of

printed pages inspired the development of dedicated mark-up software (GoldenGATE) enabling the semi-automatic mark-up of suitably clean OCRed texts.

But even this process is still time consuming and dependent on the involvement of experts. As a result, a dedicated server, plazi.org (http://plazi.org) will be launched at the TDWG meeting that will allow the community not only to retrieve the respective documents, but actively participate in the mark-up process, and to be able to retrieve digital versions of individual treatments (descriptions of taxa). Openly available services like iSpecies or EDIT's scratchpads will be able to access the treatments and incorporate them in their mash-ups or as seeds for scratchpads. For the legacy publications to become truly interoperable, TaxonX allows the inclusion of references to identifiers in the increasing number of dedicated databases (eg GBIF; bibliographic references). To bridge the gap between the idea and implementation, unique identifiers for ant names will be retrieved from the Hymenoptera Name Server (including >200K names, including all ant names) and expressed as LSIDs. For literature, handles are retrieved via bioguid.org from plazi.org's handle server, an integral part of DSpace, the respository of all the digitized legacy ant publication used to administer all the publications.

Although plazi.org currently concentrates on ants and legacy publications, it can in principle provide its services for any taxon. This all comes at high costs. What is needed in future are dedicated databases (specimens, character, names, bibliographies, etc.), unique identifiers, a program like LUCID to machine generate both a human readable text as well as the underlying xml mark up, and for publishers to integrate taxonomic specific annotations alongside a human readable text version seen in taxonomic publications.

Support is acknowledged from: NSF, DFG

8.5. Taxonomic Literature: What Next?

Anna Weitzman[1], Christopher Lyal[2]

[1] Smithsonian Institution, [2] Natural History Museum, London

Calls continue for agreed standards for taxonomic literature. Earlier work identified three key levels: microcitations, metadata and content. Broad agreement has been reached on the first of these, although the standard needs finalisation, including a decision on LSIDs. A standard for metadata is still to be agreed, and must accommodate both librarian and taxonomist needs; this is becoming urgent with the development of the Biodiversity Heritage Library (BHL) and the need to access its content. The most complex standard is for complete taxonomic literature content. Taxonomic literature accommodates many different data and information types, including those subject to existing TDWG standards. This raises the possibility of it serving as a test bed for full interoperability of taxonomic data and linking of TDWG standards. However, expression of such data and information in literature sources often differs from the source material examined by other TDWG groups. In developing a standard the requirements of meeting several goals interact: a) interoperability across data and information types; b) maximising cost-effective access to and display of information and data in a manner for the user; c) cost-effective mark-up in agreed formats. Alternative routes to interoperability include a) making different schemas congruent while reflecting properties of the data sources, and b) embedding different modular schemas within a larger container. The degree of atomisation of the content will impact both on the breadth of user needs that can be met cost-effectively, and issues of mark-up. Once interoperability is achieved, user-friendly navigation tools for the information universe thus created, and delivery of output and choices expected by different users, become issues that must be addressed.

Support is acknowledged from: The Atherton Seidell Fund of the Smithsonian Institution

Session 9. Models for Integrating TDWG: Spatial Model

9.1. Species distribution modelling and phylogenetics

Stephen Andrew Smith

Yale University

Recent advances in niche modeling methods allow us to more accurately predict the ranges of species. These models have been used extensively in building low resolution maps from georeferenced museum specimens as well as predicting future movements of invasive species, however, their full potential as a tool for evolutionary biology has not been adequately explored. With the additional development of high quality world climate layers it is possible to reconstruct not only the ancestral climate envelopes for species, but also the rates of evolution in these climatic variables. I will focus on a clade of 19 plant species endemic to western North America (*Oenothera* sect. *Anogra* and *Kleinia*, Onagraceae). I have used high resolution (1 km^2) climate data, a maximum entropy method to model the climatic tolerances of species, and phylogenies inferred from DNA sequence data. I will show a reconstruction of the evolution of climatic tolerances using standard continuous models and focus on the rate at which climatic tolerances evolve, including tests for rate heterogeneity. This approach allows me to investigate how climatic niches evolve over time scales relevant to macroevolutionary biologists. These results provide examples of both climatic niche conservation and evolution.

Support is acknowledged from: NSF

9.2. Lifemapper: Using and Creating Geospatial Data and Open Source Tools for the Biological Community

Aimee Stewart, C.J. Grady, James Beach

University of Kansas

The open source project Lifemapper 2 creates an archive of species predicted habitat maps and other spatial distribution information. Lifemapper 2 uses museum specimen data archived by the Global Biodiversity Information Facility (GBIF) and accessed through their web services, current and future International Panel on Climate Change (IPCC) climate scenarios, the openModeller niche modeling library, and a 64-node compute cluster. Applied studies using the resulting data can predict the impacts of climate change, loss of biodiversity, spread of invasive species, and emerging diseases.

Lifemapper 2 uses various open source libraries and applications to provide information to the biological community and general public. The Geospatial Data Abstraction Library (GDAL), PostgreSQL, PostGIS, and Mapserver are used in the pipeline of the system, while the cluster uses Sun Grid Engine and openModeller to create the niche models. The website provides archive browsing and data query and download, while web services provide programmatic access.

The services provided by this project will provide inputs for more user-friendly software tools for biological collection data integration and analysis. This will provide a foundation for a new pluggable, extensible architecture that will tie together the services, functions and methods of different applications. The end result will improve quality of data collected and provide support to researchers in studying species' actual and potential distributions.

Support is acknowledged from: National Science Foundation

9.3. A pilot project for biodiversity and climate change interoperability in the GEOSS framework

Stefano Nativi[1], Paolo Mazzetti[1], Lorenzo Bigagli[1], Valerio Angelini[1], Enrico Boldrini[1], Éamonn Ó Tuama[2], Hannu Saarenmaa[3], Jeremy Kerr[4], Siri Jodha Singh Khalsa[5]

[1] Italian National Research Council – IMAA and Univ. of Florence, [2] GBIF Secretariat, [3] University of Helsinki, [4] University of Ottawa, [5] IEEE and Univ. of Colorado

The Global Biodiversity Interoperability Framework (GBIF) Interoperability Process Pilot Project (IP3) addresses two Societal Benefit Areas: Biodiversity and Climate and is developed within the framework of the Global Earth Observation System of Systems (GEOSS). GEOSS is an international initiative to combine new and existing hardware and software for the purposes of supplying earth observation data and information at no cost.

The aim of IP3 is the implementation of a GEOSS Architecture through the development of relevant scenarios that draw on data and information exchange from a series of interconnected systems.

The focus of GBIF IP3 is modeling the impact of climate change on species distribution. To achieve this, heterogeneous data resources (*e.g.*, biodiversity, climatological and environmental resources) and processing services are required to interoperate by using a Service Oriented Architecture (SOA) approach.

Through the pilot some available service-based components were selected, some artifacts developed and special arrangements to facilitate interoperability demonstrated, all for specific use scenarios.

These main components are described below.

Biodiversity occurrences are discovered and accessed through web services published by the GBIF Data Portal (http://data.gbif.org) and according to the TDWG Darwin Core standard format.

Climatological data are obtained from the NCAR GIS portal which provides web access to free global datasets of climate change scenarios. These data (spanning 50 years from 2000 to 2050) have been generated for the 4th Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) by the Community Climate System Model (CCSM). The datasets are processed to generate grid coverage and served through an OGC WCS 1.0 server.

GI-cat is a federated catalog providing a unique and consistent interface that enables the interrogation of biodiversity and climatological data resources. GI-cat exposes an OGC CS-W/ebRIM interface and is able to federate heterogeneous catalogs and access servers that implement international geospatial standards (*e.g.*, OGC OWS). In addition, GI-cat implements a mediation server, making it possible to federate non-standard servers (*e.g.*, THREDDS/OPenDAP servers) by specifying "special interoperability arrangements". A special interoperability arrangement was introduced for the GBIF portal services, consisting of the introduction of a formal mapping for the GBIF data model to the ISO 19115 core metadata profile, and the GI-cat to GBIF service protocols adaptation.

The component used for processing collected data and generating future projections is the OpenModeller, an open source Ecological Niche Modelling (ENM) framework. It is accessed through a Web Services interface based on the SOAP protocol.

An AJAX client was developed to implement a user friendly interface to OpenModeller functionalities, making them accessible by any web browser. With this tool, the user is guided through the process of discovering data (by submitting queries to GI-cat), accessing selected data

(through GBIF and WCS/NCAR data servers) and running ENM projections. Finally, the results are shown.

A first demonstration dealt with the Canadian butterfly species (Amblyscirtes vialis) and its response to climate change. This demonstration was presented in the most recent GEOSS workshops.

9.4. Advances at the OGC, and Opportunities for Harmonization with TDWG Standards and Models

Phillip C. Dibner

OGC Interoperability Institute (OGCii)

Several recent developments in technology and organization at the Open Geospatial Consortium (OGC) are highly relevant to the mission and objectives of the Biodiversity Information Standards (TDWG) community.

Most fundamentally, a new institute has been created: the OGC Interoperability Institute (OGCii). Whereas the OGC's mission is to create standards that support spatiotemporal processing, the OGCii was constituted to help make these standards accessible to the scientific research community through education, technical engagement, and collaboration in research programs. The OGCii also maintains a strong relationship with the OGC's Standards and Interoperability Programs, and enjoys continued access to OGC resources.

As of this writing, the OGCii has collaborated with several academic and research institutions in scientific proposals that span a variety of domains, including biodiversity informatics. These essentially independent efforts share a common theme: enabling the integration of datasets from different domains of knowledge by harmonizing the information models employed by their respective information communities.

Several OGC specifications in the final stages of the approval process are highly relevant to these and future efforts. Prominent among them is the Observations and Measurements (O&M) standard that has been a topic in several presentations to the TDWG membership (*e.g.*, P. Dibner, 2006, Proceedings of TDWG, "An integrative, standards-compliant framework for TDWG schemata and services"), and a key component in the TDWG/OGC domain modeling and harmonization workshop conducted in Edinburgh in June, 2006. Related standards, also near release, include the Sensor Observation Service (SOS), which serves Observation objects, and Sensor Modeling Language (SensorML), which describes sensing devices and data collection processes.

There has been a recent convergence of interest throughout the scientific community in standards for describing and sharing observations. The same formats and schemata need not be adopted by every discipline; it is sufficient if the information models they use are consistent, so that real-time conversion between them is feasible. In this capacity, the O&M and related standards offer the prospect of enabling seamless integration of BIS (TDWG) data into investigations and analyses that use the growing network of OGC service implementations.

Other developments at the OGC involve products and tools that are associated with the mass market. A particularly prominent example is the project to harmonize the KML language for describing geographic information with OGC specifications, and ultimately to release it as an OGC standard in its own right. Ultimately, this will enable the distribution of scientific data via powerful, popular, and freely available tools such as the Google Earth browser.

Perhaps more significant than these technical capabilities and projects is the growing web of relationships and common interests between the two organizations. The MoU executed between TDWG and the OGC in October 2006 has continued to spawn a variety of exchanges, including speakers at meetings, collaboration in modeling exercises and proposals, and incorporation of biodiversity data in research exercises and implementation pilots. The relationship is alive, well, and continuing to grow.

Support is acknowledged from: TDWG Infrastructure Project

9.5. The BiogeoSDI workshop: Demonstrating the use of TDWG and OGC standards together

Javier de la Torre[1], Tim Sutton[2], Bart Meganck[3], Dave Vieglais[4], Aimee Stewart[4], Peter Brewer[5], Renato de Giovanni[2]

[1] Imaste-IPS, [2] CRIA, [3] Africamuseum, [4] University of Kansas, [5] University of Reading

A week long workshop was held in Campinas, Brazil during the first week of April 2007. The focus of the workshop was to develop a test-bed web application to demonstrate the interoperability of digital data and services using open standards, with particular emphasis on geospatial, taxonomic and occurrence biodiversity data.

Two versions of a prototype web application were developed using PHP and Flex. The wizard style application leads the user through a defined sequence of steps in order to acquire sufficient data to create a niche model. The process includes taxonomic validation using the Catalogue of Life, search and retrieval of occurrence data using services such as the GBIF portal or WFS, selection of raster layers representing environmental data and modeling these data using the openModeller Web Service to create a probability surface that represents areas where a species is likely to occur.

The workshop highlighted how easy it is to rapidly create a feature rich application using open access to data, free software and open standards. The workshop also highlighted areas where further work is needed to effectively blend these services into a cohesive computing platform. Finally, suggestions were made for improving OGC and TDWG standards in a report that is available at (http://wiki.tdwg.org/twiki/bin/view/Geospatial/InteroperabilityWorkshop1). The prototype will be demonstrated and the issues arising will be discussed.

Support is acknowledged from: TDWG Infrastructure Project, CRIA, University of Kansas, Africanmuseum, IMASTE-IPS

Session 10. Models for Integrating TDWG: Descriptive Model

10.1. From Xper to Xper²: comments on twenty years of taxonomic applications with descriptive and identification tools

Régine Vignes Lebbe, Guillaume Dubus

Universite Pierre et Marie Curie, Paris 6

Xper and associated programs have now existed for twenty years. They are dedicated to managing taxonomic descriptions, providing interactive free-access identification, constructing keys and diagnoses and analysing, comparing and measuring similarities between descriptions.

During these twenty years, each new taxonomic application has suggested improvement of knowledge representation, management functionalities, user interface and taxonomic tools. For example, in the past a tool to automatically write descriptions as readable text was developed to publish (in two languages) the descriptions of phlebotomine sandflies of French Guiana. Then an HTML export was added to Xper² for a quick on-line distribution of a knowledge base. Another example: a tool to compute similarities between descriptions was developed, then this tool was used to complete the taxonomic forms constructed automatically from a knowledge base to suggest the most similar taxa and the risk of misidentification. Recently the import/export in spreadsheet format appears important for practical use during an application on Flora of France.

We will discuss the positive and negative points of such developments and the gap between taxonomists' needs and computer scientists' objectives.

http://lis.snv.jussieu.fr/newlis/?q=en/resources/softwares/cai/xper2
http://lully.snv.jussieu.fr/xperbotanica/

We thank all the Xper² users for their profitable comments on the software.

Support is acknowledged from: Xper² development is funded by the BioInfo 2002 grant of the CNRS and by the French Ministère for Research and New Technology (n°04L370 – Project Xper Botanica, 2005-2007).

10.2. GrassBase – integrating structured descriptions, taxonomy and content management

Kehan Harman

Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AB, U.K.

Kew's World Grass species descriptions have recently been made available over the internet, but management of this descriptive resource is still difficult. The descriptions consist of 11,000 species and 700 generic descriptions in DELTA format using a suite of 1090 mainly morphological characters. While the CSIRO DELTA software has supported this dataset well for many years, the lack of support or development of this software has led to the need for a more sustainable tool. Furthermore the associated nomenclature database is only available through the download of an MS Access database application. The World Grass Species project has been set up to help integrate these two resources and to make them available through a web portal. Combining an industry leading Content Management System (Drupal – http://drupal.org) with

existing TDWG standards to integrate and present this data facilitates the development process, and simplifies the extension of the tool using community developed modules.

Support is acknowledged from: Royal Botanic Gardens Kew, Cranfield University

10.3. Mechanisms for coordination and delivery of descriptive data and taxon profiles in the Australasian Biodiversity Federation

Alex R. Chapman

Western Australian Herbarium, DEC

A summary of the development and delivery of descriptive data within keys and taxon profile pages in Australia, with particular focus on 'FloraBase - the Western Australian Flora' as a case study for information assembly and coordination.

Structured taxonomic descriptive data standards enable the rigorous capture of taxon data at an atomic character level, for delivery in interactive identification, information retrieval and natural language descriptions.

Coordinating the capture of coded descriptive data for one taxonomic group becomes more complex with additional contributors. Integrating descriptive data across taxonomic groups requires an agreed vocabulary and definition of characters and states even when the morphology is non-homologous; once again coordination becomes more complex with increased variety of taxonomic groups.

