[ISBN will be applied to the post-Conference version]

Biodiversity Information Standards TDWG

The Proceedings of TDWG

Provisional Abstracts of the 2009 Annual Conference of the Taxonomic Databases Working Group

9-13 November 2009
Montpellier, France
(Hosted by Agropolis and Bioversity International at Le Corum)

Edited by Anna L. Weitzman

Published by Biodiversity Information Standards (TDWG)

Montpellier, France, 2009

TDWG 2009 was sponsored by:

To be cited as:

Weitzman, A.L. (ed.). Proceedings of TDWG (2009), Montpellier, France.

This book contains abstracts of the papers, posters and computer demonstrations presented at the Annual Conference of the Taxonomic Databases Working Group held 9-13 November 2009 at Le Corum, hosted by Agropolis and Bioversity International in Montpellier, France.

The meeting attracted more than 275 participants from 30 countries and over 150 prestigious scientific research institutions, museums and companies.

The editor gratefully acknowledges with thanks the vitally important contributions of Lynette Woodburn, Gail Kampmeier, Arturo Ariño, P. Bryan Heidorn, and Lee Belbin towards the editing of this publication. We also appreciate the session chairs and numerous independent peer reviewers who have ensured the quality of the submissions.

Editor's Note: Due to the late submission of many abstracts, not all abstracts have been reviewed or edited. These abstracts are indicated with a ∞ (infinity symbol) after the title. In addition, Table of Contents, Index, and any supplemental files submitted by the authors have not been verified as present. For all abstracts, layout, formatting and proofreading are also incomplete; acronyms have not all been reviewed and sessions have not been checked for order or completion. These things will be corrected in the final version.

Published and distributed as an Adobe® Portable Document Format (PDF) document for free download from the Conference web site at http://www.tdwg.org/conference2009/

Proceedings of TDWG

1. e-Biosphere 09-Outcomes

1.1. UK Roadmap for delivery of Internet-based Taxonomy

Frank Bisby, Malcolm Scoble, Norman MacLeod, Colin Miles, Amanda Read, Walter Berendsohn, Mark Costello, Thomas Orrell
School of Biological Sciences, University of Reading, UK & Species 2000

The UK Government has asked its research funding bodies to prepare policies on various issues that are seen as impediments to successful funding for systematics. One of these is the delivery of Internet-based taxonomy. BBSRC has set up a small panel that includes UK and international participants, to prepare a 'UK Roadmap for delivery of Internet-based Taxonomy'. Included in the remit is that the panel should consult widely at international fora on some of the issues raised, and for this reason the panel was linked to the e-Biosphere 09 discussions, and is now seeking inputs from the TDWG and other communities. Later on in 2010 there will be an open meeting at the Linnean Society of London to discuss the outcomes and report for the Roadmap.

Three particular issues raised so far are i) the vision of where taxonomy should be 5 and 10 years from now, ii) the need to link initiatives with parallel funding across the funding bodies of different countries and continents, and iii) the need to bring together partnerships between funded scientific infrastructures and funded taxonomic research.

Preliminary ideas on where taxonomy should be 5 and 10 years on are these. Five years from now we believe it should be possible to have all known species on Earth effectively available and treated in Taxon Databases, and thus accessible through the aggregative or synthetic species checklist works such as Catalogue of Life via Species 2000 or via ITIS, in WoRMS, in FADA, in ToL, EoL etc. Similarly it should be possible to expose effectively all names in use through the Global Names Architecture now being set up by EoL, GBIF, CoL and the nomenclators. A contentious issue is whether increased productivity promoted by Internet-based techniques and resources may lead to, say, a doubling of monographic output into Taxon Databases by existing levels of expert staffing, and whether these Taxon Databases can be established on a sustainable basis.

There is the suggestion that 10 years from now we should expect automated or semi-automated identification systems involving both molecular and morphological techniques to become available for species in a significant number of the Taxon Databases. We would thus have the existence and concept for each species, but, more importantly for practical work, the ability to make accurate identifications by all, something that will revolutionise accessibility of good biodiversity data.

Discussions about these 5 and 10 year targets for the Roadmap bring two key issues into focus, and it is about these that we seek inputs from TDWG participants at this meeting. First: would it be possible for funding bodies in different countries to establish collaborative arrangements for funding just their part in international or even global programmes? Do any of you in the audience have experience of such arrangements between funding bodies, either in relation to taxonomy, or indeed in relation to other disciplines? Second: in some countries an impediment to funding Internet-based Taxonomy is that this necessarily combines real taxonomic research with the development of a scientific infrastructure. In the UK BBSRC and NERC fund research, but usually not infrastructures. Are there countries in which funding applications for research and infrastructures can be combined on a standard basis – something that the panel presently only know of in the European Commission's programmes?

Support is acknowledged from: The UK Biotechnology and Biological Sciences Research Council (BBSRC)

1.2. e-Biosphere and TDWG ∞

Walter G. Berendsohn, et al.
Botanic Garden and Botanical Museum Berlin-Dahlem

The e-Biosphere 09 conference in July (www.e-biosphere09.org/) brought together more than 500 individuals from 63 countries for an intense three days reflecting the status and future of biodiversity informatics. In the two days following the conference, a smaller group of representatives from several of the larger biodiversity informatics initiatives met to begin to draft a roadmap for building interoperability among their information systems. The Workshop Resolution

(www.e-biosphere09.org/assets/files/workshop/Resolution.pdf) calls for the further development of a broad-based coalition to further the efforts of biodiversity informatics. The Resolution identifies several priority action items, including:
• Creating a seamlessly connected virtual laboratory or platform for integrating, synthesizing, and analyzing biodiversity information;
• Promoting linkages with user communities that would use the platform to better model and understand the entire biodiversity of the globe; and
• Developing and disseminating a periodic outlook report on biodiversity informatics that would assess the status and future of the field.
The Resolution also identifies several initiatives for which workshop participants agreed to take leadership roles. These include:
• Creating durable, global registries for the resources that are basic to biodiversity informatics (e.g., repositories, collections);
• Completing the construction of a solid taxonomic infrastructure;
• Creating ontologies for biodiversity data;
• Developing an approach to the citation of published data and information services; Implementing active and effective outreach to the policy and research domains that rely on biodiversity informatics as a resource.
The workshop participants readily acknowledged that their meeting was only the beginning of a much broader effort. However, hopes that a process of community-wide consultation and communication could be facilitated by the e-Biosphere Online Conference Community (OCC) have not materialised. The purpose of the discussion at TDWG 2009 is to try to define TDWG's role in the process. For this, an introductory session will be held in the beginning of the conference, and individual sessions throughout the TDWG meeting will make a few minutes room to reflect on how their activities relate to the overall e-Biosphere roadmap effort. The results will be gathered and presented in the end of the conference, ideally leading to a statement clarifying TDWG's role.

2. Sharing e-knowledge on agricultural diversity worldwide

2.1. CONABIO's experience: Sharing e-knowledge, solutions and lessons learned ∞

Patricia Koleff, Cecilio Mota, Francisca Acevedo, Oswaldo Oliveros, Claudia Sánchez, Israel Martínez
National Commission for the Knowledge and use of Biodiversity of Mexico (CONABIO)

Biodiversity is an essential element for the development of food and agriculture production strategies. Genetic diversity of crops plays a critical role in increasing and sustaining production levels and nutritional diversity throughout the full range of different agroecological conditions. Free and open access to knowledge on crop science is crucial to the sustainable management of resources. Despite the wide range of available data, there are still large information gaps for countries that are centers of origin and diversification of crops, and some results of scientific research are not easily available or updated.
The National Commission for the Knowledge and use of Biodiversity of Mexico (CONABIO) has more than 15 years of experience in compiling standardized information for the National Biodiversity Information System (SNIB), based on the primary data from scientific collections (Herbaria and Museums), used as backbone for taxonomic and geographic references. Since 2006 CONABIO coordinates a large scale project to digitize and collect new data about maize and its wild relatives with the participation of numerous national institutions.
Data compilation is done using the software Biotica© based on a relational model in MS Access that normally manages nomenclatural, curatorial, geographical, bibliographical and ecological information. The system has been used over the last six years to handle data from Genetic Modified Organisms (GMO) and their wild relatives, adjusting fields to detect variation within seeds, infraspecific taxonomic categories, management techniques and use of resources, for all species whose center of origin, diversification or domestication is Mexico. It includes 1087 fields for primary diversity data and 305 new fields specific to the characterization of maize and its wild relatives. Information is gathered through external projects.
We currently have 15,300 records of maize and its relatives on hold in different national collections. New localities have been explored across different agricultural regions in the country. Up to July 2009, records for 6,355 new maize, 361 teocintles and 421 Tripsacum have been stored. Georeferenced records in gene banks represent different size areas or processes, depending on race, adaptation, uses, social perceptions or value in the regional market. Preliminary analysis shows the presence of native landraces of maize in almost all agricultural regions of the country, except in deserts in Northern Mexico, some remote mountain areas and some irrigated agricultural regions. New varieties and types of maize, new populations of teocintles, a new population of teocintle perenne (presumably a new species) and better

knowledge for species of Tripsacum have been documented. The analysis of three fields (grain color, maturity season, adaptation to soil and precipitation) has provided insights on group distributions and racial complex. So far, the results obtained have yielded a 200% increase of maize records, discovered new varieties, and provided detailed information allowing the development of diverse analysis. The data has also provided risk assessment analysis for GMOs intended as experimental crops in Mexico.
To document variation within varieties in maize and its wild relatives, essential in assessing potential risks by GMOs, it is necessary to have a standardized database, especially with different participants and sources.

Support is acknowledged from: CONABIO

2.2. Biodiversity Informatics and co-Operation in Taxonomy for Interactive shared Knowledge base (BIOTIK)

B.R. Ramesh[1], N. Ayyappan[1], D. Balasubramanian[1], P. Grard[2]
[1] French Institute, Pondicherry, [2] CIRAD

Human pressure on environment is steadily increasing and sustainable environmental management methods are engaging new technologies. In this context, BIOTIK is an initiative in the emerging area of biodiversity informatics, which has developed an ICT knowledge base centered around a tree species identification system for the two hotspots of biodiversity in South and South-east Asia: 1) the Western Ghats of India and 2) the North Anamites Mountains of Lao PDR.