While there are some notable large-scale collaborative descriptive projects scoring atomic character data across large geographic and taxonomic extents, there is more commonly a disjunction when existing data from various authoritative sources is assembled. In these cases a more generalised schema defining higher-level aggregated descriptive components can be used.

In Western Australia, with 3% of the world's vascular flora, a mixed strategy is maintained. Detailed coded data for the 1300 families and genera allows the publication of a comprehensive set of interactive keys and scientific descriptions from a single source. At the species level, some 13,000 short coded descriptions are maintained, as well as a growing set of free-text descriptions from online journals, archived floras or stand-alone projects.

Nationally, the Flora of Australia volumes have been marked up with XML schema identifying larger blocks of descriptive text. This schema may provide a common reference standard for marking up species-level descriptions from the many potential sources. With the upcoming 'Atlas of Living Australia' and global 'Encyclopedia of Life' projects, the identification of standard components on taxon profile pages and a mechanism for reliable tagging will be required.

Life Science Identifiers (LSIDs) can flag available elements for interrogation and retrieval, with the potential for automated sharing, re-use and re-assembly of authoritative content. This has the benefit of allowing the dwindling pool of specialists to focus on the development, publication and maintenance of content.

Support is acknowledged from: WA Department of Environment and Conservation; Global Biodiversity Information Facility

10.4. Using Automatically Extracted Information in Species Page Retrieval

Xiaoya Tang[1], P. Bryan Heidorn[2]

[1] Emporia State University, [2] University of Illinois

Users searching botanical texts online in currently available full-text indexes such as Google must accurately guess the vocabulary of the original author(s) to find the desired results. A large number of botanical volumes are available electronically, and many more are being made available through projects such as the Encyclopedia of Life and Biodiversity Heritage Library. However, current retrieval systems available for these collections are not able to interpret the specific information requests correctly and match them with appropriate documents. Author vocabulary often varies greatly from the user's search vocabulary. We will present a study which integrates text mining techniques into the full-text search process and automatically identifies selected plant morphological information from text to assist keyword-based retrieval. The technique could be expanded to other collections of documents.

An experiment involving users was conducted to evaluate this approach on the full-text of the Flora of North America (FNA). Thirty upper-level undergraduates and graduate students from two Illinois universities who had completed a course in botany were asked to identify ten herbarium specimens of trees of Illinois. The subjects used a full text search engine with an index of several volumes of FNA. The user search logs were used to identify the plant characteristics most frequently used by the students, independent of the usefulness of these terms for retrieving taxonomic treatments using full-text search. These characters were targeted for text extraction. A set of treatments were marked by hand to serve as training examples and a machine learning method was used to learn extraction patterns and these commonly used characters were mined from the 1637 treatments in the FNA. The accuracy of the extraction was between 60% and 100%, except for leaf shape and leaf arrangement information, which was around 50% and 30%, respectively, depending on the information type. In a new experiment one group of 12 subjects used a traditional full text search system while another group of 12 used full text plus pull-down menus and web forms that allowed them to search based on the machine extracted information. The experimental results indicate that the latter approach significantly improves keyword-based retrieval performance by allowing the users to complete more identification tasks successfully than when they had to generate their own search terms. It also increases users' satisfaction with the retrieval system.

10.5. Capturing structured data to facilitate web revisions

Dave Roberts[1], Julius Welby[1], Markus Döring[2]

[1] The Natural History Museum, [2] Botanischer Garten und Botanisches Museum Berlin-Dahlem

In order to write a taxonomic revision it is necessary for an author to assemble and consider the range of existing descriptions, bring them into a common framework (*i.e.*, standardisation) and consider how well they form delineated groups. In general the existing descriptions are in free-text blocks with associated nomenclatural and relationship information usually laid out in a structured (formatted) manner. The EU project EDIT has devised a general information-flow structure to guide the development of tools to assist taxonomists in their work and to bring the products of taxonomic effort more efficiently to the broader user community.

From a sociological perspective we consider it essential to design ways of working that mesh seamlessly with the way taxonomists work now. To that end we have investigated the natural language application GoldenGATE as a means to add structure to both the free-text descriptions and the formatted nomenclatural elements of both published and new work. The primary

intention is to capture content from manuscripts (word processor documents) rather than from published sources per se.

We will describe the information model that is guiding EDIT development and the advantage that structured data can offer in terms of increasing the efficiency of taxonomic workflow. Better tools to process taxonomic information are of significantly greater value if there is information to be processed. In other words, we need to establish a bank of structured content and demonstrate the benefits of working with structured data if we are to engage new users with this improved way of working. The goal is to motivate users to invest the effort required to understand and use structured data tools.

Support is acknowledged from: EU's Sixth Framework Programme: European Distributed Institute of Taxonomy (EDIT)

Session 11. Integrating Biodiversity Data

11.1. Removing Taxonomic Impediments: How the Encyclopedia of Life and Biodiversity Heritage Library projects can help

Graham Higley

The Natural History Museum, London

The Encyclopedia of Life (EOL) is a collaborative scientific effort led by the Atlas of Living Australia, Field Museum, Harvard University, Marine Biological Laboratory (Woods Hole), Missouri Botanical Garden, Smithsonian Institution, and Biodiversity Heritage Library (BHL), a consortium of natural history libraries[1]. Ultimately, the Encyclopedia of Life will provide an online database for all 1.8 million species known to live on Earth. When completed, www.eol.org will serve as a global biodiversity tool, providing scientists, policy-makers, students, and citizens the information they need to discover and protect the planet and encourage learning and conservation. An Advisory Board of 12 distinguished individuals from 5 countries will help guide the Encyclopedia.

The BHL has developed a strategy and operational plan to digitize the published literature of biodiversity held in their respective collections. This literature will be available at www.biodiversitylibrary.org. The partner libraries collectively hold a substantial part of the world's published knowledge on biological diversity. This body of biodiversity knowledge, in its current form, is largely unavailable to a broad range of applications including: research, education, taxonomic study, biodiversity conservation, protected area management, disease control, and maintenance of diverse ecosystems services.

From a scholarly perspective, these collections are of exceptional value because the domain of systematic biology depends, more than any other science, upon historic literature. The so-called "decay-rate" of this literature is much slower than in other fields such as biotechnology. Ongoing mass digitization projects lack the discipline-specific focus of the Biodiversity Heritage Library Project. These other projects will fail to capture significant elements of legacy taxonomic literature. The Biodiversity Heritage Library Project will actively seek to incorporate data and content from other digitization projects.

The Biodiversity Heritage Library Project will immediately provide content for multiple bioinformatics initiatives and research, including EOL. For the first time in history, the core of natural history library collections will be available to a global audience. Web-based access to these collections will provide a substantial benefit to all researchers, especially those living and working in the developing world.

Up-to-date information can be found on the 2 Web sites www.eol.org/ and www.biodiversitylibrary.org/. An EOL Newsletter will be produced shortly. Any who wishes may register to get regular email updates at www.eol.org/registration.php.

[1] Including the Smithsonian Institution, Missouri Botanical Garden, American Museum of Natural History (New York), Natural History Museum (London), New York Botanical Garden, Royal Botanic Garden (Kew), Marine Biological Laboratory and others.

11.2. Data Integration Issues in Biodiversity Research

Jessie Kennedy[1], Shawn Bowers[2], Matthew Jones[3], Josh Madin[3], Robert Peet[4], Deana Pennington[5], Mark Schildhauer[3], Aimee Stewart[6]

[1] Napier University, [2] UC Davis Genome Center, [3] National Center for Ecological Analysis and Synthesis, [4] The University of North Carolina at Chapel Hill, [5] The University of New Mexico, [6] The University of Kansas

The Scientific Environment for Ecological Knowledge (SEEK) project is developing an IT framework and infrastructure that will be used to derive biodiversity and ecological knowledge by facilitating the discovery, integration, interpretation, and analyses of ecological information. SEEK is based on a 3-layered architecture: the EarthGrid (the lowest layer) provides uniform access to biodiversity and other types of data sets; Kepler, a workflow tool (the highest layer), allows scientists to visually define, document, and execute their analyses; and the Semantic Mediation System (in the middle) uses domain knowledge represented in ontologies and databases to inform the discovery, integration and analysis of ecological data. The SEEK project has been motivated and directed by ecological analyses such as niche modelling and biodiversity studies. Example case studies have been used to explore the issues facing the researchers undertaking the analyses. This presentation will outline these issues and overview approaches used by SEEK.

Much modern research in ecology is based on the integration (and re-use) of multiple datasets. These datasets may be distributed globally, will be stored in a variety of formats, and most likely the data will have differing semantics reflecting any of the many measurements of spatial and temporal environmental factors and organismal characteristics and interactions that contribute to a given ecosystem. A typical scenario is a scientist is interested in analyzing the spread of invasive species in a certain region. S/he has distribution records in a personal database, but requires access to other potentially relevant datasets on-line. The researcher needs to be able to discover candidate datasets and then merge their relevant and compatible information. The researcher needs to resolve which datasets contain information about the species of interest or are to the timescale and locality of research. Simplistically, datasets might be retrieved and integrated on the basis of country and species name; however even simple data files can be extremely time consuming to integrate manually and complicated if at all possible to integrate automatically as a simple example will show.

In order to find and integrate suitable data, meta-data describing the content of the data sets is important, therefore SEEK requires data sets stored in the EarthGrid to be marked up with Ecological Metadata Language (EML). EML includes descriptions of the temporal, geographical and taxonomic coverage of the data sets. Much of the terminology used in EML is generically applicable to scientific data structures—such as table name or column label; while more domain-relevant terms—such as biomass or wing span, are defined in ontologies being developed by the SEEK team in conjunction with disciplinary specialists.

The Semantic Mediation System (SMS) layer in SEEK uses ontologies to expand terms for searching EarthGrid for data discovery and for supporting the scientist in semi-automatically transforming data for input to appropriate analytical components in Kepler. This is accomplished using a generalized ontology for modeling “observational data”, called OBOE. OBOE provides a framework in which the meaning and inter-relationships of observations within a scientific data set can be clarified. For example, one can use OBOE to indicate that various data sets contain both weights and wing spans of bird specimens—thus greatly facilitating effective data discovery and potential integration of those types of data sets. The SEEK Taxon group, whose work also sits in the semantic mediation layer of Kepler, has been researching the more specialized issues

associated with clarifying the semantics necessary to inter-relate the taxonomic coverage of ecological data sets.

Ecological data sets of relevance to biodiversity modeling tend to have been collected either over long periods of time or over a wide geographic range and typically use unqualified biological names for recording taxon occurrences or counts (often codes are used with biological names specified in the meta-data). However due to the ongoing work of taxonomy in classifying and naming the known organisms, the meaning associated with these names changes over time. Therefore representing the taxonomic coverage for ecological data by simply referencing names of species results in ambiguity. This ambiguity may be significantly detrimental to the results of any subsequent ecological analysis. To address this problem the SEEK Taxon group is adopting a taxonomic concept approach, as defined in collaboration with TDWG in the TCS standard. A necessary component will be formal modification of the Ecological Metadata Language (EML) to support identification of organisms to concept. We are currently developing tools to aid the ecologist in selecting appropriate taxon concepts, which will improve the accuracy of matching data for integration. The tools include a Taxon Object Server (whose model is closely based on TCS) to support the resolution of taxon names and concepts, and visual tools to enable users to compare concepts and clarify relationships among them.

Support is acknowledged from: NSF

11.3. Data Integration: Using TAPIR as an asynchronous caching protocol

Aaron Steele

University of California at Berkeley

There are over 100 million DarwinCore specimen records available on distributed networks worldwide. However, the search space for application-specific information is becoming vast and unreliable. For applications that know a priori what data are needed, asynchronous caching provides a reliable subset of data specific to a particular analysis. For example, an application generating species distribution models from Madagascar would benefit from accessing locally cached data where HigherGeography = Madagascar, instead of dynamically querying the network at run-time, which is expensive.

While TAPIR provides a straight forward caching protocol for retrieving specific DarwinCore concepts from a set of resources and integrating the results into a single database, key concerns are keeping these cached data synchronized with resources. For example, when records are inserted, updated, or deleted from resources, cached data must reflect these changes. Since TAPIR does not explicitly support syndicating these change events, they must be implicitly inferred by storing all resource GlobalUniqueIdentifier (GUID) and DateLastModified (dlm) concepts in a level-2 cache, and then periodically comparing it against the resource.

As a concrete example, suppose at time 't1' we create a level-2 cache 'C' for resource 'R'. The next day at time 't2' we create a second level-2 cache 'C2' of 'R'. Then, using 'C' and 'C2', the change events in 'R' during time period 't2'-'t1' can be defined as follows:

1) If 'C2.GUID' is not in 'C', then 'C2.GUID' was inserted.
2) If 'C2.dlm' is different than 'C.dlm', then 'C2.GUID' was updated.
3) If 'C.GUID' not in 'C2', then 'C2.GUID' was deleted.

In this way, after comparing records in the level-2 cache against current resource inventories, all change events are detected and associated with specific GUIDs. The level-1 cache then uses

these GUIDs to synchronize changes by submitting new TAPIR inventory requests (for new or updated records) and deleting cached records that have been deleted.

In this presentation I will discuss these key caching algorithms in more detail, including the process of syndicating resource changes in the level-2 cache using RSS feeds, the implementation of data harvesting, initial results of these methods in the MaNIS, ORNIS and HerpNET networks, and proposed additions to TAPIR. I will also address social and political concerns associated with caching, and provide information about free open source storage solutions including MySQL and the Google Base API.

Support is acknowledged from: TDWG Infrastructure Project, NSF, University of California at Berkeley

11.4. How to handle duplication in large datasets and import scenarios

Andreas Müller, Markus Döring, Walter G. Berendsohn

Botanic Garden Botanical Museum Berlin

When integrating, processing or querying biodiversity data, one sooner or later must address various problems raised by the existence of physical or digital duplicates. Both the creation and failure to find such duplicates may lead to information of lower quality in terms of completeness, readability or consistency of the dataset.

For the EU-funded SYNTHESYS project (A Synthesis of Systematics Resources) we developed a duplicate detection tool for the GBIF index of specimen and observation data as well as tools for importing taxonomic data into Berlin Model databases. In this context we developed different algorithms to handle such duplicates.

The current GBIF index contains about 100 million specimen and observation records. Querying such a database for duplicates online requires sophisticated techniques such as comparing each individual record which are too costly in terms of processing time. Hence an algorithm has been developed that adapts known record linkage techniques using pre-computed standardization and blocking, followed by online comparison and classification.

GBIF data are widely standardized, so little investment has been made in standardization. For blocking, a multi-channel sorted neighbourhood mechanism has been used. Records are inserted into sorted indices with a high probability of storing duplicates close to one another. When queried, this filtering component passes only those records that are in close proximity to the original record in at least one of the indices. The remaining candidates are compared by probability-based functions that work at both the attribute-level and record-level. Finally, classification depends on the type of duplicates searched for - physical or digital. The result-set is fuzzy, *i.e.*, not only exact duplicates are returned. This takes into account that data may undergo changes depending on the pathway from collecting to importing them into the GBIF index.

Avoiding duplicates during the automatic import of data into a taxonomic Berlin Model database needs more conservative comparison functions, as false positives should be avoided here. Still, records should be detected as duplicates if they differ only in the completeness of some less important attributes. To handle this problem, a rule based two-step algorithm for an object-oriented Berlin Model persistence layer has been developed to easily detect duplicate candidates and merge them if verified as duplicates. Therefore a set of rules has been proposed to handle different types of attributes and attribute groups. The rules are easy to adapt to fulfil different needs of different users.

The software developed is available on the BioCASE website (www.biocase.org).

Support is acknowledged from: the European Commission, Framework Programme 6, contract no RII-CT-2003-506117 (SYNTHESYS)

11.5. ALIS's Adventures in Wonderland

Samy Gaiji, Sonia Dias

Bioversity International

We will summarize the recent achievements of Bioversity International and the CGIAR System-wide Genetic Resources Programme (SGRP) in linking and integrating genebank information at a global scale. This was accomplished through the adoption of TDWG standards and Global Biodiversity Information Facility (GBIF) tools within the genebank community as a model. It is estimated that more than six million plant accessions are stored in ex situ collections worldwide and digitalized information on their essential characteristics have been gathered and stored in various institutions databases. Existing genebank information systems and portals, such as the CGIAR System-wide Information Network for Genetic Resources (SINGER) and the European Plant Genetic Resources Search Catalogue (EURISCO), are already major central entry points to such information. Their recent upgrade and adoption of TDWG/GBIF standards and protocols are making them more easily and readily accessible to the global community. Currently, information on over one third of the global holdings is available through the GBIF Portal through efforts from the genebank community. These efforts are aimed at contributing to the development of a global platform for the access and exchange of accession level information in support of the International Treaty on Plant Genetic Resources (ITPGRFA) and its global Accession Level Information System (ALIS) on Plant Genetic Resources, called for in Article 17 of the ITPGRFA.

11.6. Illustrating Relationships among Images, Specimens, Taxa, Ontologies and Character Matrices in the Morphbank Image Repository

Greg Riccardi[1], Austin Mast[1], Fredrik Ronquist[2], Katja Seltmann[1], Neelima Jammingumpula[1], Karolina Maneva-Jakimoska[1], Steve Winner[1], Deborah Paul[1], Andrew Deans[3]

[1] Florida State University, [2] Swedish Museum of Natural History, [3] North Carolina State University

Morphbank was designed to be a secure repository for images and metadata. It includes a rich data model that conforms to open community-supported standards, such as Darwin Core for specimen metadata. The metadata is available in Web pages and in XML and RDF. Access to content is managed by content owners and can be restricted to a specific group of users or made publicly available. The security model ensures that researchers can take advantage of the Morphbank system during the entire research lifecycle: from initial data collection and imaging through publication.

The Morphbank team has built enabling technology for biodiversity researchers, particularly users of biodiversity research collections and morphological phylogeneticists. In the process, we have built connections to other providers of functionality (*e.g.*, HERBIS, Specify) and biodiversity content (*e.g.*, ITIS, the Integrated Taxonomic Information System).

A major advance in the usability of Morphbank comes from collections management, which allows a user to accumulate digital objects (private and public) in collections. Users can accumulate all of the objects used in making scientific interpretations, including images, image

annotations, publications, operational taxonomic units (OTU), character matrices, and even other digital object collections. The digital objects that can be included in a collection are not restricted to objects within the Morphbank system, but can be any digital object with a globally unique identifier (GUID). Collections can be published, and can form the kernel around which new collections are built. Collections provide a powerful way of documenting and publishing scientific observations and opinions.

Figure 1 shows a collection that has been created to code a character in a matrix for phylogenetic study or the production of a multi-entry identification key. The user has organized the image tiles in order to define a particular feature. In this case, the character represents the occurrence of spots on butterfly wings. The states are 'many spots', 'few spots' and 'no spots'. Each blue tile and the thumbnails that follow it represent one of the states and some images that are coded with that state. The character can be used in one or more matrices and additional images can be coded and added to the state.

This strategy for representing character matrices is designed to enhance the interaction with systems that focus on matrices, like Morphobank and Mesquite. By using the images as the primary source of character information, we provide a different look and feel to the character coding process. Morphbank is a rich resource for other systems to use in the discovery and illustration of characters and states.

The presentation will emphasize the connections that have been defined by researchers among objects in the Morphbank system and external objects, and the ways that these connections illustrate important research concepts.

Support is acknowledged from: National Science Foundation grant DBI-0446224, the Florida State University School of Computational Science, and the National Evolutionary Synthesis Center (NESCENT). Contributions to the Morphbank image repository have been made by ATOL and PEET projects

Figure 1. The Morphbank Collections Interface showing a phylogenetic character and its states.

11.7. A Pollinators Thematic Network for the Americas

Michael Ruggiero[1], Antonio Mauro Saraiva[2]

[1] Smithsonian Institution, [2] University of Sao Paulo

The Inter-American Biodiversity Information Network (IABIN) has designated pollinator conservation as a major thematic area. Pollination is considered one of the most important processes for biodiversity conservation. Studies show that animal pollination can improve the amount and the quality of plant fecundation and fruit production, stimulating the use of animal pollination in environmental programs and proposals of sustainable agriculture. However, the success of these actions is based on the knowledge on pollinators, their conservation and interaction with the environment. New initiatives have been created to facilitate and to stimulate the dissemination of this knowledge.

One such initiative, the Pollinators Thematic Network (PTN), is being developed for the Americas and will link taxonomic and other content related to pollination. The Network will use established standards and protocols such as Species 2000/ITIS CoL, Darwin Core, ABCD, DiGIR, and TAPIR to link nomenclatural, specimen, and observation-related information. However, a novel approach is needed to enable the exchange of pollinator/plant interaction data. An extension to Darwin core is proposed to represent the interaction between specimens or observations in a broad sense, and a second extension is proposed to deal with specific pollinator/plant interactions. A prototype is being built to provide pollinator/plant data available from WebBee and other sources to the IABIN PTN portal based on the new proposed schemas.

The content from this network will link seamlessly to the GBIF network. A summary of progress will be presented as well as the proposed schemas to represent pollinator/plant interaction data.

Support is acknowledged from: IABIN, GBIF, NBII

11.8. Applying a Wiki system in the integration of biodiversity databases in Taiwan.

Burke Chih-jen Ko, Kun-Chi Lai, Jack Lin, Han Lee, Hsin-Hua Lin, Ching-I Peng, Kwang-Tsao Shao

Research Center for Biodiversity, Academia Sinica

During the past several years, a considerable number of biodiversity databases have been developed which provide Internet services in Taiwan. Some are education-oriented with vivid visual layouts to attract the younger community. Others are well organized research aids. Although huge amounts of digitized content of ecological/specimen distribution data, literature and species descriptions have been collected and are accessible, people can only browse them separately. The announcement of the Encyclopedia of Life (EOL) project has inspired TaiBIF, the GBIF portal of Taiwan, to become a portal for both novices and experts by integrating existing institutionally based data with others from miscellaneous sources, thus providing better coverage of biological topics.

Using a Wiki system supported by the open source community for the Taiwanese EOL (TaiEOL) will not only help information gathering and the sustainability of the website but can also help EOL obtain data of many endemic species in Taiwan. With features like customizable presentation layers of web design, the portal can serve information according to the users' knowledge level. To achieve this, we need a solution comprising standards and mechanisms: the ITIS Submittal Guidelines and Species 2000 Standard Dataset; reference management using

BibTeXML (http://bibtexml.sourceforge.net/); and DarwinCore as the exchange format between different data resources.

The features of Wiki-based software satisfy real world requirements. Version control reveals change history while verifying author credits, categorization frameworks automate biodiversity data analysis and establish semantic context, while map service extensions offer the ability to demonstrate spatial data. These together with the Wiki approach construct an ideal platform for collaborative efforts from enthusiasts and specialists. Community awareness of TaiBIF (http://www.taibif.org.tw/) and TaiEOL will be crucial to encourage users to become involved in this collaborative effort.

Session 12. Applications of TDWG Standards - 1

12.1. Scaling up The International Plant Names Index (IPNI)

James A Macklin[1], Paul J Morris[2]

[1] Harvard University Herbaria, [2] Harvard University Herbaria & Museum of Comparative Zoology

The International Plant Names Index (IPNI: www.ipni.org) serves as a critical reference to taxonomists on the status of names, and associated objective bibliographical details of all seed plants, ferns and fern allies. IPNI is a compilation of three data sources: the Index Kewensis (IK), the Gray Card Index (GCI), and the Australian Plant Names Index (APNI). This is a partnership between three contributing institutions, The Royal Botanic Gardens, KEW, the Harvard University Herbaria, and the Australian National Herbarium, which began in 1998. Today, our user community has broadened and there is increased demand for our authoritative index. The success of IPNI also raises challenges for the internal management of the data across the partners. Further, there is an urgent need to create a single subjective consensus synonymy of the plants of the world. This is a serious challenge that is achievable through a marriage of technology and taxonomic expertise. Many routes can be taken to achieve this marriage and thus we seek the feedback from our community of plant taxonomists, biologists, informaticians, IT professionals, and other potential users towards developing a new platform for IPNI.

Moving forward with IPNI requires addressing both internal issues such as organization of the datasets that will combine to produce the Index and external issues such as interfaces and services for the user community. We have begun discussing how to address the needs of participants and the user community, and this has raised numerous issues . The three current partners, and potentially others, must be able to manage their regional datasets using their current database infrastructure while seamlessly contributing a subset of the information they manage to this cohesive global project. We also need to insure that taxonomic vetting can be easily managed by both internal nomenclatural experts and by the greater pool of taxonomic experts in the community without one serving as a bottleneck for the other. A variety of user community standards, protocols, and tools may relate to IPNI and its relationships with the community and have significant potential to assist IPNI in the management and quality control of its data. Community standards and tools are also clearly the means by which IPNI should employ to interact with its users. Elements of the user interaction include a web portal that provides an interface to query the Index, both directly through the IPNI website, indirectly through web services using standards such as Life Science Identifiers (LSIDs), and access to static copies of the data for local use. The greatest challenge is to provide the appropriate infrastructure that would scale to the user community level.

12.2. What have George Bush, John Howard and TDWG in Common?

Paul Flemons[1], Michael Elliott[1], Lynda Kelly[1], Lee Belbin[2]

[1] Australian Museum, [2] TDWG

Using YouTube to try and get their message across! The phenomenon that is YouTube has captured not only the voyeur in all of us, but also those of us who have a difficult or complex message to sell, and see the short video format that YouTube has made famous as a means of doing this. The Australian Museum (AM) has been examining how the evolution in digital content creation and multi-platform distribution can create new audiences and innovative content

for cultural institutions. As a result we have had the opportunity to explore the use of short videos as a means of communicating our work. Two reasons for our involvement with short video are: potential for developing website help videos to communicate more effectively than the usual text heavy help pages, and the ability of such videos to communicate complex ideas in an easily understood format. As a result of this work the AM has been working with Biodiversity Information Standards (TDWG) to develop digital stories for the TDWG Access Protocol for Information Retrieval (TAPIR) and Life Science Identifier (LSID) standards. We will present these videos along with some observations on the process of creating them.

Support is acknowledged from: TDWG Infrastructure Project, Australian Museum

12.3. Developing an Observational Data Model to Facilitate Data Interoperability

Steve Kelling

Cornell Lab of Ornithology

Broad-scale ecological studies often require information assembled from multiple disciplines. Data heterogeneity across multiple disciplines and data sets creates major informatics challenges that include the need to better discover, access, interpret, and integrate relevant data that have been collected by others. A National Science Foundation (NSF) sponsored workshop was held to address these challenges. The major conclusion from the workshop was that a shared model for observational data would facilitate data interoperability, and would enable significant integration within and across disciplines. This observational data model would provide a flexible and ubiquitous construct for scientific data, and would be appropriate for building an interoperable data sharing framework. Promoting a shared community model for observations could lead to major implementation advantages, and facilitate tool construction and re-use. This presentation will provide more detail on the results of the workshop.

Support is acknowledged from: National Science Foundation

12.4. Moving to Fully Distributed, Interoperable Repositories for Biodiversity Information

Greg Riccardi[1], Austin Mast[1], Fredrik Ronquist[2], Katja Seltmann[1], Neelima Jammingumpula[1], Karolina Maneva-Jakimoska[1], Steve Winner[1], Deborah Paul[1], Andrew Deans[3]

[1] Florida State University, [2] Swedish Museum of Natural History, [3] North Carolina State University

The Morphbank (http://morphbank.net) research project has created a repository for images that adds significant value to stored images by managing complex metadata, organizing images for searching, and allowing users' comments and annotations to be directly linked to images and metadata.

Morphbank has more than 1 terabyte of images covering a broad spectrum of life. It has hundreds of users and images from a variety of collections. It has user interfaces that allow users to identify new taxa, characters and matrices, and *ad hoc* collections of images and other objects. The productive and dedicated Research and Development team has helped hundreds of researchers organize and use their images.

The primary Morphbank site includes the images, the database with all of the metadata, and the Web servers. Each mirror site holds some of the image files, and each image file is stored in more than one mirror site. As the number of images and amount of metadata grows, the burden of storage must be shared among the organizations that submit and use the information. A fully distributed system would consist of multiple sites, each with Web server, database, and image store. These sites would share responsibility for metadata and images. Communication between sites would use globally unique ids (GUIDs) for objects and maintain consistency of information.

In future, image and metadata repositories must be better able to communicate. For example, sharing of information between the Morphbank system and the Encyclopedia of Life (EOL) information systems is necessary so that images in Morphbank can illustrate concepts in the EOL, and vice versa. A user of an EOL Web page should be able to see images stored in Morphbank or other image repositories. Users of Morphbank should be able to find the EOL information associated with the images, specimens, taxa, and localities that they find. At present, many Morphbank Web pages contain references to GenBank and other systems and vice versa, but those references have been created by users, not through any automation or standards for referencing and data interchange. Shared ontologies must provide common language to facilitate these connections.

Morphbank is evolving into a collection of standard data exchange formats, application programming interfaces, server software, and client software that can be installed at any site. The Web and database activity of any individual site will not exceed the capabilities of off-the-shelf server technology.

The long term success of biological image and metadata repositories must occur through fully distributed systems that can freely find and share information in standard formats and must take advantage of emerging standards for information interchange, including Darwin Core, ABCD, the TDWG Species Profile Model, Web 2.0, RDF, LSID, and the W3C image annotation standard. The TDWG community must find ways to make these initiatives work for us.

The presentation will give examples of current and planned interoperability between Morphbank and other data repositories. Emphasis will be placed on how ontologies can create reliable search criteria for biodiversity information.

Support is acknowledged from: NSF, NESCENT, Florida State University

12.5. Building the German DNA bank network using TDWG standards

Gabriele Dröge, Jörg Holetschek

Botanic Garden & Botanical Museum Berlin-Dahlem

A DNA bank is a service facility for the long term storage of well documented DNA. For the German DNA bank network project, four partner institutions with complementary expertise and collections will form a pool: The Botanic Garden and Botanical Museum Berlin-Dahlem (BGBM), the Bavarian State Collection of Zoology Munich (ZSM), the Forschungsmuseum Alexander König Bonn (ZFMK), and the German Collection of Microorganisms and Cell Cultures Braunschweig (DSMZ). The main focus of the network is to enhance taxonomic, systematic, genetic and evolutionary studies.

The BGBM is responsible for coordinating the establishment of the network, designing and implementing the network infrastructure, developing databases and the web portal, and storing botanical DNA samples (plants, algae, protists). Since October 2004, a DNA bank pilot project has been in progress at the BGBM, with researchers' requests and shipping being processed continuously.

The architecture of the new network reflects the spread locations of natural history collections within Germany and aims at providing a central web portal for researchers to access data from various, distributed data sources. Each partner institution will have its own database for storing DNA data as well as the associated collection information. Most of the more recent DNA collections are stored in a variety of collection management systems, using either open standards software (*e.g.*, Specify) or proprietary software. Whenever possible, integration of those data into the new system will be done by using BioCASe (Biological Collection Access Service).

In addition, a central web portal will be developed that can be used by researchers to explore the DNA material in stock at the partner institutions, view all of the associated information (metadata) as well as high-resolution digital images of the original specimens, and request DNA samples. Communication between the network components is based on the BioCASe protocol and ABCD and uses the BioCASe provider software as well as the Unitloader package for querying distributed data providers. In a later project phase, support for the TAPIR protocol will be implemented to make sure that collection providers following the latest protocol standards will be accessible.