The existing multimedia species identification software "Identification Assistée par Ordinateur (IDAO)" was implemented to generate identification tools for tree species of each region independently. The identification system allows a user to identify a species through a visual interface, complete with graphical representations of botanical characters and their different states. It builds a virtual tree on screen, based on the character states selected by the user and also suggests possibilities for missing or erroneous information. The system works on an algorithm that operates in the same way as experts work in the field, i.e. by selecting a random sequence of observable characters. At each step of the identification process, a similarity coefficient is calculated, which compare the set of characters supplied by the user to a database of reference species descriptions. Such databases along with digital images of live specimens were created from extensive fieldworks conducted in both the hotspots. Images of reference herbarium samples were also incorporated. Once the species has been identified, it provides a resume of botanical and ecological information of the species in English (with hypertext glossary) as well as in local languages besides the photographs of characters. Particularly the use of local languages would enable wider dissemination of taxonomic knowledge and enhancement of biodiversity assessment capabilities of the two regions.

The knowledge base is available in a MS-Windows desktop version, a portable version for Ultra-Mobile PCs and also an open-to-all browser based web application (http://www.ifpindia.org/biotik/).The web application is unique in its implementation using scalable vector graphics (SVG) for image representation, php for scripting and mysql for backend database.

Support is acknowledged from: EU

2.3. Agropolis global presentation of Pl@ntNet

Daniel Barthélémy[1], Nozha Boujemaa[2], Daniel Mathieu[3], Jean-François Molino[4], Pierre Bonnet[5], Raffi Enficiaud[2], Elise Mouysset[3]
[1] INRA - Amap, [2] INRIA, [3] Tela Botanica, [4] IRD, [5] INRA

Accurate knowledge of the identity, geographic distribution and uses of plants underpins the success of agriculture and biodiversity conservation. Because their floras are still badly described and documented, most Mediterranean and tropical countries are facing a major impediment towards sustainable development. It is thus of the utmost importance to speed up the accumulation of basic data on plants, while simultaneously providing an easy and efficient access to this knowledge to potential users.
The Pl@ntNet project will contribute to this objective by providing, within a coherent platform, free, open-source, easy-access software tools and methods for plant identification and for the collection, management, sharing and exploiting potentially all kinds of data on plants. It will deeply involve citizen science, as a powerful means to enrich databases with information on plant location, phenology, ecology or uses, thus compensating for the current lack of professional botanists.
These tools, partly based on well established, or prototype software (e.g. http://umramap.cirad.fr/amap2/logiciels_amap/index.php?page=plantnote), will be proposed for individual or collaborative use, to all potential users (from general public to amateur or professional taxonomists, from farmers to

agronomists and biodiversity managers). Plant identification systems will include not only existing, morphological recognition systems (http://umramap.cirad.fr/amap2/logiciels_amap/index.php?page=idao), but also more experimental, yet potentially powerful content-based image retrieval methods (http://www-rocq.inria.fr/imedia/).
Pl@ntNet is the first flagship project of Agropolis foundation, and intends to complement other international initiatives on plant biodiversity and taxonomy.

Support is acknowledged from: Agropolis fondation

2.4. Landscape of the information standards for plant genebanks

Theo van Hintum
Centre for Genetic Resources, The Netherlands

Genebanks, conserving plant genetic resources (PGR) for use in plant breeding and crop research, have been around since the 1960's. Documentation was done in books, on cards, and other local solutions. Since the computer became accessible, these systems were converted to, in the best case, simple databases, but more often spreadsheets. All these systems had their own local solutions in regards to the structure of the data and the coding used in it. It was only when the genebank community started exchanging data that the need for standardization arose. The first data being exchanged, and so far the only, were passport data: data describing the identity and origin of the samples. On the basis of these data so called central crop databases were created, giving an overview of all PGR accessions of a crop and its wild relatives maintained in Europe or the entire world. To facilitate this exchange, the so-called Multi-Crop Passport Descriptor List (MCPDL) was compiled. This list was nothing more than a list with 28 commonly used descriptors including descriptors identifying the maintaining institute, the local ID, and the IDs given by the collecting expedition and/or the donating institute, but also the taxonomic classification, information about the location of collecting, etc. The use of some codes was required, including some defined codes defined by the MCPDL (for the biological status of the accession, for a classification of the collecting or acquisition source and for the type of storage), but more importantly two externally maintained coding systems: the 3-letter ISO 3166 country codes and the Institution Codes as maintained by the FAO. The first system worked relatively well, although the fact that it did not maintain historical codes, such as the one for the Soviet Union, caused problems, since these appeared and still appear frequently in the datasets. The second system, the one for institution codes, proved more problematic. It appeared incomplete and poorly maintained because of the bureaucratic way it was managed. However the MCPDL created solutions for both problems, by extending the standard ISO 3166 list and by allowing an alternative for institutes without institution code.

When the European genebank community started sharing passport information on a routine basis in the EURISCO database, the MCPDL was adopted for data exchange extended with a few new descriptors mainly for administrative purposes and to link the PGR accession to additional information. Now users have access to passport data of c. one million accessions maintained in Europe. Recently, initiatives to include characterization and evaluation (C&E) data in this database were taken. The issues regarding data quality in EURISCO and regarding the complexities due to lack of standardization of C&E data will be discussed.

2.5. Value of a coordinate: geographic analysis of agricultural biodiversity

Andy Jarvis[1], Julian Ramirez[2], Nora Castañeda[2], Samy Gaiji[3], Luigi Guarino[4], Hector Tobón[5], Daniel Amariles[5]
[1] International Centre for Tropical Agriculture, CIAT, Cali, Colombia and Bioversity International, Regional office for the Americas, Cali, Colombia, [2] International Centre for Tropical Agriculture (CIAT), Cali, Colombia, [3] The Global Biodiversity Information Facility (GBIF) Copenhagen, Denmark, [4] Global Crop Diversity Trust, Rome, Italy, [5] International Centre for Tropical Agriculture, CIAT, Cali, Colombia and Universidad de ICESI, Cali, Colombia

The presentation demonstrates the value of a geographic coordinate for exploring ecogeographic patterns, understanding the biology of a species, and ensuring adequate conservation policies for a given genepool. We focus the presentation on the use of geographic analyses on agricultural biodiversity, and specifically on crop wild relatives. Primary biodiversity records accessible through the Global Biodiversity Information Facility (GBIF) portal can be coupled with modeling and earth observation to assist in decision-making to conserve wild genetic resources. This is critical to maintaining food security in the face of significant threats to genetic diversity including from climate change.
Wild relatives of modern and traditional crops have proved to be a useful source of genes for crop breeding for developing resistance to a range of biotic and abiotic stresses. The wild relatives harbor an abundant supply of useful genes, and demand for wild relatives in crop improvement programmes is on the increase. A significant proportion of these invaluable genetic resources have already been lost due to anthropogenic activities and these pressures are increasing.
First, we present a gap analysis on 17 wild genepools from 16 important crops worldwide: cassava, groundnut, potato, rice, chickpea, common bean, barley, cowpea, wheat, maize, sorghum, pearl millet, finger millet, pigeon pea, faba bean,

and lentil. The gap analysis identifies the state of conservation for these species (in situ and ex situ), and prioritises possible sites for future collecting. We analyzed 28,751 herbarium and genebank species occurrences accessible through GBIF for 643 wild taxa belonging to the 17 genepools. We used a maximum entropy climate envelope model to create distribution maps for each species under current climates using the WorldClim database. The results show the species and geographic regions where high priority for genetic resource collection exists. Full results are available at http://gisweb.ciat.cgiar.org/gapanalysis/.
We then present analyses on the impacts of climate change on these 17 wild genepools. The likely shift in geographic distribution is performed using 18 statistically downscaled Global Climate Models (GCM) under the emission scenario A2 (business as usual). The results are used to indicate areas that should be further prioritized for either ex situ and in situ conservation or both.
The presentation concludes by discussing the issues of data quality in geographic information in plant collection databases, and some of the means by which International Centre for Tropical Agriculture (CIAT) and GBIF are going about providing scripts to check errors and correct erroneous coordinates.
A concerted conservation effort is needed to address the possible extinction risks that may come from climate change, particularly in wild species and especially for those taxa more likely to provide novel adaptations to biotic and abiotic stresses or those that are not adequately represented in genebanks. Ex situ conservation efforts should start in those areas that currently hold a considerable amount of diversity, and in which changes in species diversity will be of high significance. Evaluations of the status of ex situ collections and in situ conservation (gap analyses) of wild relatives of important food crops and vulnerable wild genepools are also needed to assess priority species and areas to be conserved either ex situ or in situ.

Support is acknowledged from: Global Crop Diversity Trust; World Bank GPG2;[9:33:22 PM] Andy Jarvis: UNEP-GEF

3. Global networks

3.1. The GBIF Global Names Architecture ∞

David P Remsen, Markus Döring,
GBIF

All information about a species is tied to a scientific name. Scientific names are labels for taxa, and taxon definitions are expressed in taxonomic catalogues and monographs, that provide direct or indirect circumscription information. This may include additional names and synonyms that are no longer accepted for the taxon. The informatics challenges presented by taxon names are well-known but effective mechanisms for addressing them are not part of a uniform biodiversity informatics landscape. One of the more effective mechanisms for addressing the names problem in biology is in promoting the use of taxonomic identifiers, instead of a scientific name alone, as the means for labelling an information or data object related to a taxon.

In order to facilitate this practice, effective and simple mechanisms for publishing information about taxa and names, starting from originating nomenclatural acts to subsequent taxonomic catalogues, need to be in place. Furthermore, access to these published data, given its wide applicability in managing information about species, should be consistent, liberal , and resolvable to authoritative sources . This requires common data exchange standards, scalable and low-latency access methods, and a global discovery mechanism for published resources.