By using protocol standards and standard software components the DNA Bank Network ensures that all data generated will be accessible to the international biodiversity data networking initiatives such as GBIF and BioCASE.

http://www.bgbm.org/bgbm/research/dna/
http://www.gbif.org
http://www.biocase.org

Session 13. Applications of TDWG Standards - 2

13.1. Development of a TAPIR-based protocol for the Global Invasive Species Information Network

Jim Graham[1], Annie Simpson[2], Michael Browne[3], Thomas J Stohlgren[4], Greg Newman[1], Catherine Jarnevich[4], Alicia W Crall[5]

[1] Colorado State University, [2] National Biodiversity Information Infrastructure, [3] IUCN Invasive Species Specialist Group, [4] US Geological Survey, [5] University of Wisconsin-Madison

Invasive species are a global problem and managing them is inefficient at best. Being able to access data on invasive species globally will: 1) provide information on potential invasives, 2) allow access to a wide variety of information on effective management techniques, and 3) enable research on the nature of invasive species at a level that is impossible with isolated databases. There is an existing community of organizations with databases that contain information on invasive species that would be valuable to a much larger audience. Surveys and interviews with these organizations have shown that they have limited resources and technical knowledge to participate in data exchange. At the same time there are few resources available for centralized development and support of this activity. What is needed for the Global Invasive Species Information Network to be successful is a protocol that these organizations can implement with a toolkit that can be easily customized and supported by a very small staff. We have defined a protocol based on a subset of TAPIR and implemented a test system that has shown to be much simpler than existing approaches. This GISIN protocol is very robust and will provide high performance. The next steps will be to begin adding data providers to the network while further refining the system and to develop a design for a toolkit that meets the needs of the invasive species community. For the test system, see http://squall.nrel.colostate.edu/cwis438/websites/GISINDirectory/.

Support is acknowledged from: US National Science Foundation, NASA, Colorado State University, the Global Biodiversity Information Facility, TDWG Infrastructure Project and the US National Biological Information Infrastructure.

13.2. Marking and Exploring Taxonomic Concept Data

Paul Craig, Martin Graham, Jessie Kennedy

School of Computing, Napier University, Edinburgh

We will present a summary of our work in developing graphical tools for marking up and exploring relationships between overlapping taxonomies. Two visual tools are presented: the first demonstrates how concept relationship sets between classifications can be constructed using a drag and drop tool, and the latter allows exploration and comparison of multiple, inter-related classifications through these concept relationships.

The first application, Concept Relationship Editor, allows taxonomists to create, edit and delete relationships between pairs of classifications. Users can select a pair of concepts, one per classification, and choose a relationship type to construct between them through a drag and drop metaphor in the user interface.

The second tool, TaxVis, allows the exploration and comparison of multiple such classifications. Comparison of sub-groups or an entire classification can be made against the rest of the classification set either through name matching or through the defined concept relationships,

allowing the degree of overlap and the difference between naïve and explicit matching to be observed. Overlapping groups of concepts are indicated using colouring. Explicit relationships for specifically interrogated concepts are drawn as links between the corresponding concept representations in the display. A linked panel showing details of specific concepts and their relationships in text form can also be viewed.

The data sets we use in these prototypes are defined under a subset of the Taxonomic Concept Schema (TCS) TDWG standard, with particular focus on its concept relationship mechanism. The tools are designed to improve accuracy beyond naïve name matching when mapping between related taxonomic classifications. The advantage of using mainly graphic representations to convey such classifications and their inter-relationships is that it allows creating, querying and interpreting results to be performed through point and click operations rather than requiring detailed knowledge of the TCS schema and associated XML mechanisms. The dataset also appears as a cohesive whole rather than a succession of atomic information items as would be returned by a traditional text-based system.

We performed initial usability tests on the prototype applications by asking taxonomists to try simple tasks with the tools, and we are interested in accommodating further user-centred development through empirical testing of the prototypes and associated qualitative feedback.

Support is acknowledged from: National Science Foundation (NSF) through the SEEK project and also by the Engineering and Physical Sciences Research Council (EPSRC).

13.3. From National Plant Checklist to Chinese Virtual Herbarium (CVH)

Keping Ma, Haining Qin, Lisong Wang

Institute of Botany, Chinese Academy of Sciences

The world's largest Flora, the Flora Republicae Popularis Sinicae (FRPS) has been completed after 45 years of extraordinary effort by 312 Chinese botanists. The Flora documented more than 31,000 species of vascular plants native to China, with more than half of them found nowhere else. The Flora and its voucher herbarium specimens preserved in the herbaria of China and other countries are important baseline data to biodiversity research in this region. However, these data were not easily accessible to botanists and public users through an internet-based environment before our present project, the Chinese Virtual Herbarium (CVH) http://www.cvh.org.cn/. The initial goals of CVH are therefore to integrate these primary biodiversity data into a database, and provide web-based services through standard information technology.

There are two key elements in the present CVH, the Catalog of Life, China Plants and the Virtual Herbarium.

1) The Catalogue of Life, China Plants brings together all published scientific names of Chinese plants from FRPS, the (English language) Flora of China (FOC), local Chinese floras, taxonomic monographs, and journal papers. This comprehensive reference system of Chinese plant names represents the botanical portion for the Species2000 China Node. It now contains more than 95,000 records covering Chinese mosses, ferns and seed plants. The accepted name for each plant and its synonyms are flagged according to appropriate taxonomic opinion. Each name has been validated against the original literature by taxonomic experts. Geographic distribution, economic uses, conservation status etc. are also linked to each accepted name. These data will be put on-line and published as a CD ROM at the end of 2007.

2) The Virtual herbarium will bring together the label data from all Chinese herbaria through a distributed network. Emphasis is being placed on collection event information, such as when, where and by whom the specimen was collected. Presently, there are more than 2.7 million

specimens in the 17 Chinese herbaria that are participating in CVH. High-resolution images have been linked to each specimen. Georeferencing and GIS-based distribution maps are also in preparation.

In order to provide an integrated information platform, CVH also includes electronic versions of local Chinese floras and important taxonomic bibliographies, such as the Illustrated Flora of Higher Plants in China, Flora of Tibet and Flora of Hainan, and The Bibliography of Chinese Systematic Botany, 1949-1990. A gallery database including more than 27,000 color images belonging to 3,800 native Chinese species also has been linked to accepted names.

Our ultimate goal for the CVH is to integrate all available data sources, including herbarium specimens, observational data, taxonomic bibliography, e-flora, and field color images, and to make it the corner-stone of Chinese plant biodiversity research.

Support is acknowledged from: The Ministry of Science and Technology, P.R.China; The Chinese Academy of Sciences

13.4. The Central African Biodiversity Information Network (CABIN): a Contribution to the Sub-Saharan African Biodiversity Information Network (SABIN)

Patricia Mergen[1], Charles Kahindo Muzusa-Ngabo[2], Michel Louette[1], Franck Theeten[1], Bart Meganck[1]

[1] Royal Museum for Central Africa, Leuvensesteenwaeg 13, 3080 Tervuren, Belgium, [2] Université de Kisangani, Centre Universitaire extension Bukavu, DR Congo

The aim of this multi-donor initiative is to establish a thematic sub-Saharan African Biodiversity Information Network of African institutions and research centers that hold and manage information about local biodiversity and natural environments.

For historical reasons as well as current cooperation and development projects that include biodiversity investigation, a large proportion of scientific knowledge and specimens from sub-Saharan Africa are located and managed in European and North American research facilities. These facilities are members of international initiatives, such as the Global Biodiversity Information Facility (GBIF), Biodiversity Information Standards (TDWG) or the Consortium of European Taxonomical Facilities (CETAF), whose goals are to make biodiversity and scientific collection Information and knowledge available to all. The Royal Museum for Central Africa (RMCA) and several African partner institutions are active collaborators in these initiatives.

These networks have developed informatics tools and standards to share biodiversity and associated geospatial information. For example, more than 120 million of records, from more than 1000 collections institutions are freely accessible via the GBIF network. According to rough estimates more than 1.5 million of the specimen/observation records served by GBIF concern sub-Saharan African species. However, except for South Africa, it appears that no African institutions are providing information to GBIF directly. Instead, the information resources of African institutions are usually hosted by institutions outside Africa. Recently a GBIF national node has been setup in Tanzania in the framework of the GBIF CEPDEC (Capacity Enhancement Programme for Developing Countries) in collaboration with the Danish cooperation and development agency.

The intention is to enhance the capacity of African partner institutions to serve information to the GBIF network using TDWG standards and informatics tools. These goals are intended to be achieved by:

1. Assessment of the current ICT infrastructure in African partner institutions and prioritization of their needs in collection management systems and biodiversity information sharing;
2. Cataloguing of African biodiversity collections and research projects;
3. Enhancing the ICT infrastructure and implementation of GBIF/TDWG informatics tools at several identified African partner institutions based on the results of points 1 and 2;
4. Setting up a thematic data access and data usage portal targeting sub-Saharan Africa; and
5. Organize capacity building and training sessions in collaboration with GBIF on how to become a data provider, on how to access and process biodiversity information data available through the GBIF network with a special focus on georeferencing techniques.

The Afro tropical Zoology and Geological Departments of RMCA are currently joining forces and will use funding from the Belgian Cooperation and Development and the Belgian Science Policy offices to enhance access to both biodiversity and geological information resources in Rwanda and the Democratic Republic of Congo and to upgrade the ICT infrastructure of institutions managing those resources.

The resulting Sub-Saharan African Biodiversity Information Network will enhance access to African biodiversity information, to support joint scientific research and publication. Free access to biodiversity information will also improve local decision making in conservation and environmental protection.

Support is acknowledged from: Belgian Development Cooperation, JRS Biodiversity Foundation, GBIF, Belgian Science Policy Office

13.5. The potential key role for promoting the use of Biodiversity Information Standards by a consortium of research institutions in the Eastern Democratic Republic of Congo (DRC) in Central Africa.

Charles Kahindo[1], Dudu Akaibe[2], Upoki Agenong'a[2], Ulyel Ali-Pato[2], Patricia Mergen[3], Michel Louette[3], Erik Verheyen[4], Jérôme Degreef[5]

[1] Université Officielle de Bukavu, Democratic Republic of Congo, [2] Université de Kisangani, Democratic Republic of Congo, [3] Royal Museum for Central Africa, Tervuren, Belgium, [4] Royal Belgian Institute for Natural Sciences, Brussels, Belgium, [5] National Botanic Garden, Meise, Domein van Bouchout, Meise, Belgium

Research institutions in Africa can achieve the goal of efficient biodiversity information sharing and a number of benefits could be generated. As centers of knowledge such institutions play, now as before, a key role in generating and applying knowledge for economic growth and local development.

Many African universities and research centers perform well, despite the challenge of limited funding by initiating revitalization programs, but the role of international support is of paramount importance in the present context.

In the Democratic Republic of Congo (DRC), five collaborating institutions have been identified, based on geographical location, historical background, infrastructure, networking and resource availability. This consortium includes two universities and three research centers:

- Kisangani University is one of the three leading universities in DRC. Created in 1963, its specific focus is on training and research in biological fields, for which it is the leading facility in the country.

- Bukavu State University was founded in 1993 as a branch of Kisangani University; but expanded rapidly and developed into an autonomous university, recruiting students in the far east of the country.
- Created in 1935, the National Agricultural Research Center at Yangambi, located in the eastern part of the Congo Forest block near Kisangani, was once the biggest agricultural research center in sub-Saharan Africa. The largest herbarium collection of tropical Africa is hosted in Yangambi and collection is still in good condition for future botanical research.
- The National Research Center in Natural Science of Lwiro, located in Eastern Congo, 40 km north of Bukavu, was created in 1947. It was meant to be a reference research station in Central Africa for natural science studies especially biodiversity surveys, seismology, documentation and tropical medicine. It holds important biological collections.
- The Hydrobiology Research Centre of Uvira, located at the extreme northwestern end of the Lake Tanganyika, was created in 1949 by the colonial power in order to carry out research activities in biology and ecology on Lake Tanganyika.

The newly-elected DRC government recognizes scientific research as a necessity for improving human conditions. It supports the cultural and economic development of small entities and would welcome support in the field of biodiversity information access and exchange for consideration when resolving biodiversity problems and issues.

A TDWG initiative to stimulate the role of the eastern DRC research institutions in this field would encourage and facilitate their interaction with local, national and international stakeholders. It would also ensure continued research and public access to the most current research materials from one of the world's biodiversity hotspots.

In our presentation we give a detailed overview of the resources available in this region and the initiatives we are undertaking in collaboration with local, Belgian and other international partners.

Support is acknowledged from: Federal Public Service Foreign Affairs, Foreign Trade and Development Cooperation Belgium - DGDC

13.6. An Anthropology Extension to the ABCDEFG Schema

Charles J.T. Copp

Charles Copp Environmental Information Management

Physical Anthropology collections are held in a great number of museum and university collections world-wide. They are important because they are the basis of our understanding of human evolution and diversification. Specimens range from the fossil remains of hominids and pre-hominids right up to recent remains of modern people. In addition to the expected bones, teeth and skulls, there is a wide range of other anthropological material including mummies, preserved tissues, genetic samples, models, facial reconstructions, images, census datasets, pathology collections and molecular data. Many of the specimens are associated with artefacts ranging from stone tools to elaborate grave goods and some specimens may also have known biographical data.

The ABCD schema includes many of the elements needed to document anthropological specimens and the EFG extension provides elements to handle the fossil material. There are, however, a number of types of data that will require their own elements and it was therefore proposed to develop an anthropological extension to the ABCD schema within the framework of the European SYNTHESYS project.

An international workshop was organised in Budapest (August 2006) to develop a preliminary data model and schema for Physical Anthropology collections, building on the existing ABCD and EFG models. The requirements for an extension for Anthropology were documented in a subsequent report in November 2006 (Zsuzsanna Guba, Standardisation of anthropological collection data NA-D 2.52). The Hungarian Natural History Museum, Budapest and the Natural History Museum, London then commissioned the author to develop an integrated XML schema based on the findings of the report. The first draft of the resulting schema is now complete and will be placed on the SYNTHEYS website, along with full documentation, for comment in time for the TDWG 2007 conference.

One issue in creating the anthropology schema is that the specimens, their associated artefacts and collecting context can link them closely to ethnographic and archaeological data, which could have a major impact on the size of the schema. The development of an archaeological schema with full coverage of excavation methodologies or the full description and analysis of cultural artefacts are major efforts in their own rights, requiring wide consultation. It is possible to limit the links into these other domains by restricting the information recorded or limiting it to simple text elements, but this could undermine the extensibility of the schema and miss the opportunity to provide the greatest scope for data interchange. The solution adopted in the preliminary schema has been to create extensible and replaceable elements that allow the schema to be developed further as needed. An example extension for pottery artefacts was developed as a proof of concept.

Support is acknowledged from: SYNTHESYS Network Activity NA-D Developing and maintaining databases

Session 14. Communication, Education and Outreach

14.1. TDWG Communication

Lee Belbin

TDWG Infrastructure Project

TDWG has not been effective in communicating its work within its potential community of interest. Users of TDWG standards need quality software and documentation. Organisations working with natural history information / biodiversity must also be aware of the significance of TDWG's work.

While TDWG does have an internationally significant role, it has been marginalized within its client community. The lack of effective communication is a key cause. The TDWG Infrastructure Project has invested heavily in improving communications but even the best technologies do not guarantee effective communication.

The project has built a collaborative environment and attempted to improve communications, but that project ends this year. How will TDWG continue to build its reputation? I suggest that it is by the quality of our outcomes and the effectiveness of communicating those outcomes to our clients.

The obstacle is however the extremely small ratio of membership to activities. At the time of writing, TDWG has 17 individual members, 18 institutional members and at least 16 groups and 18 standards in various forms. This situation cannot assure quality outcomes. TDWG must therefore either build membership and/or focus on fewer core areas.