The Global Biodiversity Information Facility (GBIF) has worked in concert with other initiatives to develop common methods for publishing and accessing taxonomic information and tying it to information about species. These efforts include the refinement and adoption of new Darwin Core taxon terms and simple text-based publication formats to promote simplified and low-latency publication of taxonomic data, the development of tools and methods for publishing data in these formats, and the use of the entire GBIF informatics suite to enable uptake among different stakeholders. A common registry with the ability to filter registry entries into sub-networks provides the basis for a global architecture.

GBIF is proceeding to use this architecture to promote the publication of taxonomic catalogues, nomenclatural data, and species inventories. We are extending our indexing capacity to include a comprehensive index of these published resources and developing services that allow them to be referenced and facilitate the use of taxonomic and nomenclatural identifiers within new synthesised resources and in the annotation of primary biodiversity data.

3.2. The Global Invasive Species Information Network: Its latest advances and its barriers

Michael Browne[1], Jim Graham[2], Christine Fournier[3], Annie Simpson[3]
[1] Global Invasive Species Information Network, [2] Colorado State University, [3] National Biological Information Infrastructure

In 2004, funding from the US State Department initiated a global network for managing freely-available, digitized, invasive species information: the Global Invasive Species Information Network (GISIN) (Simpson 2004). The need for such a global network had been described for at least five years prior to receiving funding, with experts claiming, quite correctly, that "regional databases are insufficient for tracking harmful species that can suddenly appear in areas far from their native ranges" (Ricciardi et al. 2000). Now five years after its initial start-up, with a new URL (http://www.gisin.org), the GISIN is on the brink of broad implementation of a system to cross search data in a way that is compliant with the TDWG Access Protocol for Information Retrieval (TAPIR) (Graham and Jarnevich 2008) and compatible with the Global Biodiversity Information Facility (GBIF) Integrated Publishing Toolkit (IPT). Since 2005, four standards working group meetings have been held with additional seed funding from organizations such as GBIF, the Group on Earth Observations (GEO), the Convention on Biological Diversity (CBD), and the US Geological Survey's National Biological Information Infrastructure, yet inadequate funding is still the largest barrier to full implementation of the GISIN (Simpson et al. 2006). With the limited funding that has been made available to date, the GISIN standards meetings produced six data models for a system to share invasive species information. The GISIN system builds on the components of Darwin Core and Dublin Core, but defines additional concepts that are important to invasive species science and are not included in either of these standards (see: http://gisin.org/cwis438/websites/GISINDirectory/tech/Protocol_Home.php?WebSiteID=4). Services provided by other biodiversity organizations are very important to the eventual success of the GISIN system. The GISIN will use the GBIF registry for the generation of organization codes to create Globally Unique Identifiers (GUIDs), is considering the Global Names Architecture (which will tap into disambiguation databases such as the Catalog of Life and the Global Names Usage Bank) for scientific name disambiguation, and uses its own invasive species on-line databases list for data provider recruitment (Sellers et al. 2004). For better system performance, a data cache has been implemented (Graham et al. 2008). Although toolkits in the programming languages PHP, ASP, and the IPT (which is java-based) are available, for the easiest implementation, GISIN data providers are encouraged to post their data for upload to the cache as flattened, tab-delimited text files. As agreed at the TDWG annual meetings a year ago, the GISIN strives to be a viable test case for TDWG standards and protocols, and we seek further advice from TDWG members about our implementation methods. To this end, an invasive species working session will take place at TDWG 2009, and depending on participant numbers and expertise, will work on the problems data providers might have with the various tools and generate ideas for solutions, for example: 1) mapping the GISIN protocol concepts to the GBIF IPT star schema; 2) the data cache protocol; and 3) flat text file upload vs. Web services. The major problems and the solutions found will be reported back to the plenary on Friday and used in the improvement of the GISIN system.

Graham J, Jarnevich C. 2008. Building a TAPIR-Lite Toolkit for the Global Invasive Species Information Network (GISIN). Proceedings of TDWG. Accessed 25Sep09: http://www.tdwg.org/proceedings/article/view/401
Graham J, et al. 2008. Vision of a Cyberinfrastructure for Nonnative, Invasive Species Management. BioScience 58(3): 263-268. Accessed 07Oct09: http://www.gisinetwork.org/pubs.html
Ricciardi A, Steiner WWM, Mack RN, Simberloff D. 2000. Toward a global information system for invasive species. BioScience 50: 239–244. Accessed 24Sep09: http://sgnis.org/publicat/papers/riccstei.pdf
Sellers E, Simpson A, and Curd-Hetrick S. 2004. Draft List of Invasive Alien Species (IAS) Online Databases and Databases Containing IAS Information. Accessed 24Sep09: http://gisin.org/cwis438/websites/GISINDirectory/DatabaseDirectory_Table.php
Simpson A, Sellers E, Grosse A, Xie Y. 2006. Essential elements of online information networks on invasive alien species. Biological Invasions 8: 1579–1587. Accessed 24Sep09: http://i3n.iabin.net/documents/pdf/Online_info_networks.pdf
Simpson A. 2004. The global invasive species information network: What's in it for you? BioScience 54: 613–614. Accessed 23Sep09: http://gisin.org/WebContent/cwis438/GISIN/Documents/04_July_Viewpoint_Simpson.pdf

Support is acknowledged from: US Geological Survey, US National Biological Information Infrastructure, Global Biodiversity Information Facility, Secretariat of the Convention on Biological Diversity, Group on Earth Observations, The Polistes Foundation, Colorado State University.

3.3. The Global Biodiversity Information Facility (GBIF): The decentralised architecture

Samy Gaiji, Éamonn Ó Tuama, David Remsen, Vishwas Chavan, Tim Robertson
Global Biodiversity Information Facility

In advancing the Global Information Facility (GBIF) "from prototype to full operation", the community recognises the need to move to a more distributed and decentralised model based on the active engagement of more self-sufficient

participants nodes.

Such logical evolution of the GBIF network architecture is aimed first at increasing its capacity to rapidly mobilise and share a larger amount of biodiversity related information covering not only the existing taxon point occurrence data records but also other data types such as spatial, multimedia, names, and associated metadata.

To achieve this, the GBIF Secretariat developed in 2008 a first blueprint of its decentralisation strategy, which was presented at the TDWG 2008 Annual Conference. To achieve this ambitious evolution of its network architecture, the focus has been on simplifying the process of contributing data from existing and new data publishers as well as improving the indexing frequency. Since then, GBIF has been actively engaged in delivering the first core components of its Informatics suite, namely: the Integrated Publishing Toolkit (IPT), the Harvesting and Indexing Toolkit (HIT) and the Global Biodiversity Resources Discovery System (GBRDS).

Towards decentralising its architecture, GBIF has made some radical and fundamental decisions in its approach to information networking. First, it has recognised that the existing information retrieval protocols based on federated search (e.g. TAPIR, DIGIR, BioCASe) were unsuitable for the scale of growth expected within a global network. More importantly, GBIF recognised that the actual Registry implementation based on UDDI was not scalable or rich enough to meet the growing needs of a network that requires discovering much more than single endpoints. Through the GBRDS, GBIF intends to bring forward the concept of a biodiversity network compass to be available for use by all informatics initiatives to discover, locate and index a large variety of biodiversity data resources, services, schemas etc… Finally, to enable successful interoperability and to facilitate accurate citation of data provenance, GBIF will provide solutions for the use of persistent and stable identifiers.

This presentation will provide an overview of the GBIF decentralised architecture and focus on how other informatics initiatives and networks (e.g. agro-biodiversity, invasive species etc.) could fully benefit from these achievements.

3.4. EDIT: experience and impact

Dave Roberts
The Natural History Museum

The European Distributed Institute of Taxonomy (EDIT; http://www.e-taxonomy.eu) is the collective answer of 28 leading European, North American and Russian institutions to a call of the European Commission, issued in 2004, for a network in "Taxonomy for Biodiversity and Ecosystem Research". The EDIT consortium agreement started on the 1st of March 2006 and will last 5 years. The underlying purpose was to change the way that taxonomists work in Europe, to be achieved by museums agreeing common strategic goals and policies in key areas of activity to build collaboration, rationalising various tasks and working to common standards, and integration of taxonomic work at researcher's level. We found that taxonomic institutions operate in quite different ways in terms of governance, funding models, strategy and policy. Indeed, on a detailed level, their written policies had very little in common beyond their taxonomic output. Surprisingly the first time that the European taxonomic institutions sat together to discuss common problems and approaches in detail was through the committees initiated by EDIT. Moreover, it was through EDIT's board of Directors that so many taxonomic institutions had been represented for the first time at directorial level. We now consider that much more than the 5 year timeframe available to EDIT is needed to form common policy as a mechanism to federate taxonomic institutions. Certain areas have proved amenable to integration, particularly the care of collections and IT.
A primary driver towards collaboration and standardisation was the introduction of computer-based tools, specifically the Internet Platform for Cybertaxonomy and Scratchpads. The Internet Platform for Cybertaxonomy is a large software initiative that provides a coherent workflow from data creation to publication, specifically aimed at the needs of taxonomists; it has, of course, taken time to create. Scratchpads were created in response to the immediate need to engage users and publish data to the web. They are an agile development (http://en.wikipedia.org/wiki/Agile_software_development) which continues to evolve functionality while offering a developmental framework in which individual communities evolve their own products.
Scratchpads provide a data publication platform, and the Platform for Cybertaxonomy, a suite of tools. The former are only loosely structured while the latter are properly structured, but that structure is hidden from the user. Both are very lightly branded and data remain the property and responsibility of the data provider. Experience with the Scratchpads, which has the larger user base, will be discussed in the light of a sociological survey that has just been conducted. A key observation was that the barrier to uptake of this new way of working was primarily sociological rather than technological.