Support is acknowledged from: The Gordon and Betty Moore Foundation

14.2. EDIT scratchpads as a vehicle for community building and outreach.

Dave Roberts, Vince Smith, Simon Rycroft

The Natural History Museum

The idea behind the EDIT (European Distributed Institute of Taxonomy) Scratchpads is simple enough: make web sites easy to get, easy to use, and easy to read by computers and humans alike. Scratchpads are an intuitive web-application, enabling communities of taxonomists collaboratively to build, share, manage and publish their data on the web. Scratchpads are easy to use, adapted to taxonomists' needs and provide powerful tools for managing biological information. Data are stored in an underlying database that is connected to a series of user-controllable modules which determine the sites functionality.

Major features include:

- Unlimited content (pages, images, maps, bibliographies, phylogenies, DNA sequences, specimen lists, classifications, forums) or define your own content type;
- Workflows for data import (includes MS Excel support) and editing;
- User controlled hierarchies to classify content;
- Unlimited and controllable site membership;
- Public, private and user defined access groups;

- Context sensitive user profiles;
- Rich text (WYSIWYG) editing & intuitive administration; and
- User defined web addresses.

Scratchpads are independent and unconnected, allowing communities to create distinct customized sites tailored to their needs. They are built on the content management system Drupal (http://drupal.org/), and managed on servers at the Natural History Museum, London. Content is archived such that it can be cited like a traditional publication.

In the first four months of operation, 18 communities of users have established sites to capture and disseminate thematic information about various taxonomic groups. In the next phase of development, data from these independent sites will be connected and fed through to a common data structure, such as that proposed for EDIT's Cybertaxonomy platform or that generated by the Encyclopaedia of Life (EoL) project.

For more information, explore the following links:

- Current Scratchpads: http://www.editwebrevisions.info/SiteList
- Sandbox Site: http://sandbox.editwebrevisions.info/
- Scratchpad Functionality: http://www.editwebrevisions.info/featureList
- Scratchpad Help Screencasts: http://www.editwebrevisions.info/help
- Scratchpad Frequently Asked Questions: http://www.editwebrevisions.info/aboutus
- Apply for your Scratchpad at http://editwebrevisions.info/signup

Support is acknowledged from: European Community Sixth Framework Programme: Network of Excellence "EDIT"

14.3. KeyToNature: a European Project for Teaching Biodiversity

Pier Luigi Nimis, Stefano Martellos

Dept. of Biology, University of Trieste

KeyToNature is a 3-year targeted European project, approved in the framework of the e-Contentplus Programme, addressing important issues of the ongoing digital revolution. It strives to achieve a pan-European approach to teaching biodiversity, focusing on the identification of organisms.

Species identification has been based mainly on paper-printed tools, such as classical dichotomous keys which are sometimes based on systematic hierarchies. Such keys have several educational drawbacks. Several software packages have been developed in recent decades, which enable the rapid and easy creation of interactive identification tools which are not necessarily based on systematics. Such tools have a high educational content, they may be much more user-friendly than the traditional paper-printed keys, and can be easily adapted to different educational levels. Their introduction into the educational world will overcome one of the most serious gaps in biodiversity education: the lack of identification tools adapted to user-specified needs. The new tools require the connection of different, presently scattered, databases, including those of images, sounds, textual descriptions, and thesauri of scientific and common names.

KeyToNature aims at improving the searchability and usability of existing digital contents to support the emergence of a European educational service related to teaching and learning

biodiversity with novel, advanced, powerful approaches, filling a serious gap at European Union (EU) level. The new technologies raise a series of novel issues and problems, which require solutions at the European level.

The main objectives of KeyToNature are to:

1) Increase access and simplify use of e-learning tools for identifying biodiversity;
2) Improve interoperability among existing databases for the creation of identification tools;
3) Optimise educational efficiency and increase quality of educational contents;
4) Add value to existing identification tools by providing multilingual access; and
5) Suggest best practices against barriers that prevent the use, production, exposure, discovery and acquisition of the digital contents required for designing the identification tools.

A selection of primary and secondary schools and university courses in the EU will be involved in testing, using and accessing the educational products of KeyToNature. The project mobilises 14 partners from 11 EU countries. It includes leading centres in biology, pedagogy and education and information technology, plus three small/medium enterprises (SMEs). International data standards will be used to make existing e-contents more accessible, usable and exploitable in formal education, both for face-to-face and distance learning. Using the business model to be developed, the educational tools will be accessible beyond the end of the project.

14.4. Using New Technologies for Education

P. Bryan Heidorn

University of Illinois

When the Internet first came into use, only two tools were available at a reasonable cost to facilitate education: web pages and PowerPoint slides. Both tools produced relatively static learning objects that were designed to be passively read. However, learning is more effective in a dynamic and interactive mode. New social computing Internet tools are increasingly being used by academic and other organizations to facilitate two-way communication among communities of learners. I will review ongoing projects and tools categorized as digitally mediated information services. I will focus on delocalized participation in biodiversity standards meetings and tools that foster broader participation in standards development and adoption groups throughout the year. I will review the application and limits of Voice over IP (*e.g.*, Skype), desktop sharing, SecondLife, MySpace, UTube and other tools.

I will focus on two toolkits. One tool may be used to allow students (and non-experts in TDWG standards) to participate in the 2008 Biodiversity Standards Meeting (BSM). A second toolkit might be used to facilitate interactive training through the year. Ordered by depth of learning, students will need to be able to identify prerequisite knowledge, knowledge sources, access to standards and associated experts, viable exercises and professional development in internships and eventually employment. Bandwidth limitations and knowledge of students suggests the use of the simplest technology that meets the needs. Prior to major BSM sessions speakers should provide 1) an abstract with references (pointing to background knowledge necessary to understand the session). During the session, technologies should include 2) slide broadcast or remote desktop sharing, 3) audio broadcast at variable sound quality, 4) two way multiple participant “chat” for remote participants to text questions and comments, 5) VoIP input patch to speaker system, and 6) all modes of communication should be recorded for later use.

Many useful technologies are already used by some standards development groups but more could be added to improve and simplify communication. For asynchronous communication, Wikis provide a good method of text communication but are poor for synchronous communication. Phone conversations do not leave a record for others who did not participate in the original call. All technologies listed above for annual meetings should be supported for standards meetings. All participants should be able to share both desktop and voice. Standards are a particular challenge because of the complexity of the content and the requisite background knowledge needed by students. Annotated demonstrations and exercises are much more effective than static documents at effectively communicating the issues and process involved in standards. TDWG's experience with DiGIR training is a good model of the kinds of participatory learning that should be the goal for education at a distance or asynchronous learning.

14.5. Available Communication Tools

Lutz Suhrbier

The ability to communicate and cooperate over long distances appears to be the most important feature of the Internet for research. For example, email represents one of the most widely known and used communication tools.

Using current web technologies, several applications have evolved, providing facilities which alleviate the organisation and coordination of cooperating work groups. In particular, the new emergent techniques are often subsumed within the term "Web 2.0". These offer fascinating visions of individually focused information flows for producers and consumers, allowing a new experience as applications are aggregated into personal "Web Desktops".

The aim of this presentation is to give an overview of currently available communication tools outlining those of potential interest for applications within a biodiversity context.

The presented communication tools result from an evaluation effort for the European Distributed Institute of Taxonomy (EDIT; http://www.e-taxonomy.eu) to prepare a Web-based platform supporting communication and cooperation processes adapted to taxonomists' needs. The evaluation results may be of value to other biodiversity disciplines and the presented tools may inspire work groups to integrate them into their environment.

Support is acknowledged from: European Community Sixth Framework Programme: Network of Excellence "EDIT"

14.6. Mapping Biodiversity Specimen Data: Opportunities for Collaboration

Gail E. Kampmeier[1], John Pickering[2]

[1] Illinois Natural History Survey, [2] University of Georgia

Making data available to a broad audience is desirable and even required by funding sources supporting our research and collections. GBIF (Global Biodiversity Information Facility) plays no small part in leading this charge not only in assembling an electronic catalog of names, but with the debut of its new portal (http://data.gbif.org/), with information from over 220 data providers and nearly 1500 datasets that may be mined. While laudable, the steps to make these datasets available to GBIF are often beyond the scope of those without robust information technology support, making these datasets vulnerable to being lost as grants end, data and database stewards change priorities, retire, or leave the field. However, one way to capture and integrate these datasets is through Discover Life (http://www.discoverlife.org/), whose mission is "to assemble and share knowledge in order to improve education, health, agriculture, economic

development, and conservation throughout the world". With nearly 1.2 million species represented, its major strengths include mapping and on-line illustrated identification tools. Mapping of taxa, specimens, and collections is in collaboration with TopoZone.com. As with GBIF, Discover Life (DL) does not take ownership of data provided to it, but attributes it back to its source either by drilling back to a provider's database or denoting its ownership throughout the display process.

Our data on the fly family, Therevidae, is an example of a mature database (http://www.inhs.uiuc.edu/research/mandala/TherevidWebMandala.html) that has been working its way towards being served to GBIF, but was able to be mapped and represented with DL beginning in 2003. Discover Life accesses exported text files of over 1,300 valid (accepted) taxonomic names (http://www.discoverlife.org/mp/20q?search=Therevidae) and nearly 123,000 georeferenced specimens, which it updates daily. Users choose a taxon and where specimens exist, scalable distribution maps are automatically generated, with clickable data points, allowing users to see details about individual specimens. The real power of the system is in the customizable mapping (http://www.discoverlife.org/mp/20m?act=make_map). Users can map one or more taxa from multiple data sources or entire datasets, restrict or expand mapping by data source(s) or points, center maps by clicking or using fixed latitude/longitude or UTM coordinates, and make maps for display or publications in color or black & white. Satellite, topographic, and for some areas of the globe, photo maps allow visualization of the landscape.

Currently, as has happened with many initiatives, development of GBIF and DL has been taking place largely in parallel, often targeting slightly different audiences, with somewhat different goals. One of the strengths of GBIF is its commitment to the history of taxonomic names and its adoption of TDWG standards. A major weakness is the difficulty, real or perceived, for many users to get their data to GBIF. Discover Life can quickly map specimens of one or more taxa, drawn from a single or multiple data sources. Datasets do not need to be independently available on the internet: database owners may provide DL with a delimited text file with basic standards-compliant output. For both data providers and the intended audience and/or user groups, it is time to recognize strengths of various systems and endeavor to work together towards a collaboration that benefits all.

Support is acknowledged from: National Science Foundation; Schlinger Foundation; Polistes Foundation

Session 15. Building Biodiversity Data Content

15.1. Integrating the catalogue of Mexican biota: different approaches for different client perspectives

Diana Hernandez, Susana Ocegueda, Patricia Koleff, Sofia Escoto, Rocio Montiel
CONABIO

The long-standing and ambitious aim of creating a unique index of the species of the world has stimulated the creation of local catalogues for biodiversity assessments. One of the main goals of CONABIO has been to gather, manage, analyze and disclose information on Mexican biodiversity, which is mainly based on specimen data acquired from national and international collections. But there is an outstanding problem: how to analyze biodiversity without standardized taxonomic information? Scientific names are the key data to organize and retrieve biological information however most of these names are still sparsely located in independent publications. In 1997, CONABIO started the creation of the Taxonomic Authority Catalogues. These hierarchical databases are organized by taxon and include valid names, as well as their synonyms, species distribution, and relevant additional information. They are always reviewed by expert taxonomists. To date, we have compiled a list of 32,000 valid species, 4,000 infraspecies, and 20,000 synonyms, representing around 18% of the estimated names of Mexican flora and fauna.

The catalogues are organized for separate taxonomic groups into a specialized information system developed by CONABIO called Biotica. Integration of all catalogues into a single index of the whole Mexican biota is our next goal with the aim of implementing generic user functions and facilitating cross-links with external information systems. The information is available in different electronic formats to encourage its diffusion to a broader audience. So far we have received positive responses from our users with respect to the use of multiple download formats and forms. These formats are available through our webpage http://www.conabio.gob.mx/informacion/catalogo_autoridades/doctos/electronicas.html. They are commonly used by scientists, high school and undergraduate students, decision makers, government officials and other relevant specialists. When constructing the index of the world's biological diversity it is important to consider the widest range of users of this information in order to guarantee its uniqueness, reliability and utility.

Support is acknowledged from: CONACyT

15.2. Moving Targets: Integrating semistructured data

Pepe Ciardelli, Marc Geoffroy
Botanic Garden Botanical Museum Berlin-Dahlem

We will present our experience importing nomenclatural, taxonomic, bibliographical, and distribution data from text files into a relational database, as part of the Euro+Med Plantbase project. The aim of the project is to present a taxonomic inventory of vascular plants in Europe and the Mediterranean countries on the web.

There is a cultural gulf in taxonomic computing: few taxonomists truly comprehend that in order to develop adequate computer software, every conceivable case must be dealt with in advance. Even when users have agreed in advance to a detailed data format, the consequences of not sticking rigorously to this format are seldom truly appreciated.

Taxonomists expect software to follow their normal working processes, while at the same time being able to access the results of their work in all possible formats. In practice, this may mean importing a document file (usually Microsoft Word), in atomized form, into a database. Although these data appear to be in a structured format, *e.g.*, tables with pre-agreed spacing, special signifying characters, etc., they remain text, meant for humans, not computers, to read. There is normally no mechanism to check the validity of input at the instant it is typed; errors are first recognized when the document as a whole is parsed and loaded into the database. In our experience, importing a number of such files required us to continuously adapt import software.

The first complication arises from the fact that it is in practice impossible to iron all typographical errors out of a 400 page document. Programmers are expected to build error-tolerant software, and in fact, all errors described above can, in our experience, be corrected programmatically after a few iterations. The larger problems lie in: variation between taxonomists' standards; the peculiarities specific to certain taxonomic groups; lack of agreement on extra-taxonomic notations; and lack of communication between programmers and taxonomists.

For example, there may be only one group in a series of imports where species are allowed to be included in another species – if a programmer without any taxonomic background is given this file to import without warning, the entire algorithm for taxonomic inclusion may be thrown into chaos. Notoriously complex and exception-rich groups like Pilosella and Hieracium require a great deal of extra notation, and subsequently communication between taxonomist and programmer is of the utmost importance.

There is no way to program for 100% of exceptions and the programmer hours that would need to be invested are better spent putting unresolved cases in catch-all fields in the database, then allowing experts to parse the data later manually. In the end, we must recognize that the taxonomist is always right, and learn to adapt to his/her work methods. However, the learning process must take place on both sides. Many taxonomists are not ready to invest the time in checking the success of the import with, for example, a web interface. In the best case, the programmer and the taxonomist would develop such a tool according to the taxonomist's preferences. While the younger generation of taxonomists has a substantially greater level of comfort with computers and appreciation of their capabilities, the problem of moving targets will likely continue to exist for the foreseeable future.

Support is acknowledged from: Euro+Med Plantbase project from the Mattfeld-Quadbeck Foundation, the Association of Friends of the BGBM, and the Global Biodiversity Information Facility (GBIF)

15.3. Global Compositae Checklist: Integrating, Editing and Tracking Multiple Datasets

Christina Flann[1], Aaron Wilton[2], Kevin Richards[2], Jerry Cooper[2]

[1] Wageningen University, [2] Landcare Research

The Global Compositae Checklist is an ambitious project aiming to integrate existing electronic data sources for one of the largest plant families in the world to provide definitive nomenclatural information and up to date taxonomic concepts. Purpose-built Checklist Software has been designed and developed by Landcare Research in New Zealand. The Checklist Software 1) imports multiple existing datasets, 2) integrates datasets together using rules, and 3) provides a transparent digital audit trail for the integration process and subsequent manual annotation and editing. Datasets have been contributed from many different providers from major botanical institutes around the world. Datasets are imported via TCS (Taxon Concept Schema standard), one of the first real uses of this TDWG standard, or a defined fixed MS Access format. The

provider records are then integrated into the Checklist using a simple algorithm that tests each record for possible existing matching records. A tool for matching variant author abbreviations is included in the Checklist Software. This can continually be updated with the correct abbreviation. Any given variant should be referred to, allowing the matching of names with varying author citations. Using these methods 'provider' records are linked to a new or to an existing consensus record. The nomenclatural data are then verified by an editor, with any edits being added as an editor's provider record. Following each integration or edit the data for the consensus record are re-calculated using a majority consensus from the linked provider records to determine the value in each field calculated. However, this simple majority is overridden by an editor's record which has priority over other provider records. Taxonomic concepts are also included when they are included in the data sets by data providers and are integrated following the same principles as for the nomenclatural data. The data from the Checklist will be available through the Checklist website and via a TCS mediated web-service.

Support is acknowledged from: Global Biodiversity Information Facility (GBIF), Netherlands Organisation for Scientific Research (NWO), Systematics Association

15.4. The changing role of publishing biodiversity data for Northern Ireland on the internet

Susan Fiona Maitland

National Museums Northern Ireland

The Centre for Environmental Data and Recording (CEDaR) was established as part of the Sciences Division of the Ulster Museum (now part of National Museums Northern Ireland) in 1995, in partnership with Environment and Heritage Service (EHS) and the local recording community. Functioning as a Local Records Centre, CEDaR facilitates the collection, collation, management and dissemination of biodiversity and geodiversity information for Northern Ireland and its coastal waters. One way that CEDaR disseminates this information is via the suite of web sites at www.habitas.org.uk.

Initially web sites were developed on an *ad hoc* basis however, in March 2004 CEDaR centralised the site development role into one post. Since this was done the workload of this post has increased dramatically. Existing web sites need to be updated and standardized and new sites created for each new CEDaR initiatives.

One of the first web sites to be developed was the Flora of Northern Ireland. Species pages are displayed dynamically, pulling information directly from an Access database containing all the information required for the species pages including image and map details. This method has been adopted for most of the CEDaR web sites, except where the number of species represented is small.

Each site has been designed to allow easy navigation and utilises similar formats for the display of information. The species accounts are written by local experts and museum staff using content and layout standards. The accounts are then proof-read and formatted for the web. This process is laborious and along with the preparation of images to illustrate the species accounts adds a great deal of time to the process of publishing the biodiversity information on the web sites. This presentation will help delegates to better understand how CEDaR staff solved the technical problems.

Support is acknowledged from: National Museums Northern Ireland, Environment and Heritage Service

15.5. The role of networks in a cyberinfrastructure

Zack Murrell, Derick Poindexter

Appalachian State University

The World Wide Web (WWW) has changed science and, in turn, how systematic biology, conservation and biogeographic studies are conducted. The Semantic Web will further revolutionize science by building a mesh of information that will allow scientists to gather and analyze data in a more thorough fashion. The Semantic Web will also provide opportunities for information retrieval by the general public in complex and automated systems.

The innovations of the Semantic Web enable the development of "virtual communities" of scientists. Such a virtual community is SERNEC, the SouthEast Regional Network of Expertise and Collections. This network of herbarium curators provides an electronic database of herbarium specimen labels and images. As this database is built, its contents will be reviewed by the collective taxonomic expertise of this virtual community. This process will result in an increasingly accurate portrayal of the biogeography of the region.

The SERNEC virtual community includes information scientists, social scientists, educators, and artists, as well as the taxonomic expertise of the region's curators. With the Semantic Web, power comes from the development of "high quality" information. The quality database developed by SERNEC will attract the public, government decision-makers, corporations and educators. This will, in turn, increase the value of curatorial expertise.

Collaboration and innovations such as interactive keys and mapping developed within this virtual community will provide many positive outcomes. Complex information will be delivered in intuitive ways to a range of user groups. We hope that this system will stimulate interest in plant systematics, conservation and biogeography.

Session 16. Where to from here: Evolving and Emerging Standards

16.1. New Standards from Old - reconciling HISPID with ABCD

Peter Neish[1], Ben Richardson[2], Greg Whitbread[3]

[1] National Herbarium of Victoria; Royal Botanic Gardens Melbourne, [2] Western Australian Herbarium; Department of Environment and Conservation; Perth, [3] Australian National Herbarium; Centre for Plant Biodiversity Research; Canberra

Australian herbaria have interchanged specimen data using the HISPID TDWG standard for nearly 20 years. HISPID is an existing TDWG standard describing the data elements required for the meaningful interchange of herbarium specimen data sets. ABCD is a generalised standard for the interchange of specimen and observation records. Our attempts to implement Australia's Virtual Herbarium (AVH) using ABCD resulted in some loss of information and enforced relaxation of content standards compared to HISPID.

We will describe the process used to reconcile these two TDWG standards through the extension mechanism of ABCD and the restriction mechanism of XML Schema. We also discuss the challenges encountered enforcing data integrity in ABCD, while at the same time meeting the needs of the Australian herbarium community in a 'real-world' federated application.

Support is acknowledged from: TDWG Infrastructure Project

16.2. Biodiversity Portals: Implications for TDWG

Donald Hobern

Global Biodiversity Information Facility

Many of the TDWG standards were first developed to support federated searches. The intended use case was as follows:

- A user submits a search request (*e.g.*, occurrences of Chiroptera from before 1950);
- A workflow application passes the request to relevant data providers;
- Each provider responds with at least the first page of matching records; and
- The workflow application returns the combined results to the user (with support for retrieving records not returned in the initial request).

This approach requires the workflow application to maintain the following information:

- Basic technical metadata for each dataset (*e.g.*, endpoint, data standards);
- Session information to support paging through matching data sets;
- Ideally - knowledge of each dataset's content so requests can be forwarded only to relevant data providers; and
- Ideally - domain knowledge to enhance requests (*e.g.*, to use synonyms as well as the accepted name for a species).

Most biodiversity data portals, including GBIF, have used a cached index of key information retrieved from the various data providers to provide quick answers to most search requests. This solves several problems which appear as network sizes increase:

- It is wasteful to forward every request to every potentially relevant data provider;
- Many requests are too general for any datasets to be excluded in advance;
- At any time some providers will be off-line; and
- Some providers cannot handle complex requests, or respond very slowly.

The decision whether a network should use an index/cache will depend on several factors:

- The size of the network;
- The robustness and availability of the data providers;
- Whether the data providers have the desire/capabilities to maintain a live server;
- Whether search requests require joins between multiple data sets; and
- Whether the portal needs to pre-process data to make queries more reliable.

This approach has some implications for future TDWG standards development:

1. TDWG should produce recommendations on the use of TAPIR and/or OAI-PMH (or some similar harvesting protocol) for maintaining central caches of records. The approach should be selected to minimise the burden on data providers.
2. TDWG should continue to revise its standards to minimise alternative representations for the same information and to stabilise key concepts. The LSID vocabularies promise to provide stable properties which could be recognised wherever they are used. This will simplify building index databases.
3. TDWG should adopt or develop metadata standards for on-line datasets. As well as standard Dublin Core properties, these metadata should document the taxonomic, geographic and temporal coverage of the dataset (with references to standard taxonomies and vocabularies) and the methods used to gather the data (atlasing projects, amateur observations, etc.). Even when portals use local indexes, good metadata can simplify selection of datasets and improve the quality of the indexing process.
4. TDWG should ensure that protocols and data standards make it easy to provide attribution for each individual record. The GBIF portal offers interfaces for searching aggregated data but cannot use the same interfaces used by the original data providers. This is because the existing output models were developed for individual data providers (with the same metadata for all records) rather than for composite documents with records from a number of datasets. This problem can be solved with standard record-level properties for attributing any record to its source dataset.

16.3. Building an index of all genera: A test case in interchange

David P. Remsen[1], David J. Patterson[2]

[1] GBIF, [2] Encyclopedia of Life

A challenge facing the Global Biodiversity Information Facility (GBIF), the Encyclopedia of Life (EoL), and other initiatives that manage large amounts of species information is in organizing the data in a way that makes biological sense. Data mobilized within these initiatives are associated

with taxon names, many of which are no longer accepted or may be misspelled. The GBIF ECAT program draws upon the Catalog of Life (CoL) as the major component of a taxonomic infrastructure but cannot effectively assess names not present in the catalogue. The primary goal of the CoL is to compile the currently accepted names of living taxa, not to organize all taxon names associated with past and present biodiversity data. Thus, additional components of the organizational infrastructure are needed to complement, not complicate, existing efforts.

One component is a catalog of all biological genera. Genera provide potential organizational value in species data because a genus name is a component of every species combination. A complete catalog of all genera offers a number of compelling benefits from both an organization as well as a referent biological perspective. First, it provides the means to identify and quantify all generic-level homonyms. Such a compilation may prevent future homonyms. The inverse identification of all uniquely spelled genera is also valuable as it implies that any species combination referencing that genus name was unambiguously assigned to it. Assigning all genera to a provisional consensus taxonomic position is a relatively accessible ambition, already underway. A genus provisionally placed within a higher taxon links all associated combinations with that higher group. Such a structure can serve useful disambiguation functions for taxon references that are otherwise undifferentiated. Coupled with high-quality data mobilization efforts, like the Biodiversity Heritage Library, it enables taxon experts, for example, to be alerted to previously unknown taxon references in their area of expertise.

In order to achieve this index, the All Genera Index (AGI) must coordinate with a number of existing nomenclatural initiatives that already catalog large subsets of generic names. It must also reconcile overlap, where it exists, which requires informed lexical comparison algorithms that can confidently distinguish spelling variation in genus-author combinations from true homonyms. Interchange is bidirectional, as the index itself may serve as a point of origin for previously uncatalogued names. All of this is dependent upon access to flexible standard messaging systems that enable the exchange and synchronization of both verified and unverified nomenclatural records between systems that may employ different implementation mechanisms. Such an exercise, focused on genera and a small number of providers, will serve as a useful test case prior to attempting this on a larger scale with species checklists.

The AGI will serve a useful function as a staging interface between authoritative nomenclators and currently relevant digitization activities such as the Biodiversity Heritage Library that will undoubtedly uncover novel taxon references with name combinations currently outside indexed compilations. The AGI will not only provide a provisional repository for such unverified records it will serve a vital organization role to support the verification and assembling of a complete catalog of taxa and associated names.

Support is acknowledged from: Global Biodiversity Information Facility

16.4. Catalog of Fishes 2.0: improving user services and preparing for community participation

Stanley Blum[1], Richard Pyle[2]

[1] California Academy of Sciences, [2] Bishop Museum

The “Catalog of Fishes”, edited by W.N. Eschmeyer, was published as a three-volume work in 1998. Since then, basic information about taxonomic names has been expanded to include recent opinions about taxa and classification, and the database has nearly doubled in size. Eschmeyer currently maintains the Catalog and is committed to doing so through at least 2008, but the long-term future of this critical resource needs to be addressed. We have evaluated several alternative models for maintaining the Catalog and believe that a community participation model is the most

likely to succeed. In this model, authors of taxonomic acts (or someone acting for them) will have the ability to add records for taxa that have been newly described or revised, while peers and editors will have the ability to review and validate records.

With funding from the United States National Science Foundation we have recently begun a two-year project to prepare the Catalog for maintenance under that model. In particular, our goals are: (1) to migrate data contents into a form compatible with the recently adopted Taxonomic Databases Working Group standard for taxonomic data; (2) to improve end-user facilities, such as browsing and querying interfaces and support for downloads in widely used formats; (3) to develop policies for arbitrating disagreements about database content; (4) to develop a web application that enables remote users to login and create and edit database records; and (5) to develop web services to make the contents of the Catalog directly available to other databases that include basic taxonomic information about fishes.

Many of the challenges we will confront in this project will be common to other taxonomic databases. As the Catalog of Fishes moves to an open model of content management, we need to ensure that the credibility of the database is not diminished. Policies for managing content must be reviewed and agreed to by the larger taxonomic community. These policies will begin with commitments to have the database reflect only the content of the published literature and to have conflicting opinions represented fairly and openly. Another issue to be solved will be the partitioning of responsibility among databases with overlapping content, such as ITIS, Species 2000, ZooBank, FishBase, and the incipient Encyclopedia of Life.

Support is acknowledged from: US National Science Foundation

16.5. Summary of upcoming challenges

Anna Weitzman[1], Christopher Lyal[2]

[1] Smithsonian Institution, [2] The Natural History Museum, London

Earlier in TDWG 2007, we presented a talk and diagram: "Taxonomists at Work: Relationships of Process and Data". Using this as a basis, combined with the presentations, discussions, and results of TDWG 2007, we will present a summary of what TDWG has accomplished in Biodiversity Data Standards and what the challenges are for the future.

Session 17. Computer Demonstrations

17.1. Usability Evaluation of Tools for Marking and Exploring Taxonomic Concept Schema Data

Martin Graham, Paul Craig, Jessie Kennedy
Napier University

We will demonstrate graphical tools for marking up and exploring relationships among related taxonomic concepts and request that potential users participate in a usability test. Two visual tools will be evaluated: the first demonstrates how relationships between concepts in different classifications can be constructed using a drag and drop function, and the second allows exploration and comparison of multiple, inter-related classifications through these concept relationships.

The first application, the Concept Relationship Editor, allows taxonomists to create, edit, and delete relationships between pairs of classifications. Users can select a pair of concepts, one per classification, and choose a relationship type to construct between them through a drag and drop function in the user interface.

The second tool, TaxVis, allows exploration and comparison among multiple classifications. Comparison of sub-groups or an entire classification can be made against other classifications either through name matching or through pre-defined concept relationships. Name matching uses differential colouring of child taxa to indicate the degree of overlap in the selected taxonomic concept(s). When linking by pre-defined relationships is chosen, links depicting the particular relationships are drawn between the corresponding concept representations in the display. A linked panel shows details of relevant concepts and their relationships in text form. Similarities or differences in agreement between the name matching and explicit relationships can be observed.

The data sets we use in these prototypes are defined using a subset of the Taxonomic Concept Schema (TCS) TDWG standard (http://www.tdwg.org/activities/tnc/tcs-schema-repository/), with particular focus on its concept relationship mechanism, designed to improve accuracy beyond naïve name matching when mapping between related taxonomic classifications. The advantage of using mainly graphic representations to convey such classifications and their inter-relationships is that it allows creating, querying, and interpreting results to be performed through point'n'click operations rather than requiring detailed knowledge of the TCS schema and associated XML mechanisms. The dataset also appears as a cohesive whole rather than a succession of atomic information nuggets as would be returned by a traditional text-based system.

The demonstration will consist of a series of usability tests. Volunteer users will be asked to sign up for available usability sessions and complete a pre-test questionnaire. In a session, each user will be trained with a demonstration of the tool, following which they will be asked to complete a task-based usability test, providing comments as they go, which will be recorded for later analysis. Users will then be asked to complete a post-test questionnaire.

Support is acknowledged from: U.S. National Science Foundation (NSF) through the Science Environment for Ecological Knowledge (SEEK) project and also by the Engineering and Physical Sciences Research Council (EPSRC)

17.2. Machine Learning to Produce Structured Records from Herbarium Label OCR

P. Bryan Heidorn, Qin Yin Wei

University of Illinois

In this session, we will demonstrate the learning process of HERBIS, the XML (extensible markup language) schemas used in learning and markup, the principles for the use of the web interface and the web services interface, and discuss future developments. In the current version of HERBIS, all machine learning is run by the project programmer. End users provide raw OCR (optical character recognition) output to the classifier and the system returns an XML document. In the new version, we will permit users to provide accuracy feedback to the system allowing the performance to improve with system experience.