3.5. The Global Biodiversity Information Facility – strategic objectives and perspectives on building the biodiversity informatics commons

Nick King
Global Biodiversity Information Facility

Current global environmental changes are unprecedented in cause, in that they are increasingly understood to be man-made. As the architects of these changes, we also need to find the solutions. Finding solutions will not be simple if we do not have enough scientifically-sound data readily accessible, nor the tools to analyse these, in order to effect improved policy- and decision-making in response to climate change and other major drivers. Mobilising disparate yet related datasets worldwide to create global environmental datasets is increasingly made possible through IT developments, and in particular, mobilising the estimated billions of analogue primary biodiversity records already in existence around the world is a critical component of establishing baseline knowledge of species and ecosystems, allowing time-series analyses of changes and modelling of future environmental trends to improve decision-making. To date however, progress in discovery and mobilisation of these primary biodiversity data has been slow and largely opportunistic in approach.

The Global Biodiversity Information Facility (GBIF) is an inter-governmental organisation (of currently 54 countries and 43 international organisations) set-up and paid for by countries, in recognition of their common needs in this arena and to help build a common research infrastructure to mobilise the required data. GBIF has been instrumental in catalysing agreements on many of the standards, protocols and infrastructure components required to make disparate datasets compatible, discoverable and accessible worldwide. As of October 2009, some 190 million records from 8000 datasets from more than 280 institutions worldwide are now accessible through the GBIF data portal (data.gbif.org). Whilst data are increasingly captured digitally, less than 40% of these records are collection specimen-based - this calls for expediting digitisation and access of an estimated more than 2 billion specimens housed in natural history collections worldwide.

To this end GBIF is developing various strategies and tools, including a 'Global Strategy and Action Plan for mobilisation of natural history collections data (GSAP-NHC)', and, to facilitate discovery of biodiversity data resources, a Global Biodiversity Resources Discovery System (GBRDS). Earlier in 2009, the Integrated Publishing Toolkit (IPT) was launched, allowing for efficient sharing and hosting of occurrence data, taxonomic and nomenclatural checklists, and general dataset metadata.

Thus, the global biodiversity commons is now a reality, allowing access to previously inaccessible records and datasets, and analyses which were previously impossible. With 'proof of concept' secured, in order to enhance our ability to analyse planetary change and find solutions, there is an increasing need for members of professional societies such as TDWG not only to continue to participate and drive development of this common infrastructure, but also to communicate and encourage uptake throughout the wider biodiversity informatics community to contribute and take full advantage of this emerging 'global good' infrastructure to expedite the mobilisation of biodiversity data from all sources. The call is made for participation by all who hold and collect biodiversity information to make their data available. In addition, the call is being made to all funding agencies, both public and private to mandate within their grants that the data generated are captured to these global standards and made freely available.

4. Ontology and Life Science Identifiers: The state of the play

4.1. Vocabularies - Managing Them ∞

Kehan Tristam Harman[1], Roger Hyam[1], David P Remsen[2]
[1] Natural History Museum, London, [2] Global Biodiversity Information Facility

Biodiversity science and applications require community managed vocabularies which give some kind of definition of terms, how they relate to one another and ways of locating further information on them. Both within TDWG (Biodiversity Information Standards) and more generally there are standardised lists of terms that have been developed - these include for example the ISO (International Standards Organisation) lists of country (ISO:3166) and language (ISO:639) names and the TDWG Geographic Regions. Having these standardised lists makes it possible to attach information to uniquely identified terms. Community management of vocabularies can be tricky from a social perspective and it is difficult to find an optimal workflow. A tool has been developed by GBIF (Global Biodiversity Information Facility) on top of the Scratchpads (http://scratchpads.eu) framework which allows communities to collaborate on both the development of new vocabularies/thesauri and the enrichment of existing vocabularies by adding multilingual translations of the normative forms that standardised lists require. In addition to this vocabularies can be extended with user defined fields. Work is under way to make these vocabularies compatible with developments in the TDWG Ontology. The tool is accessible at http://vocabularies.gbif.org.

Support is acknowledged from: Global Biodiversity Information Facility; Natural History Museum, London

4.2. The TDWG Life Sciences Identifiers Applicability Statement

Ben Richardson
Western Australian Herbarium - DEC

A review of the TDWG Life Sciences Identifiers Applicability Statement is in progress, having been re-initiated in May this year.
The Applicability Statement seeks to provide guidance on how to use Globally Unique Identifiers (GUIDs) and, in particular, Life Sciences Identifiers (LSIDs) in the biodiversity informatics community.
The review process has provided the following:
• A separation of the Statement into two documents, a top-level GUID Applicability Statement containing recommendations for a successful implementation of GUIDs in general, and an Applicability Statement specific to the implementation of LSIDs. As LSIDs are one form of GUID technology, it makes sense to deal separately with the technical requirements for GUIDs in general and the specific case of LSIDs.
• A general consensus for the adoption of GUIDs.
• A general disagreement over which one of the GUID technologies is most suitable.
• The metadata for a GUID, irrespective of technology, should be represented as Resource Description Framework (RDF) formatted in XML. The implication is that we have a completed RDF-based vocabulary for our data.
• There are some technical impediments to the implementation of LSIDs, including:
• Their more complex hosting requirements, as compared to HyperText Transfer Protocol Uniform Resource Identifiers (HTTP URIs).
• The need for extra effort to support the semantic web community, and the emerging trend in the TDWG community to the adoption of semantic web technology.
• The data versus metadata problem for biodiversity data.
• That there is likely to be more than one GUID technology in use for life science data.

4.3. Persistent Identifiers for Biodiversity Informatics: Status and Likely Directions

Greg Riccardi[1], Richard White[2], Phil Cryer[3], Roger Hyam[4], Chuck Miller[3], Nicola Nicolson[5], Éamonn Ó Tuama[6], Rod Page[7], Jonathan Rees[8], Kevin Richards[9]
[1] Florida State University, [2] Cardiff University, [3] Missouri Botanical Garden, [4] Natural History Museum, London, and PESI, [5] Royal Botanic Gardens, Kew, [6] GBIF, [7] University of Glasgow, [8] Science Commons, [9] Landcare Research, New Zealand

The Global Biodiversity Information Facility (GBIF) has identified the provision of identifiers for biodiversity objects as one of the central challenges to developing a global bioinformatics infrastructure. One of the stated goals in the GBIF strategic plans document "GBIF Plans 2007 – 2011 from prototype towards full operation" (http://www2.gbif.org/strategic_plans.pdf) is to consolidate the underlying enabling infrastructure and standardization for global connectivity of biodiversity data and information through an activity to "develop a system of globally unique identifiers and encourage their use throughout biodiversity informatics". The GBIF plans envisage using TDWG standards to "allow all data objects to be identified using standard actionable globally unique identifiers" and provision of a GBIF web service and user interface to allow users "to locate and view any data object with a standard globally unique identifier".
GBIF convened a task group, the "Life Science Identifiers (LSID) and Globally Unique Identifiers (GUID) Task Group" (LGTG) to explore the issues and offer recommendations on the way forward, with particular reference to the GBIF network that will enable GBIF to provide architecture leadership and best practices for implementation. The principal objective of the group was to provide recommendations and guidelines on deployment of identifiers on the GBIF network with particular reference to the potential role of GBIF as a stable, long term provider of identifier resolution services.
The report of the task group is available at http://biodiversity.cs.cf.ac.uk/gbif/PersistentIdentifiers.html
The task group focused on two over-arching use cases that make identifiers effective for users:
• Uniqueness of reference: An identifier can be used to aggregate information about the identified object. For example, information received from multiple sources associated with a single identifier is information about a single object.
• Action: An identifier can be used to find further information about the object, concept or data to which it refers. This information might be interpreted directly or used to support services.
Effective identifiers make a vital contribution to facilitating the use of biodiversity data by software agents, so that data can be used by and become embedded in an unlimited number of future information systems, as the world moves towards Web 2.0, the Semantic Web, Linked Data and the e-Science Grid.
The LGTG recommended that the GBIF Secretariat take action in support of persistent identifier technologies by taking a leadership role in driving the application and use of identifiers, providing educational and promotional programs,

encouraging the use of appropriate identifier technologies, in particular LSIDs and HTTP URIs, and demonstrating good practice in its own use of identifiers in the GBIF data portal.
The presentation of the report of the task group will include an overview of the characteristics of effective identifiers and of the recommendations for action by GBIF.

Support is acknowledged from: Global Biodiversity Information Facility (GBIF)

4.4. Linked Data

Roger Hyam
TDWG

The implicit goal of biodiversity informatics is to join up the world's biodiversity data: to allow researchers and decision makers access to data from multiple sources that they can combine with their own data for analysis and decision making. For this to occur, data around the world must be both retrievable in a standardised form and also cross linked. Species occurrence data, for example, includes notions of place and taxon that are not fully expressed in the occurrence record itself but must be linked to. There has been much debate as to how we can achieve this linking together. During this debate the World Wide Web has been growing. Google now estimate that there are around 1 trillion (1 million million) resources on the web. This is a proven, scalable system that could be used to link up biodiversity data, indeed all kinds of data.
Tim Berners-Lee, the inventor of the web, advocates the publication of data using standard web technologies through a set of best practices collectively termed "Linked Data"[1].
The basic concept of Linked Data is that every data resource, and indeed every real world object of interest, should be given an HTTP URI (HyperText Transfer Protocol Uniform Resource Identifier often considered synonymous with a URL – Uniform Resource Locator) as a web name. When these HTTP URIs are looked up on the web they give back useful information. If the user is human, they get data back in the form of a web page or other human readable resource. If the user is a machine, they get data in RDF (Resource Description Framework) format. Importantly, the data contains more links to other data objects thus building a global web of information objects just as the World Wide Web today is a web of documents.
Linked Data uses well established standard technologies, principally HTTP and RDF. It is widely documented with tutorials and validation mechanisms. From the point of view of TDWG the advantages of the Linked Data approach are:
- It is easy to implement at a technical level because it uses standard web technologies all developers are familiar with.
- It should be easy to write client applications because the data resolution mechanisms are built into all web aware devices.
- If more complex technologies, such as LSIDs, are required in the future they can easily be layered on top of HTTP URIs.