As presented elsewhere (1), supervised machine learning (SML) techniques and learning by example can be used to transform herbarium specimen label data to digital format. In the HERBIS project the objective of SML is to make a computer system that can recognize patterns in the OCR output of scanned herbarium labels, and convert them into 36 XML components including, for example, family, genus, species, author, variety, location, collection date, annotations, and others for convenient ingestion into museum databases. To accomplish this, the human trainer gives the computer properly classified examples to learn from. The computer generalizes from these examples to properly extract information from previously unseen examples. While a computer is accomplished at never forgetting an example that it has seen, like a savant child, the computer cannot recognize something it has never seen before. For example, the determiner on a label might be indicated by "Determiner:", "DET", or "Det.", all of which are different from the point of view of the computer. Therefore, it is the job of the human trainer to provide carefully-selected examples that are representative of the future tasks that the computer will be asked to perform, including typical OCR errors. The trainer must tell the computer how to classify strings like "DFT:", where a faded "E" was misread by the OCR as an "F" as well as other numerous but systematic errors. Using a combination of Rote Patterns Learning, Naïve Bayes classification, Hidden Markov Models, and other techniques, HERBIS reaches high accuracy on some elements but not all. Through improvements in the algorithms and improvements in training examples, performance is being enhanced. With a little practice, botanists can learn to provide training examples for the computer to allow the HERBIS SML System to efficiently convert herbarium label data to database format.

(1) Heidorn, P. Bryan, Wei Yin Qin, Beaman, Reed and Cellinese, Nico (2007). Learning by Example: Machine Learning and Herbarium Label Digitization. Joint Plant Science and Conference Botany 2007, Chicago Illinois. July 7-11, 2007.

Support is acknowledged from: National Science Foundation

17.3. Federated Authentication and Authorisation with Shibboleth

Lutz Suhrbier[1], Andreas Kohlbecker[2], Markus Döring[2]

[1] Freie Universität Berlin, [2] Botanic Garden & Botanical Museum Berlin-Dahlem

Shibboleth is a project of the Internet2 Middleware Initiative (http://middleware.internet2.edu), which provides an architecture and open-source implementation for federated identity-based authentication and authorization infrastructure. Federations may be built from groups, organisations, or projects who agree on common security policies and practices. Using the SAML (Security Assertion Markup Language) and Shibboleth (http://www.e-taxonomy.eu)

protocols allow for cross-domain single sign-on and remove the need for content providers to maintain usernames and passwords.

This computer demonstration shows the current single sign-on approach used for the federation of several taxonomic data and service providers within the European Distributed Institute of Taxonomy (EDIT; http://www.e-taxonomy.eu). Currently, our approach implements a central Shibboleth Identity Provider using an extended metadata schema focusing on taxonomy particularly. Relying upon the Shibboleth security components (http://dev.e-taxonomy.eu/wiki/SecurityComponents), multiple applications (*i.e.*, "service providers" in Shibboleth terms) such as Drupal, Subversion, and TRAC can be adapted to specific user preferences as well as to meet security concerns of service providers.

The scenario is subjected to enlargement by further applications or services and is applicable to other TDWG groups or institutions as well. Finally, it has the potential to become the security framework building up a common TDWG federation in the near future.

This computer demonstration is presented in conjunction with the talk "Shibboleth, a potential security framework for the TDWG architecture".

17.4. Integrated Open Taxonomic Access (INOTAXA) Pilot

Anna Weitzman[1], Christopher Lyal[2], Cynthia Sims Parr[3], Farial Shahnaz[3]

[1] Smithsonian Institution, [2] Natural History Museum, [3] Information International Associates

Both taxonomists and those who need taxonomic information require greater access to material held in natural history museums and similar large biological repositories and their libraries. These repositories hold a wealth of inadequately accessible resources that describe and explain the diversity and complexity of life on earth. Mining these data for research, conservation, drug discovery, protected area management, disease control, education, enjoyment of the natural world, etc., is difficult, time consuming, and often leads to redundant efforts. What should be a seamless, open "book" of knowledge consists, instead, of disparate, unintegrated sets of data - some in electronic form but most still on paper, and both published and unpublished.

Information held in museums centers on the following types of biological datasets: specimen collections, taxonomic databases, published taxonomic literature, geographical information systems, and unpublished archival materials. Making these information sources available is part of a larger, worldwide effort to enable easy access to the complete range of data required to understand individual species and their environmental and evolutionary relationships. This will require the establishment of cross-linkages between, and simultaneous access to, datasets from such information sources throughout the world.

As a start on this important task, we are in the preliminary stages of developing the INOTAXA portal, which uses an XML schema, taXMLit, for literature markup. The portal will be a web workspace in which taxonomic descriptions, identification keys, catalogues, names, specimen data, images and other resources can be accessed simultaneously according to user-defined needs. It will allow access to data held in multiple servers, and will use a distributed data model. If, in the future, the various nomenclatural Codes permit web publication of new taxonomic names and acts, INOTAXA will be able to integrate single descriptions placed on servers worldwide, so long as they are indexed through a registry such as the one operated by the Global Biodiversity Information Facility, GBIF. The portal will be built on open source software that will be made freely available to easily set up at sites, as desired, worldwide. We will demonstrate the software and solicit feedback on the interface and functionality.

Support is acknowledged from: The Atherton Seidel Fund of the Smithsonian Institution

Session 18. Posters

18.1. NCD Toolkit: Storing and Exchanging Natural Collection Descriptions Using the NCD Schema

Wouter Addink, Ruud Altenburg

ETI

The National Nodes of the Global Biodiversity Information Facility (GBIF) store and/or manage information about Natural Collections. To aid the creation and aggregation of data to describe those collections in a standardised format, a toolkit is being developed using the Natural Collection Descriptions (NCD) standard under development within Biodiversity Information Standards (TDWG).

GBIF commissioned the development of an NCD toolkit which consists of a MySQL database and web-editor in PHP. Collections descriptions can be maintained with the toolkit on any platform that is able to run a web browser. The toolkit supports multiple languages and allows import of legacy data in NCD schema compliant XML documents. The toolkit will also include an easy to use web service for extraction of data in the XML format as defined by the NCD standard. Entering a correct URL in a web browser will return an XML file with data to the user without the need for additional technology (this is called a REST style web service). In addition an Open Archives Initiative (OAI) harvesting interface will be developed. This means that users of OAI can use standard OAI queries to get NCD data.

The NCD toolkit will be available for download as open source software at the GBIF website. Installation executables for both Macintosh and Windows platforms will be available, and a helpdesk will be established by NLBIF (the national GBIF node in the Netherlands) to support users.

A central database will be established at the Berlin Botanic Gardens (BGBM) pre-seeded with data currently in the BioCASE NoDIT database (a database developed in the European BioCASE project containing information about natural collections in Europe). This database will be used in testing and to act as a repository for data from organisations that do not wish to host their own.

Support is acknowledged from: GBIF

18.2. Patterns in biodiversity records digitised from literature

Arturo H. Ariño, Estrella Robles

University of Navarra

Until the advent of taxonomic and biodiversity databases, manual literature searches were the main source for studies of biodiversity distribution. Recently that has changed with the wide availability of primary species occurrence data that initiatives such as TDWG (Taxonomic Databases Working Group, now Biodiversity Information Standards, http://www.tdwg.org/) and GBIF (Global Biodiversity Information Facility, http://www.gbif.org/) are facilitating. Articles, reports, and other literature items had to be parsed, and the relevant records (in its most basic form, taxon names occurring at given localities) extracted from the references. Finally, occurrence tables, usually targeted at certain taxonomic groups, had to be built.

Work towards the automation of such constructs was, thus, a logical event. Taxonomic databases started to incorporate records parsed from literature, but apparently to a lesser extent than the

incorporation or federation of databases based on specimens or observations. This is somewhat striking, as the wealth of information to be garnered from published items (especially peer-reviewed papers) is both large and of high quality. Nevertheless, several databases, both due to purely scientific initiative, or commercial ones such as Zoological Record, exist and could eventually be encouraged into some federation scheme under new or existing standards, such as those pursued by TDWG.

However, both the nature of such existing databases, and the underlying digitisation processes (be it manual or automatic, the latter generally through optical character recognition and a combination of taxon name discovery, georeferencing, and natural language processing), should be well studied in order to ease their convergence or determination of their fitness within any proposed standard. Emerging patterns on records already digitised from literature sources could provide valuable insights into possible digitisation issues.

In search of such patterns, we have analysed Zootron 4, a vintage taxonomic database including sampling and observational data and about 200.000 worldwide occurrence records of fauna, which were manually digitised from scientific literature over a period of more than two decades. In this poster, preliminary results are introduced.

Record patterns belong to two main conceptual categories, which are not readily distinguishable upon analysis: those existing in the literature, and those arising from the selection or digitisation processes. In turn, these main categories can be cross-divided into at least four broad classes of patterns, according to the aspect being analysed: taxonomic, geospatial, human-dependent, and chronological patterns. Statistics, maps, and conceptual plots help recognise and visualise these patterns, while geostatistics, record analysis and categorisation, and cross-referencing facilitate their discovery.

The techniques introduced in this poster could potentially be used in other databases to confirm or disprove the existence of these patterns and, more importantly, to clearly separate patterns in the literature from patterns in the digitisation of the literature.

18.3. Biodiversity Information Standards (TDWG): A Poster

Lee Belbin

TDWG Infrastructure Project

Biodiversity Information Standards (TDWG) needs to build its membership. Resources that could help with TDWG's outreach efforts have been placed on the web site http://www.tdwg.org/about-tdwg/. Among those resources are PowerPoint® presentations, brochures and a poster. All these resources provide a relatively simple introduction to TDWG and its work, and are designed for members to use at meetings related to TDWG's domain.

The brochures and the poster are in Adobe® PDF format. The brochures are designed to be printed in a tri-fold format on A4 paper. The poster is designed to print on A1 paper (594mm wide × 841mm high). Like the other resources, the poster is simple in design; portraying TDWG products as the centre of a complex network of biologically-related projects and their dependency on sharing data. The aim of the poster is to portray TDWG as the international 'enabler' for the sharing of biological data.

We encourage members and friends to use these resources wherever possible.

Support is acknowledged from: The Gordon and Betty Moore Foundation

18.4. Recorder 6 and its collection management extensions

Guy Colling[1], Tania Walisch[2], Charles Copp[3]

[1] Luxembourg national museum of natural History, [2] Luxembourg National Museum of Natural History, [3] Environmental Information Management

The Luxembourg Natural History Museum (MNHNL) studies and documents the natural heritage through the recording of biological and geological field occurrences and the collection of specimens. Since the 1980's various efforts have gone into the digitisation of such information at the Museum, leading to numerous 'home made databases'. Eventually, the MNHNL opted for an integrated database solution in 2000 and implemented Recorder (http://www.recordersoftware.org/), a software package for the collection, collation and reporting of biological field records, developed by the Joint Nature Conservation Council in the UK. The main reasons for the Museum's choice in favour of Recorder were the quality of the underlying National Biodiversity Network (NBN) data model; the open, flexible build, allowing new functionality to be added, and the integrated data transfer format.

Recorder was not suitable for earth science information and natural history collection management in general. In 2001, Copp extended the NBN data model, integrating museum collections. The NBN data model included a greater majority of attributes required for the recording and management of biological field data. Recorder had, however, only limited facilities for recording details of specimens linked to records and none at all for museum specimens lacking collecting event data. Recorder was also unable to manage earth science data such as names of rocks, minerals or stratigraphic information about fossils. Consequently, the Museum engaged in the development of a collection management module for Recorder 6, which uses an MS SQL-Server database. The module was based on the extended data model, and now handles the following extra classes of information: accessions, collections, specimens, documents as objects, images as objects; storage details like buildings, rooms, specimen cabinets; loans, exchanges and valuations. It also allows museum staff to document conservation checks of specimens; to describe and prioritise conservation tasks, and to keep track of conservation jobs and materials used. Finally, it allows documentation of funding sources for jobs or acquisitions as well as enquiries from the public. The collection module also includes a quick data entry form for specimens and a powerful thesaurus to manage all kinds of term lists used in Recorder: for example, lists for taxonomy, geology and for collection management (specimen condition types, acquisition methods, etc.), as well as gazetteers, keyword lists, etc.

The number of users of Recorder software has steadily increased over the past six years and the scope of the software has been greatly extended. The addition of collection-related functionality and the current work on one or more web-based versions has attracted new large-scale users in Europe and interest around the world. No single individual or organization has full ownership or intellectual property rights associated with Recorder software, add-ins, or its theoretical basis (model and standards). Recorder has been developed for the public benefit and should be retained in the public and open source domains. During the second international Recorder conference, which will be held in spring 2008, the creation of an international Recorder Foundation is planned. The Recorder Foundation's main role will be to simplify the management of partnerships, the coordination of initiatives, and to assure coherence of the development path.

Support is acknowledged from: Luxembourg National Museum of Natural History, eLuxembourg, National Biodiversity Network

18.5. TOQE - A Thesaurus Optimized Query Expander

Niels Hoffmann, Patricia Kelbert, Pepe Ciardelli, Anton Güntsch

Department of Biodiversity Informatics and Laboratories, Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin

Websites and web portals allow users to find information on unit data. Unfortunately, most search engines only return data literally matching search terms, and do not look for concept-related data.

At the same time, there are already several thesauri available online, such as taxonomic checklists, country lists, etc. These thesauri allow users to retrieve a list of concept-related elements for a given term or a given concept.

As long as web portals are not integrated with these available thesauri, the user will find searching frustrating and imprecise. A thesaurus integrated into the search engine would make searching much more efficient. A search engine equipped with a thesaurus interface would first query a thesaurus for concepts related to the given search term, and then perform the original query "expanded" with the results from the thesaurus.

Such an interface should include operators to retrieve:

- a list of relationship types or methods implemented by the thesaurus and
- a list of semantically-related concepts for a given search term.

The service should also be able to connect to any kind of structured thesaurus database, such as XML Topic Maps (XTM), Simple Knowledge Organisation Systems (SKOS), Relational Database Management Systems (RDBMS), etc.

The Thesaurus Optimized Query Expander (TOQE) has been implemented as a web service. TOQE provides the client with a fixed set of methods, thereby hiding the complexity and structure of the underlying thesaurus. The service transforms requests into queries applicable to the underlying thesaurus. Results are then transformed into a well-defined XML schema and returned to the client. The process of querying a thesaurus becomes transparent, generic, and independent of the thesaurus used.

TOQE is already used by the SYNTHESYS portal for access to specimens and observations (http://search.biocase.org/synth-ui), and with the Euro+Med taxonomic checklist as the thesaurus database. The web service can be accessed at http://ww3.bgbm.org/toqe/.

18.6. The Atrium® Biodiversity Information System

John P Janovec, Amanda K Neill, Jason H Best, Mathias Tobler, Anton Webber

Botanical Research Institute of Texas

Atrium (http://www.atrium-biodiversity.org) is a technology platform for revolutionizing biodiversity information management by enabling researchers and organizations to share, synthesize, manage, and publish biodiversity data in a collaborative, online environment. Atrium provides a broad range of tools for research organizations as well as an unparalleled, open-source framework based on industry standards, which facilitates the development of powerful applications and tools for the biodiversity community. The development of the requirements, design, and implementation of Atrium have been funded in part by grants from the Gordon and Betty Moore Foundation, the Beneficia Foundation, the World Wildlife Fund, and the Amazon Conservation Association.

Atrium provides a web-based platform for merging specimen/collection data with species information via species pages, literature citations, GIS layers, ecological data, environmental data, and images. As of August 2007, the current version, Atrium 1.5, provides:

(1) tools that facilitate the collection, organization, and sharing of organismal, environmental, geographic, and ecological information;

(2) tools for worldwide, real-time collaboration to connect taxonomic experts who contribute updates to identifications of specimens and images;

(3) a digital herbarium with many tools for entry, sorting, filtering, and analysis of botanical collection data (collaborators can view and download complete collection data and high-resolution images, print labels, annotate collections through online determinations, and produce annotation labels remotely);

(4) a dynamic distribution mapping system that links the georeferenced botanical dataset with Google Maps and Google Earth for desktop and online mapping of collections and species;

(5) an innovative module for automated user-generated plant checklists to the families, genera, and species of plants by geographic region, research site, research project, habitat, and habit;

(6) a GIS Data and Metadata Server that allows users to browse, search, visualize, download, and upload an extensive collection of GIS data layers – points, lines, polygons, and images;

(7) a bibliography and library module for managing, browsing, searching, and downloading literature references;

(8) sophisticated tools for the production of image-rich botanical field guides in digital format that can be designed and printed on-line and on-demand by the user; and

(9) a system for managing data access and protection that is facilitated by a sophisticated permission-management system that controls over 50 different levels of data access.