From the point of view of specialisation of Linked Data for biodiversity data, all that is required is the definition of a set of vocabulary terms and a system for managing of those terms. TDWG has already achieved this for our key resources.
In conclusion Linked Data is an approach that answers many of our problems concerning the representation and exchange of biodiversity data and is already being implemented within some major projects. For example, PESI (the Pan European Species Infrastructure) are proposing its adoption as a standard means of exchanging taxonomic hierarchies.
[1] http://linkeddata.org/

5. EDIT Cyberplatform

5.1. ENHANCING THE VISUALIZATION OF BIOLOGICAL DATA: EDIT GEOGRAPHIC TOOLS

Pere Roca[1], P. Sastre[1], J. M. Lobo[1], B. Meganck[2], Franck Theeten[2], P. Mergen[2], Andreas Müller[3], A. Kohlbecker[3], S. Dusan[4], D. Mikiewicz[4]
[1] Museo Nacional de Ciencias Naturales (CSIC), [2] Royal Museum for Central Africa (RMCA), [3] Botanical Garden and Botanical Museum (BGBM), [4] Hungarian Museum of Natural History (HNHM)

The use of the vast amount of data compiled by thousands of taxonomists from all over the world during the last two centuries is basic for scientific and applied purposes. The emergence and progress of Geographic Information Systems (GIS) and the development of technologies for the storage and access of primary biological data provide new opportunities for taxonomists. However, many taxonomists still have difficulty managing GIS technologies. The European Distributed Institute of Taxonomy (EDIT) application, among others, aims to deliver simple and easy to use online visualization tools to help taxonomists in the spatial representation of their biological data.

The team devoted to developing the Geographic Platform Component of the EDIT application is assisting taxonomists by:

1) developing an easy-to-use web-application (MapViewer) allowing displaying, analyzing and printing of georeferenced information directly from simple data sources. Planned features of the MapViewer application are the provision of print-quality maps (over 600dpi), depicting the probable location of both well- and poorly-surveyed regions, and a direct link to GBIF (Global Biodiversity Information Facility) online tools to plot data occurrences.

2) building generic mapping services (MapRest Services) providing distribution maps from a URL (Uniform Resource Locator). Users may choose among a large range of symbolization parameters. To use these services, the user simply needs to construct the URLs and the MapRest Services generate a map with the selected features, symbolized as the user chooses. The current implementation only accepts TDWG regions (at all levels) as the distribution layers to be printed. MapRest Services can return an image or a file that can be used by a client-side webmapping application (like Openlayers).

Point data can also be plotted from coordinates using MapRest Services. Some All Taxa Biodiversity Inventories (ATBI) sites like Mercantour/Alpi Marittime and Gemer are currently using it.

Some EDIT dataportals using these tools to map taxa distributions are PalmWeb and Chichorieae

These tools differ mainly in the degree to which human intervention is involved in handling the data. While EDIT MapViewer works with user data, EDIT MapRest Services accept feeds from other EDIT web services.

However, these tools share the use of Open Source tools and can be easily linked.

All the geo-software used is OpenGIS Consortium (OGC) compliant: Geoserver and MapServer provide map images, legends and scale bar. Openlayers lets users visualize and navigate this geographic information. The JQuery javascript framework and some of its plug-ins provide dynamic tools to the EDIT MapViewer user interface.

On the server side, PHP interacts with PostGIS to upload point data, dynamically creates the legend and, when requested, performs spatial analysis on EDIT MapViewer. ImageMagick produces images in different formats (PNG, JPEG, GIF, TIFF, greyscale, etc.) and resolutions.

http://edit.csic.es/geo/mapviewer/edit.html
http://dev.e-taxonomy.eu/trac/wiki/MapRestServiceApi
http://atbi.eu/mercantour-marittime
http://atbi.eu/gemer/
http://dev.e-taxonomy.eu/dataportal/palmae/
http://dev.e-taxonomy.eu/dataportal/cichorieae/

Support is acknowledged from: EDIT

5.2. The Introduction of EDIT's Community Single Sign-On System ∞

Lutz Suhrbier[1], Andreas Kohlbecker[2], Andreas Müller[2]
[1] Freie Universität Berlin, Department of Computer Science, Networked Information Systems (http://www.ag-nbi.de), [2] Freie Universität Berlin, Botanic Garden and Botanical Museum Berlin-Dahlem (BGBM)

The European Distributed Institute of Taxonomy (EDIT) platform, as well as biodiversity providers in general, provides a multitude of web-based taxonomic applications and services. Also, the diversity of service providers reflects the highly distributed, cross-national organisational infrastructure of taxonomic institutions and collections. This results in a problem of identity management. While the service provider's system administrators have to maintain individual access control policies and potentially different user directories for each offered service, users have to enter and remember a variety of login/password combinations in order to access all these different services.

Therefore, EDIT dealt with that problem and was introducing the Community Single Sign-On (CSSO) security infrastructure. To its final extend, CSSO shall protect and provide access to any connected EDIT platform component based on a single identity per user. That way, users need to remember only one login/password combination to use EDIT's platform facilities. On the other hand, service providers keep the sovereign power on their data collections and information infrastructures by defining individual access control policies, but at considerably reduced administrative costs.

Support is acknowledged from: EDIT

5.3. Crossmedia publishing with the CDM ∞

Niels Hoffmann
Botanischer Garten und Botanisches Museum Berlin-Dahlem

One of the main bottlenecks in the taxonomic workflow is the amount of time consumed by the interplay between the taxonomist and the publisher in preparation for going to print. Integrating a system to provide print publishing functionality into the EDIT Platform for Cybertaxonomy, and thus turning the database system into the true "master database" all the way up to the point of publishing, would be a substantial improvement in the workflow. Crucial to the success and acceptance of a publishing system is that the user be given full control over the format and appearance of the printed product.

To comply with these requirements, we are proposing a system that outputs data stored in the EDIT Platform's Common Data Model (CDM) into an OpenDocument format (ODF) document. ODF is an ISO-certified open standard that has been adopted worldwide by numerous organizations. The benefit of this approach is that the ODF document can then be edited using a word processor (e.g. OpenOffice.org), giving full control over the layout of the data to the user. The macro functionality of word processing or desktop publishing programs make it possible to further automate editing of the output, e.g. to delete unwanted parts of the data. Besides using the standard layout template, it will be possible to create specialized templates for individually laid-out publications, for instance to facilitate the process of publishing for a series or journal with specific editorial rules. Using so-called styles, similar to the Cascading Style Sheets (CSS) used in XHTML documents, a layout may be specified for every class of elements, thus changing the layout of the data in a consistent manner.

The proposed system will give taxonomists the opportunity to lay out their data using tools with which they are likely already familiar.

Support is acknowledged from: European Union, EDIT

5.4. Fieldwork today with data acquisition tools

Alexander Kroupa
Museum fuer Naturkunde

Every day probably more than 100.000 biological datasets (observations, collected specimens) are newly generated in the field. Many of these data are still not captured digitally and the majority of these data are not recorded using standard protocols or proper referencing. The goal should be that all datasets are recorded digitally from the outset and that all individual field records are accurately geo-referenced with exact date & time or interval. Therefore the use of electronic field recording tools and software should be promoted also to help minimize error rates, in particular to avoid mistakes right at the beginning of the recording chain. Many errors may be avoided by using authority lists, e.g. for countries, habitat-types or taxa that can already be determined in the field.
Automated geo-referencing and recording of date and time in standardized formats already in the field will also avoid errors when importing or retyping such data into a database. The advantage of using digital field tools is to simplify data recording as well as to improve data quality. Relevant software should be usable for tools such as mobile phones with GPS (Global Positioning System) functionality up to water resistant PDAs - Personal Digital Assistant (e.g. Magellan - Mobile Mapper; Trimble – Juno, Nomad).
For ArcPad software (by ESRI Inc. - Environmental Systems Research Institute) specific applications have already been developed for recording species in the field for different use cases, e.g. one application for birdwatchers monitoring bird sites near Gainesville, Florida, and another application with customized ArcPad forms for inventorying earthworms in Pictured Rocks National Lakeshore, Michigan. Another new tool to record biological occurrence data in the field is the software application "DiversityMobile". developed at the SNSB (Staatliche Naturwissenschaftliche Sammlungen Bayerns) in Munich, Germany for PDAs with GPS functionality.
The example presented here for more efficient electronic data recording in the field is the application for mobile recording with customized forms for ATBI+M (All Taxa Biodiversity Inventories + Monitoring; www.atbi.eu) sites developed within the EC-funded EDIT (European Distributed Institute of Taxonomy; www.e-taxonomy.eu) project. This is a general approach for recording geo-referenced species data using customized forms are for ESRI ArcPad applications. For locality information, two customized forms have been developed to record descriptive data and the geo-referencing of the locality. For each locality, multiple recording events can be created via a form listing all events for one locality. Each event is characterized by a unique EventCode and the start and the end date (& time) of the event. For each

event, a species list of observed or collected individual specimens can be created. The species names can be selected from a taxonomic authority list provided in a file in dBASE-format. Such files can be easily created and exchanged to allow individual researchers to use regional or otherwise customized species lists. Fields and field formats correspond to ABCD standards so that exports of recorded locality, event and species data can be directly integrated into a central database and applications for individual ATBI websites (e.g. www.atbi.eu/mercantour-marittime/ or www.atbi.eu/gemer/).

Support is acknowledged from: European Union, EDIT

5.5. The Taxonomic workflow in the EDIT Platform for Cybertaxonomy

Andreas Kohlbecker, Pepe Ciardelli, Niels Hoffmann, Katja Luther, Andreas Müller
Botanic Garden & Botanical Museum Berlin-Dahlem

The European Distributed Institute of Taxonomy (EDIT) is an EU-funded project designed to help integrate the traditionally disparate field of scientific taxonomy as practiced in Europe.
The EDIT Platform for Cyberaxonomy, henceforth called “the Platform”, brings the taxonomic workflow to the Internet, providing an open architecture to connect and integrate existing applications and developing new tools only where necessary.
The Platform provides taxonomists but also life science in general with a set of loosely coupled tools to facilitate fieldwork, analyze data, assemble treatments, and to publish efficiently. Reliability and reusability of data are key requirements for each of these tools and thus for the Platform as a whole.

In order to guarantee reusability of data and to facilitate full interoperability between the various Platform components, a new data model, the “Common Data Model” (CDM) was developed. The CDM is strongly influenced by both the TDWG Ontology (http://wiki.tdwg.org/twiki/bin/view/TAG/TDWGOntology) and the Berlin Model (http://www.bgbm.org/BioDivInf/Docs/bgbm-model/). Other models and standards have influenced the modelling as well. On top of the CDM, the CDM Library - a Java library - has been built. It offers a local API, web services and transformation services for all major taxonomic standards from and to the CDM, making the CDM the “glue” between Platform components.

The workflow the Platform seeks to optimize is in essence the “revisionary” process by which an existing classification of a group of organisms is revised, and by which previously unclassified organisms are assigned to a “taxon”. The workflow begins either in the field with the collection and transport of specimens, or with a review of existing literature and specimens. Final results of the entire process are the publication of revisionary treatments, publication of new taxa, and preparation of new specimens to be stored and curated in natural history collections.

Bottlenecks and areas for improvement within the taxonomic workflow were identified early and made a focus of the design process for the Platform and its components. Time consuming and error prone data entry is improved by the EDIT Desktop Taxonomic Editor, with its modern approach to the daily taxonomic work process and use of techniques such as drag&drop, on-the-fly parsing, and passive warning instead of intrusive alerting.
Bottlenecks related to publication are avoided by the EDIT publication tools. The individually configurable EDIT data portal is the preferred tool for publishing data hosted in a CDM (Common Data Model) Store via the web. Ready-to-use distribution and occurrence maps are created and provided by the EDIT MapServices. An integrated print publishing service will accelerate and improve the creation of professional, printed publications directly from the CDM Store, putting publication into the hands of the practicing taxonomist.

Locating specimens often is time consuming. The EDIT Specimen Explorer helps find taxonomically relevant specimen and observation data by searching on the GBIF (Global Biodiversity Information Facility) index by using checklist-based thesauruses to receive more complete results.

A special fieldwork data base and field tools like water-resistant GPS/GIS handhelds with integrated camera for efficient data acquisition in the field complement the set of tools which together enable the EDIT Platform for Cybertaxonomy to fully support the taxonomic workflow .

5.6. e-Taxonomy with CATE and EDIT

Benjamin Richard Clark
Royal Botanic Gardens, Kew

Creating a Taxonomic eScience (CATE) is a project conducted by a consortium of scientists from the Royal Botanic Gardens, Kew, the Natural History Museum, London, and the University of Oxford. The project has explored how

groups of taxonomic experts can use the web to deliver a dynamic, consensus classification for a particular taxonomic group via the web. Two families were chosen as the focus of the exemplar groups, one zoological (Sphingidae, the hawkmoths), the other botanical (Araceae, arum lilies & their relatives). Each group of taxonomists has published a "web-revision"; resources that serve the same function as a paper monograph or flora, taking advantage of the web to enrich the data in the web-revision and provide tools to enable users to navigate the content in a variety of ways.

Web-revisions, as envisaged by CATE, are dynamic resources, continually being improved online by a community of taxonomists as new names are published and new data becomes available. In order to serve as a useful checklist, each web-revision preserves changes to the data, allowing the prior state of the web revision to be viewed and changes between versions to be tracked.

CATE is built using the European Distributed Institute of Taxonomy "Common Data Model" (CDM) and CDM Java Library, a low level software library for interacting with the CDM that is the foundation of several tools and services produced by the EDIT project. Most of the data covered by the CDM can be published and edited using the CATE application. CATE uses much of the core functionality provided by the CDM Java Library, such as versioning, audit, persistence and search as its foundation. In addition, CATE uses components from the CDM Community Server application, a web application that provides web services exposing CDM data to software clients and allowing data published by a CATE site to be integrated with other biodiversity data. CATE sites also use a number of other services and tools, such as the EDIT Map REST (Representational state transfer) Service, the Lucid and Xper2 Interactive Key players, and EDIT Scratchpad sites.

This presentation will introduce the CATE project and explain how the CATE application uses the CDM and the generic services exposed by the CDM Java Library to create a specific application for a particular purpose. It will show, through examples, how the CDM library has enabled the development of the CATE web application and how communities of taxonomists have used the CATE exemplar sites to publish and curate taxonomic information. Developing a generic software library to support multiple applications is a challenge, and this presentation will explain how the design of the CDM Java Library has been influenced by the requirements of the various applications that it supports.

The CATE exemplar web revisions can be found at http://www.cate-araceae.org and http://www.cate-sphingidae.org . The project software can be downloaded from http://www.cate-project.org . Information about the EDIT Common Data Model can be found at http://dev.e-taxonomy.eu.

Support is acknowledged from: Natural Environment Research Council (Grants NE/C001532, NE/C51588X/2 and NE/C515871/1), European Distributed Institute of Taxonomy

5.7. Using the CDM to build Europe's largest species database

Marc Geoffroy, Anton Güntsch, Andreas Kohlbecker
BGBM

PESI (a Pan-European Species-directories Infrastructure) defines and coordinates strategies to enhance the quality and reliability of European biodiversity information. It is a joint initiative of two Networks of Excellence: EDIT (European Distributed Institute of Taxonomy) and MarBEF (Marine Biodiversity and Ecosystem Functioning); funded by the European Union under the Framework 7 Capacities Work Programme: Research Infrastructures and led by the University of Amsterdam. It started in May 2008, will last three years and involves 40 partner organisations from 26 countries.

One of the goals of PESI is to taxonomically integrate and secure the main pan-European species checklists, starting with Fauna Europaea (FaEu), the Euro+Med plantbase (E+M), and the European Register of Marine Species (ERMS). With more than 200,000 species, together they provide the largest and most comprehensive regional species inventory in the world.

The integration of these three checklists, currently stored in separate databases with their different data models, relies on the Common Data Model (CDM) which was developed within EDIT with a goal of ensuring it could be mapped to most of taxonomic databases. The CDM essentially follows the TDWG Ontology (http://wiki.tdwg.org/twiki/bin/view/TAG/TDWGOntology), but modelling was influenced by other models and standards, such as the Access to Biological Collections Data (ABCD) schema, the Taxonomic Concept Schema (TCS) and the Structure of Descriptive Data (SDD) schema, as well.

The CDM Java library implements all classes in the CDM, and is the primary interface for applications communicating with CDM data stores. Import routines will be created as necessary for each of the PESI source databases; all source data

will then be merged into a single PESI CDM store instance. If one or more of these databases is maintained externally, the import routine must be run regularly. Rules for data quality control concerning the syntax of terms and the structural and relational integrity of data will be implemented at the CDM level and applied to the complete data set of the PESI CDM store. Overlapping and inconsistent data stemming from disparities among the source databases - for instance, a handful of animal species with brackish water habitat are stored in both ERMS and FaEu - will also be detected. These conflicts and discrepancies can then be resolved by PESI taxonomists using the EDIT Taxonomic Editor, which will play an important role for the maintenance of participating checklists and therefore complement the quality checker rules.

PESI data will be regularly exported from the CDM store into a denormalised relational database management system (the "PESI data warehouse"), following maintenance and improvements to data quality. Here the CDM's versioning capability will also provide substantial support. The "PESI data warehouse" is optimized for queries from the World Wide Web portal and PESI web-services. The new PESI portal will make the content of the major European taxonomic infrastructures available and support the use of the pan-European species data in the e-science domain.

Using the CDM as the decisive layer for an ambitious project such as PESI represents a major step towards establishing the CDM as a possible standard for taxonomic databases and applications.

PESI (http://www.eu-nomen.eu/pesi)
CDM (http://dev.e-taxonomy.eu/trac/wiki/CommonDataModel)
EDIT (http://www.e-taxonomy.eu/)
ERMS (http://www.marbef.org/data/erms.php)
Euro+Med plantbase (http://www.emplantbase.org/home.html)
FaEu (http://www.faunaeur.org/)
MarBEF (http://www.marbef.org/)

5.8. The EDIT Taxonomic Editor - More Features, Less Code

Pepe Ciardelli
Botanic Garden Botanical Museum Berlin-Dahlem

In these days of slavish devotion to all things web, it may strike some as odd that the flagship product of the EDIT (European Distributed Institute of Taxonomy) Platform for Cybertaxonomy is a desktop application. The decision to go desktop was made fairly early in the project, nearly two years ago. Has experience proven this to be the right decision? And what lessons can other TDWG member institutions draw from this experience?

The Taxonomic Editor was implemented using the Eclipse Rich Client Platform (RCP). Eclipse is an integrated development environment (IDE) much loved by Java developers for its myriad user interface shortcuts and gestures, which accelerate the programming process substantially. The RCP makes the nuts and bolts of Eclipse generic, and offers a set of Java libraries that can be used as the building blocks for any application that involves the editing of resources.

A few lessons learned from using the Eclipse RCP:

Start with the strengths of the library and design the software based on these strengths. We often started a given development task by trying to bend the Eclipse Platform to our ideas about the best way to treat the taxonomic workflow in a user interface. However, we tended to come around in the end to a more Eclipse way of doing things, as we repeatedly found that their solutions to common design problems were more sophisticated than ours.

Use the library as an instructional tool. It is therefore helpful to consider Eclipse as an implementation of best practices in interface design. Eclipse is also a way to introduce developers to object-oriented programming best practices, design patterns, etc. This is especially relevant in our field, which attracts people who start in the natural sciences then teach themselves programming, often without any formal study in the information sciences.

Find ways to teach the TDWG community how to use the new technology. The crucial question is whether Eclipse development – and desktop development in general – is beyond the resources of most institutions in our community. A great deal of time was spent simply learning the ropes of the Eclipse Platform, which like many open-source technologies is notoriously under documented. If the TDWG community is to adopt more sophisticated technologies for software development, person-to-person knowledge transfer among institutions will be crucial to overcoming this early, largely trial-and-error phase. The EDIT developer workshops taking place as part of this year's TDWG meeting are a step in the right direction.