Support is acknowledged from: The Gordon and Betty Moore Foundation, Beneficia Foundation, World Wildlife Fund, Amazon Conservation Association

18.7. Botanicus - A freely accessible, Web-based encyclopedia of digitized 18th, 19th and early 20th century botanical literature

Chuck Miller, Chris Freeland, Doug Holland, Robert Magill

Missouri Botanical Garden

Digitizing, indexing, and annotating historical scientific literature is vital to future research in systematic botany, the science of the identification of plants. Like other natural history disciplines - but unlike the physical sciences - systematic botany is built upon and requires frequent reference to the literature of its past. To conduct carefully documented and authenticated research, botanists must spend weeks in library collections searching the published botanical literature for data to develop a new project or substantiate their recent observations.

Comprehensive collections of botanical literature are only available in a handful of libraries, all located in North America and Europe. For botanical researchers, these library-centered literature searches, while a crucial requirement of any project, delay hypothesis development or recognition and publication of new plant discoveries. For those traveling in remote parts of North America or stationed overseas, lack of access to library resources compounds these difficulties. Further, no matter how scrupulous the search, when scientists must work manually through an array of journals and books it is impossible to be sure that all historical facts have been located and all published observations have been seen.

Over 67,000 systematic botanical publications exist, but only those most recently published are in digitized form. Botanicus now makes many of these publications freely available world-wide via the Internet.

Support is acknowledged from: Keck Foundation

18.8. Challenges and tradeoffs in the management of geological context data in paleontological collections.

Paul J. Morris

Harvard University Herbaria/Museum of Comparative Zoology

Paleontologists collect fossils from exposures of rocks on the Earth's surface and by other means such as drill cores. As these fossils are curated into museum collections, information systems need to handle the geological context from which the fossils were collected. The geological context for a fossil may include the geologic time unit from which it was collected, a description of the rocks from which it was collected, the formal (lithostratigraphic) name of the rock unit from which it was collected, and a wide variety of geologic zones that represent fine-grained hypotheses about the ordering of events in geologic time. There are multiple different ways to handle these data in relational databases and in data interchange standards, with tradeoffs between different choices. I will examine several different approaches to the management of information related to the geological context for paleontological specimens and the tradeoffs that result from different choices.

In abstract terms, the geological context of a fossil is an attribute of the place or locality from which it was collected. In practice, geologists think of a locality as a place in a two dimensional coordinate system where a slice of geologic time is exposed, and where multiple different rock units and geologic time units may be exposed and collected from that locality. Treating geological context as an attribute of a locality may result in the entry of redundant locality records that differ only in their geology, and may create problems for the management of legacy data with poorly resolved geographic and geologic data. Geological context may be treated as an attribute

of a specimen, and this may work well for capture of legacy data and collections where material enters as bulk samples that are tracked and processed, but may make it very difficult to manage changes in hypotheses about the geological context of an outcrop or a drill core. Geological context can be treated as an attribute of a collecting event. Collecting events are often unknown for legacy data. Tying geological context to a collecting event creates challenges for cleaning and editing legacy data by raising the possibility of introducing errors that propagate through many related specimens, however, the nature of a collecting event as a sampling visit in time for a locality may make it the best fit for geological context information.

A geological context may be thought of as a single objective attribute related to a specimen, or as a hypothesis that can change over time. Large scale geological information (placement in a high level geologic time unit, or in a high level rock unit), is not likely to change over time, whereas fine scale divisions of geologic time and rocks are hypotheses that are more likely to change. The relationship of a specimen to a geological context can be thought of as a determination made at some point in time, with a changing history as the placement of boundaries of time and rock units change. Treating a relationship between a collecting event and a geological context as a one to many relationship makes for less complexity (in databases, code, exchange standards, and user interfaces) than does treating it as a many to many relationship, and there may or may not be enough data of high enough quality to merit tracking histories of changes of the geological context of a specimen.

18.9. RDF123 and Spotter: Tools for generating OWL and RDF for biodiversity data in spreadsheets and unstructured text

Cynthia Sims Parr[1], Joel Sachs[2], Lushan Han[2], Taowei David Wang[1], Timothy Finin[1]

[1] University of Maryland, [2] University of Maryland Baltimore County

OWL (the Web Ontology Language) and the related RDF (Resource Description Framework) are XML-based languages designed to represent the semantics of data. These languages enable systems to go beyond simple controlled vocabularies and specify the contexts and logical relationships among terms. Formal ontologies use classes (*e.g.*, Species A) and properties (*e.g.*, is a member of, or eats, or has body mass) to represent concepts and relationships as assertions. For example, two assertions might be "Species A is a member of Family B," and "Family B is a taxon whose members eat plants". A machine can then use logic to reason that all individuals of Species A should also eat plants; other assertions would make it clear that, in this context, plants are also organisms and not factories. The Semantic Web is the collection of web documents using semantic languages such as OWL and RDF. On the Semantic Web, specialized search engines can use such data assertions to more sensibly find and integrate information. For example, applications can determine if a web document refers to a "crow" that is a bird, or the "Crow" that is a Native American tribe. They can merge data for "mass" from different body mass datasets but ignore data related to other meanings of the word "mass".

Although many authors have claimed that OWL and RDF will solve data discovery and integration issues, keen problems in biodiversity science, adoption of these formats has so far been largely limited to computer scientists, database administrators, and highly trained ontologists. The SPIRE project has developed two tools designed to make it easier for individual scientists to convert their information to RDF and OWL. We report on tests from using these tools with biodiversity data.

RDF123 (http://rdf123.umbc.edu/) is a highly flexible open-source tool for transforming spreadsheet data to RDF. It is intended for use with ontologies in any content area. Two RDF123 interfaces are available. The first is a graphical interface that allows users to map their

spreadsheet columns and rows to ontology classes and properties in an intuitive manner. The second is a web service, intended for machine-to-machine communication, that takes as input a Google spreadsheet and an RDF123 map, and provides RDF as output. RDF123 was tested using spreadsheet data from the first annual Blogger BioBlitz in 2007. This biodiversity survey involved sightings of a broad range of taxa in 17 localities in April 2007. We mapped spreadsheet columns to concepts in SPIRE's ETHAN and observation ontologies so that RDF123 could generate OWL representations. The resulting OWL data was posted on the web where it was indexed by Swoogle, the semantic web search engine.

Spotter (http://spire.umbc.edu/firefox/) is a Firefox RDF-based extension designed for observations made by citizen scientists from unstructured sightings of organisms (*e.g.*, in web blog entries, discussions, photo-sharing sites). The user fills out a simple form and pastes a link in their comment or blog entry. By following the link, semantic web crawlers then generate and index the appropriate RDF. We continue to test Spotter on our own blog, http://ebiquity.umbc.edu/fieldmarking, and in cooperation with an environmental education summer camp.

In both RDF123 and Spotter, the RDF data is able to be discovered and integrated, using our TripleShop application or a mapping application, with related data collected in different contexts. For example, it is possible to conduct queries such as "What invasive species were observed in the Blogger BioBlitz?" or "Where have people observed frogs this year?" We found that of 1200 Blogger BioBlitz observations, 47 of them were of species defined as "of concern" by the US Fish and Wildlife Service. We plan to extend this work by taking advantage of existing technologies such as RSS for alerting subscribers to new data of interest on the Semantic Web.

Support is acknowledged from: US National Science Foundation

18.10. Encouraging Users to Share Biodiversity Information

Katja Seltmann[1], Greg Riccardi[1], Austin Mast[1], Fredrik Ronquist[2], Neelima Jammingumpula[1], Karolina Maneva-Jakimoska[1], Steve Winner[1], Deborah Paul[1], Andrew Deans[3]

[1] Florida State University, [2] Swedish Museum of Natural History, [3] North Carolina State University

The Morphbank research project (http://morphbank.net) has created a repository for images that are useful to biodiversity researchers. The project adds significant value to stored images by managing complex metadata, by organizing images for searching, and by allowing users' comments and annotations to be directly linked to images and metadata. Its sophisticated user interfaces allow users to identify new taxa, characters and matrices, and *ad hoc* collections of images and other objects. The productive and dedicated research & development (R&D) team has helped scientists organize and use their images productively on the Web.

Morphbank has more than one terabyte of images covering a broad spectrum of life. It has users and images from diverse collections, taxonomic groups, and research projects. This broad user base has stressed the need for flexibility in the Morphbank system. To succeed in attracting users, recognition of the effort it takes for scientists to participate in Web dissemination activities must be acknowledged in biology research circles. Some examples of helpful strategies are:

- Credit for service. Each user's logo, a list of contributors, and links to major projects are prominently displayed. All are useful in helping to promote the biologists' work and to provide evidence of scholarly activity.

- Big results with little effort. Many of the new users to the system are attracted by the ability to link collections to publications. It's these big results with little effort that will persuade users to participate.
- Helping users learn. Straightforward user manuals, workshops, and a contact phone number help users learn the system. Many need someone to consult and it's important for users to feel welcome. Workshops may also include discussions of the value of data beyond the scope of the researchers' individual projects, and provide important feedback for the Morphbank R&D team.
- Need for control. Biologists who are not skilled in programming often feel controlled by programmers or left out of the process. The differences in the way programmers and biologists approach issues may lead to difficulties. For example, biologists wish to freely edit data in the system, but developers worry that data structure and the ability to modify and augment data needs to be protected and controlled. Biologists agree with this but believe they are the ones best suited to make these decisions, not the creators of the database.

This poster will illustrate how Morphbank influences the users' visions of the data they are importing into the system. There is a correlation between the level of difficulty in getting the data into the system and the perceived value of the data by the biologist. This manifests itself on many levels: 1. desire to keep data hidden, 2. wanting to be sure only the best image is used (unwillingness to have imperfect data), and 3. a feeling of being overwhelmed by the technology. Generally by making access, upload, and maintenance easier, users will feel they have control over the data and will be more apt to share. Thus, effective user interface design is crucial to encouraging participation.

Support is acknowledged from: US National Science Foundation, Florida State University

18.11. Biodiversity Information Infrastructure of the Royal Museum for Central Africa (RMCA)

Franck Theeten, Bart Meganck, An Tombeur, Danny Meirte, Patricia Mergen, Michel Louette

Royal Museum for Central Africa

This poster will describe the technical infrastructure installed at the RMCA and show our contribution to biodiversity information networks such as the Global Biodiversity Information Facility (GBIF, http://www.GBIF.org) and the global network of herpetological collections (HerpNET, http://www.herpnet.org).

This infrastructure is built on the foundation of a combination of several databases and web-services, which were developed and installed at different times, and which must meet both internal needs (curatorial management, extensive taxonomical description for the researcher) and external needs (integration of the data into standards currently in force within the scientific community for the exchange of taxonomical and geographical information).

We will consider the following points:

- A presentation of the data standards used by RMCA for the exchange of taxonomical and geographical information, and their corresponding network protocols.

- The issue raised by the interaction between a legacy system that contains the reference data and providers using modern data standards based on XML in order to connect networks of distributed databases.
- A presentation of the contribution of the RMCA to the development of new protocols such as:
    - SYNTHESYS –NAD 3.7 (Itineraries): Project for the visualisation and assessment of the accuracy of geographical data on collector's pathway. (http://synthesys.africamuseum.be/). This is a GIS project based on the analysis of historical data and whose purpose is the determination of several possible alternative routes followed by a collector during a collecting event.
    - Herpnet: Presentation to GBIF and Herpnet of RMCA's Herpetological collections (http://www.herpnet.org/portal.html, http://data.gbif.org/datasets/provider/147).
    - GNOSIS, a Web Map Server implementation financed by the Belgian Science Policy Office, developed collaboratively by major Belgian scientific institutions (the RMCA, the Royal Belgian Institute of Natural Sciences and the Royal Meteorological Institute) and the IT company GIM (Geographic Information Management) (http://www.gnosis.be).
    - MIDAS: a project managed by the RMCA for the development of a centralized database for taxonomical data, gazetteers and curatorial data, which features a scalable on-line interface, and standardized instructions for taxonomical operations (such as synonymy and check on the availability of names).
- We will review the advantages of the modular architecture which is used in the tools developed by the TDWG (DiGIR, TAPIR and BioCASe): the lifetime of each module is improved as the underlying database containing the data and the application which centralizes the research of data by the means of a common interface and a central register for the indexation of data are kept separated. This architecture allows the contributing institution to retain the intellectual property of its data and to keep the control over its technical implementation.

However, the preparation of the data and the mapping of fields into the web-service may require a considerable preliminary work, and we will discuss techniques improving the development time while keeping the structure of the original data intact with the use of SQL views and functions.

Support is acknowledged from: GBIF, EDIT, SYNTHESYS, Belgian Science Policy Office

18.12. Machine Learning to Produce Structured Records from Herbarium Label Text

Qin Yin Wei, P Bryan Heidorn

University of Illinois at Urbana-Champaign

Digitization of herbarium labels should be more than "scanning images" and perform OCR (Optical Character Recognition) on them. It is necessary to make the data available in a structured format that is easily imported into local databases and made available through data federation frameworks. The HERBIS (http://www.herbis.org) project speeds up herbarium label digitization using supervised machine learning technology. Machine learning (ML) techniques for information extraction are very useful where there are "learnable" patterns in data, but where

the irregularities are varied enough to make hand programming of regular expressions cost-prohibitive. Digitization of herbarium labels is such a domain.

Supervised Learning is a method that "operates under supervision by being provided with the actual outcome for each of the training examples"[1]. In other words, the machine learning algorithms get the knowledge from the examples and then use the knowledge to classify new examples. In order to gain satisfactory performance, the learning schemas (*e.g.*, decision trees, Naïve Bayes) need sufficient numbers of training examples to represent situations likely to occur in real use. The best situation is one where the examples are representative of the testing data, meaning each label element and label type in the full dataset should be represented in about the same proportion in the training set and in the real-world conditions where the ML model will be used[1]. Also, the learner assumes all the information in the training set is correct. If the input data has many errors, the correct learning is impossible. Each error would be misleading or confusing to the learner, leading to faithful reproduction of errors and inconsistencies. For example, if training data is provided from institutions with different policies, (*e.g.*, where one institution codes training data to save only the latest determination while another codes training data to save all determinations) for the use of the data on herbarium labels, contradictory training data can result.

Therefore the best training set should contain correct, representative, and numerous enough training examples. Unfortunately, the digitization of herbarium labels is made difficult by the high variability of label formats, OCR errors, training errors, and the open class nature of some of the elements. "Open class" means there is no reasonably finite set of instances of elements that could be learned by the ML and placed into a lookup table. In HERBIS we extract 36 independent elements of information from these labels. Most of them are Dublin Core metadata. Our training data consists of digitized OCR records from the Yale Peabody Herbarium with multiple label formats randomly selected from the labels. We have developed a Relax NG Schema which contains 36 independent fields, allowing all elements occurrence to be optional, potentially multiple times and in any order as is required by the variability in the input data. Naïve Bayes and Hidden Markov Models are used as learning schemas. Our results indicate that machine learning is encouraging with only a few hundred training examples (several thousand are typically a reasonable number) with an F-score of 86% using the same data set[2]. In order to improve the performance, our future work will add training examples, allow multiple data exchanging schemas, and support multiple machine learning schemas.

[1] Witten, I. H., and Frank, E. (2005). Data mining: practical machine learning tools and techniques (2 ed.). Boston, MA: Morgan Kaufmann Publishers.

[2] In the Information Retrieval domain, F-score is the weighted harmonic mean of precision and recall, where precision is the proportion of retrieved documents which are relevant to all the documents retrieved, and recall is the proportion of relevant documents that are retrieved to the full set of relevant documents available. F-score we used is F=2*Precision*Recall/(Precision+Recall). Generally speaking, the higher F-score, the better results.

Index to Authors