We are only now beginning to leverage the full power of the Eclipse Platform, and we hope upcoming releases of the Editor offer a truly pleasurable experience, one in which the application practically disappears and the user is fully engaged in the taxonomic workflow.

5.9. The EDIT Platform for Cybertaxonomy - User Workshop

Pepe Ciardelli[1], Niels Hoffmann[1], Andreas Kohlbecker[1], Alexander Kroupa[2]
[1] Botanic Garden Botanical Museum Berlin-Dahlem, [2] Museum für Naturkunde Berlin

The European Distributed Institute of Taxonomy (EDIT) is an EU-funded project designed to help integrate the traditionally disparate field of scientific taxonomy as practiced in Europe.
The EDIT Platform for Cybertaxonomy, henceforth called "the Platform", brings the taxonomic workflow to the Internet, providing an open architecture to connect and integrate existing applications and developing new tools only where gaps in the workflow exist.
The Platform provides not only taxonomists but also life science in general with a set of loosely coupled tools to facilitate fieldwork, analyze data, assemble treatments, and publish efficiently. Reliability and reusability of data are key requirements for each of these tools and thus for the Platform as a whole.

The purpose of this workshop is to introduce potential users to various components which make up the Platform. The focus will be on components which require installation within the user's institution or on the user's desktop, although the role of web services within the Platform will also be addressed.

Participants will be introduced to the Common Data Model (CDM), the backbone for those components which require that the user's data be stored. The philosophy behind the CDM will be explained, and participants will learn how to get their data in and out of the CDM using native import / export functionality. The software tools implementing this functionality will be demonstrated.

Specific components which will be the subject of the workshop include:

ATBI Fieldwork Tools – Participants will learn about EDIT's efforts to fill gaps in ensuring efficient data acquisition in the field. Tools to be demonstrated include standardized forms for data acquisition, a special fieldwork database and field tools such as water-resistant GPS/GIS handhelds with integrated camera.

The EDIT Desktop Taxonomic Editor – Participants will learn how to install the Taxonomic Editor on their desktops, and how to move seamlessly from working on a single-user, locally installed database to working on a remote database together with a community of users. The precise role of the Editor in the taxonomic workflow will be explained: which types of data are best edited here, which types are better left for other, established tools, and how can data from these tools then be integrated into the CDM?

The EDIT Data Portal – Combining a rich feature set culled from the community's experience with taxonomic web portals with the ease of configuration offered by its Drupal backend, the EDIT Data Portal displays data in a CDM data store. Participants will learn what is required to get an EDIT Data Portal up and running in their community, and how to configure CDM data to prepare it for this pre-publication state.

6. Agriculture including Biocuration, Disease

6.1. Semantic Web for Ecosystem Approach applied to Fisheries

Julien Barde, Pascal Cauquil, Abdel Dkhissi
IRD

Ecosystem Approach to Fisheries (EAF) is an application of sustainable development that aims to improve the management of (over)-exploited marine resources by taking into account the numerous relationships between marine ecosystems' components (like food webs). It is very important in the context of EAF to share existing informational resources (IR) among stakeholders to set up effective management schemes. However these IR are distributed in many organizations and thus difficult to share. Moreover, they have been, so far, managed with heterogeneous formats in systems that are difficult to access, even at the scale of a single organization like ours. Standardization is thus required.

The Mediterranean and Tropical Halieutic Research Center (CRHMT), in Sète, France, aims to improve and enhance EAF by sharing its IR more effectively. For years, CRHMT's work has focused on just a few top predators exploited by fisheries in a few marine ecosystems. This is just a small part of the whole set of IR needed for an EAF at a global or even regional scale. Therefore, a new IR management and sharing application will enhance our ability to collect the

necessary resources and contribute to EAF worldwide.

By setting up a new information system (Ecoscope), built on top of existing ones, the CRHMT aims to inventory, manage and share the various IR acquired during previous, ongoing and coming projects. This system aims to comply with current best practices in terms of standardization so as to become a node of a global network which facilitates the exchange of these IR among institute researchers, with our collaborative partners and to the wider public.

The architecture we implemented takes into account some of the recommendations of the World Wide Web Consortium Semantic Web activity to facilitate knowledge and metadata sharing, as well as recommendations of the Open Geospatial Consortium and TDWG that are specifically relevant to reach our goals. However, re-engineering our previous datasets handled in heterogeneous ways to serve them properly with standard data formats and related protocols is a challenging task. Indeed enabling syntactic and semantic interoperability requires a lot of work and attention to detail. Moreover, as our approach is generic and could be implemented in similar ecosystem management cases, we aim to share, or duplicate, this application in other contexts.

We set up an ontology schema for our domain by reusing specific parts of various existing schemas. For example, we aim to match with i) SPIRE ontologies for ecological concepts, with ii) Friend Of A Friend (FOAF) RDF (Resource Description Framework) schema for social networks concepts, with iii) Simple Knowledge Organization System (SKOS) RDF schema for controlled vocabularies concepts, and with iv) Dublin Core Metadata Initiative RDF schema for a simple metadata element set.

This presentation will focus on our ongoing work on food webs in marine ecosystems, the resulting knowledge base conceptual model (a UML (Universal Modelling Language) class diagram) and technical aspects of our current implementation (using the Web Ontology Language and Jena APIs (Application Programming Interfaces)). We will then give some examples of a Graphical User Interface we set up to satisfy different use cases. The need to aggregate raw data on fact sheets is addressed with geographic and network representations (e.g., foodwebs). Systems interoperability is assured by delivering part of our IR through web services like the Web Catalog Service (CSW) and the Web Map Service (WMS).

Support is acknowledged from: IRD, Research Institute for Development

6.2. Integrating Animal Health and Biodiversity Informatics Standards

James Case
University of California -- Davis

The value of sharing biological, biodiversity or health related data is well-recognized. The value is especially keen when addressing a natural or introduced environmental or health related emergency. The need for real time access to high quality, validated data from a wide variety of data suppliers is key to the effective detection and response to large scale events that threaten human, animal and environmental well-being. The heterogeneity of local data has typically been an obstacle to routine data interoperability. This report discusses the approach of the National Animal Health Laboratory Network (NAHLN) to achieve near-real time data interoperability of animal health data through the implementation of a set of selected messaging, terminology and identifier standards.
While most veterinary diagnostic laboratories employ sophisticated Laboratory Information Management Systems (LIMS) in their operations, most are highly customized to the local needs of their clients and reporting agencies. The resultant heterogeneity in data structures and terminology is an impediment to the rapid collation and aggregation of data that are needed for both ongoing disease surveillance activities and response to animal health emergencies. In many cases integration of related data such as biodiversity (both flora and fauna), meteorological, water and air quality, and transportation are important considerations in the appropriate response to animal health events both in commercially produced animals and wildlife. The impact on public health is also an important consideration.
To address the initial need for integration of health related data, the (NAHLN) evaluated and selected a set of health data and identifier standards, many from the human health domain as the basis for exchanging data both between laboratories and a central data repository. The selected standards; Health Level Seven (HL7), Logical Observation Identifiers Names and Codes (LOINC), Systematized Nomenclature for Medicine (SNOMED-CT), National Animal Identifier System (NAIS) and ISO object identifiers have proven to be robust, extensible and capable of meeting the interoperability needs of the NAHLN. These standards were extended by the NAHLN to meet the specific needs of animal health and provide the foundation for increased interoperability for other health related initiatives such as the Integrated Consortium of Laboratory Networks (http://www.icln.org/), which includes human, animal, environmental and military oriented networks providing surveillance activities.
Terminology management has proven to be of primary importance in the effectiveness of the NAHLN, due to the requirement to eliminate post-processing of aggregated data. The NAHLN adheres to good vocabulary practices that

allows for extension, retirement and traceability of terms over time, as well as the partitioning of terminology subsets based on the underlying data model used to create the messaging platform. The application of this approach to biodiversity informatics may provide a readily available mechanism to achieve interoperability among biodiversity databases.

6.3. Catalogue of Life Phase 2 and the new 4D4Life Project

Richard J White[1], Frank A Bisby[2], Meirion Jones[3], Wouter Addink[4]
[1] Cardiff University, [2] University of Reading, [3] Botanic Gardens Conservation International, [4] ETI BioInformatics, University of Amsterdam

The Catalogue of Life (CoL) partnership recently announced plans for its Phase 2 programme from 2009 – 2014. These include the ambitious EU-funded project "Distributed Dynamic Diversity Databases for Life" ("4D4Life") to establish a state-of-the-art e-infrastructure with an array of new electronic services.

The 4D4Life project is presently consulting with various user groups concerning the services wanted, before drawing up a priority listing for March 2010. At TDWG we seek input from system developers as to which web services and application programming interfaces (APIs) they would wish for enhanced connectivity to the Catalogue of Life information.

At present, users of the CoL can search the database to find data about the taxon for which they have a name, check the spelling of scientific names, or find related taxa. They can use "synonymic indexing" both ways – to locate a valid species from a name now placed in synonymy, or to see all synonyms listed for a particular taxon. Users can also make copies of the database for local use or for providing their own new services and products, as the Global Biodiversity Information Facility (GBIF), the Encyclopedia of Life (EoL) and others are doing.

One issue for discussion is the way the CoL taxon Life Science Identifiers (LSIDs) are allocated, which will support services for users who have different definitions of what they mean by a change in a taxon. We plan to support services and automated alerts which can inform users of name changes, concept changes (which may affect the reliability of sharing data for what is apparently the same taxon), and changes in other data which the CoL holds about a taxon, including common names, classification and distribution.

New services under consideration include:
(i) synchronising users' data sets with the Catalogue of Life to keep them up-to-date with new names, new taxa, and changes in taxa and classification, or to use the CoL as a backbone for their annotations and extensions;
(ii) submitting batches of names not presently in the CoL for scrutiny and inclusion by the panels of taxonomic experts working with the provider databases;
(iii) incremental cross-maps between the annual editions;
(iv) creating products and services which automatically gather CoL data and perhaps their own data and present it in new ways;
(v) a range of distribution, query and download services using standard protocols and data models; and
(vi) special services to support novel applications on hand-held devices.

The issue for discussion at this meeting is how to prioritise the development of these and additional services which may be requested by system developers or required for linkage to other initiatives such as EoL, GBIF and the Global Names Architecture.

Support is acknowledged from: EC Framework 7

7. Identifying Biodiversity for Food and Agriculture

7.1. Impact of Citizen Science projects on biodiversity policies :Tela Botanica conference results ∞

VIOLETTE ROCHE

Scientist projects, including citizens are largely developed particularly in ecology and the environmental sciences since the 20th centuries. Thousand of volunteers collect data for scientist projects that have been specifically designed or adapted to give amateurs a role, either for the educational benefit of the volunteers themselves or for the benefit of the project.

Projects on climate change, invasive species conservation biology, ecological restoration, population ecology or monitoring of all kinds require citizen science as they need to collect large volumes of field data over a wide geographical area.

On of the consequences of this recent explosion of citizen's sciences is the largescale of projects developed in biodiversity with different protocols, networks, technologies or ethics.

This is because citizen's sciences projects have never been confronted
in France that we organized the first meeting of "Citizen's sciences &
biodiversity" this last 22 and 23 of October (colloquescb.tela-botanica.org).

7.2. Plant diversity, functional traits and ecoinformatics ∞

Eric Garnier
CEFE

Understanding the dynamics of biodiversity requires the integration of organism responses to changes in their environment at the level of populations, communities, ecosystems and landscapes. This can be assessed through a functional characterization of organisms using their "traits", defined as morpho-physio-phenological features of organisms which impact fitness indirectly via their effects on growth, reproduction and survival. A better understanding and prediction of the response of vegetation diversity to environmental drivers (including land-use and climate changes) on the one hand, and the effects of these changes on ecosystem properties on the other hand, can be achieved by combining a functional (through traits) and a taxonomical (through relevés) approach to biodiversity.

The potential of this approach will be shown in a case study, and how it could be generalized by coupling the thousands of computerized vegetation relevés (giving species abundances in local communities) with existing trait databases (giving species functional characteristics), environmental data and ecosystem properties will be discussed. A prerequisite is that we succeed in integrating the fragmented data sources and make them available to address these questions. A first step towards data integration through the development of a trait-based ontology for plant ecology will be presented.

Support is acknowledged from: CNRS

7.3. Data integration and its impact on genebank management

Christopher Richards, Gayle Volk
USDA ARS

The management of genebanks has become more complex in the last decade. Over the last few decades, molecular geneticists have begun to understand the genetic underpinning of complex traits that are critical for agriculture. There has been a universal recognition that continued progress is dependent on making use of the natural variation contained within the world's genebanks. The primary mission of providing validated, living materials for research is ongoing; but the use of natural diversity for gene discovery programs is giving rise to a different stakeholder group with additional expectations of collections. In many ways the data contained within these collections is as important as the materials they describe. Developing ways of integrating these data with other data sets including geospatial, ecological and genetic data requires implementation of data standardization. These interests can easily be aligned with ongoing TDWG efforts. We present several case studies of how data standards are critical to data integration, and how successful integration provides key services for gene bank management.

Support is acknowledged from: USDA. ARS National Center for Genetic resources Preservation

7.4. Semantic Standards for Genomic Analyses of the South and Mediterranean Plants: the Generation Challenge Program Use Case

Manuel Ruiz
CIRAD

The Generation Challenge Programme (GCP) platform was developed to meet the challenges of data acquisition, computational resources, and software interoperability and integration across a globally distributed consortium of partners. This platform includes: (i) shared, public platform-independent domain models, ontology and data formats (ii) web service and registry technologies (iii) platform-specific middleware implementations of the domain model integrating a suite of public databases and software tools into a workbench to facilitate biodiversity analysis, including the comparative analysis of crop genomic data.

The cornerstone of the GCP platform is the development of common standards for GCP data. Major concepts in the domain of crop research - for example, concepts like "germplasm" and "genotype" - can be expressed in terms of a general blue print of such concepts-as-entities and of their relationships to one another, within a so-called domain model. The GCP development team has specified such a domain model to drive development of a "model driven architecture" within which tools and data sources may be efficiently connected to one another.

The GCP domain model is not a complete embodiment of semantics in GCP software systems since specifying such a complete model would not be practically feasible. Therefore, in addition to the GCP domain model, the GCP development team has also specified a formal framework to manage a GCP ontology that complements the semantics of the domain model.

Several applications and integrating tools have been developed, such as a GCP web query and display application ("Zeus"), a GCP Ontology browser, and the stand-alone molecular breeding components MBDT (Molecular Breeding Design Tool) and MOSEL (Molecular Selection Tool). Our team was particularly involved in the development of GenDiversity, a query and analysis Web application combining genotyping data from diverse data sources, developed in support of diversity studies. Furthermore, GCP components can also be used by non GCP projects. Indeed, we present Orylink, a personalized integrated system for rice functional genomic analysis.

The infrastructure of the platform is complex, and it still may discourage developers from using it. Therefore, we need to establish better training and documentation for users of the platform.

Support is acknowledged from: GCP

7.5. Sketches from an anthropologist's fieldnotes on local knowledge about agrobiodiversity: a primer for biodiversity information specialists

Eric Garine
Université Paris Ouest

The hope that indigenous local ecological knowledge (LEK) could contribute to biodiversity conservation and sustainable development has spurred proposals that this knowledge be stored in international databases, where it can be diffused, shared and used. These proposals have been controversial, for reasons often poorly understood by those interested in biodiversity more generally. I will outline some of the empirical, methodological and ethical dilemmas associated with LEK databases, illustrating my points with results from our own multidisciplinary studies of LEK about agrobiodiversity. LEK is dynamic, not static. Its conservation, like that of biodiversity itself, cannot be ensured by its ex situ storage in archives, but only by preservation of the biological and cultural processes responsible for its maintenance and its continuing evolution. LEK is not easily divisible into transposable bits of information, but forms part of local knowledge systems. Removed from its environmental, social and cultural context, LEK may lose its meaning, and its promised contributions to conservation and development may prove illusory. Finally, storage in international archives could undermine the control that local peoples exercise over their knowledge.

7.6. New challenges for visual information retrieval in biodiversity applications

Raffi Enficiaud, Nozha Boujemaa
INRIA Paris-Rocquencourt

Species identification is a key step for most projects in biodiversity conservation and sustainable development. In this context, automated methods are often crucial in order to overcome some of the many difficulties inherent to the identification process. These primarily include the time required to make an identification, and the level of domain expertise needed to do this correctly.

Multimedia search engine technologies based on "Content-Based Image Retrieval" (CBIR) mechanisms allow envisaging a fully automated, scalable and wired method for taxonomic identification. In this context, the image retrieval system IKONA, developed in the IMEDIA project at INRIA (http://www-roc.inria.fr/imedia/), has been successfully applied for identification of orchid species. The system has been shown capable of consistently identifying (95% identification rate) 40 species from standardized sets of images derived from leaf scans, with a signature computed on the overall visual appearance of a given image. The system describes shape, colour and texture features, providing a visual appearance modelling.

If these techniques are to be used on a wider range of species, several difficulties that directly affect the overall efficiency and robustness of the system response must first be solved. First, the digital representation of the species

images in signatures should be able to cover all the visual cues with which the biodiversity experts are already used to working during the identification process. This represents the signature property of fidelity to visual content; through this property we can measure the accuracy of a search engine. To capture all this expertise, the amount of information and operations should be increased. Geometrical relation modelling on the retrieved elements is one promising prospective research path to accomplish this.

Ideally, such a system should distinguish between relevant and irrelevant information. How can human and application-specific expertise be transferred to the system in an efficient manner? Current research explores the statistical learning of the notion of "relevance", including modelling users' feedback into the system.

As increasing numbers of people become involved in the building of taxonomic databases, the knowledge they create becomes, by nature, highly heterogeneous. Images are no exception to this rule: shots of the same plant usually encompass very different views (leaves, flowers, entire trees, etc.). A CBIR system should also be able to recognize different views of the same plant, and hence be able to use all the pieces of knowledge.

The growing amount of data requires scalable approaches in handling both indexing and searching stages. This is why great effort is devoted to structuring feature spaces as well as to optimized signature-computing methods. Finally, intuitive navigation and visualizing of the databases should be proposed to the users to facilitate exploitation of the available information.

In conclusion, specification work done through the Pl@ntNet project has identified several promising research directions, including active multi-class learning, geometric consistency, and scalable search and structuring. Our presentation addresses these research topics within the Pl@ntNet framework.

7.7. The concept of "Networked collection" or "Virtual collection": revisiting the classical delineation between "in situ" and "ex situ" conservation and its consequences on database management

Roland Bourdeix[1], S F Weise[2], S Planes[3], L Guarino[4], T Bambridge[3], C Lusty[4]
[1] French Agricultural Research Centre for International Development, [2] Bioversity International, [3] Insular Research Center and Environment Observatory, [4] Global Crop Diversity Trust, Rome

A networked collection, also called a virtual collection, is located at more than one geographical/institutional site, spans the genetic diversity of a given species (genepool) and gathers stakeholders having a mutual interest in rationally conserving and exchanging germplasm. In the extreme application of this concept, several accessions could be conserved, each at a distinct site. Many intermediate strategies are also conceivable.

The global coconut conservation strategy (GCCS) was developed by the International Coconut Genetic Resources Network (COGENT) and the Global Crop Diversity Trust. This strategy is mainly based on ex situ conservation in five large regional field gene banks. The implementation of a networked collection could allow this system to involve more countries, sites and stakeholders.

In order to make the germplasm affordable to stakeholders, the Polymotu concept was integrated as a new approach in the GCCS. Several accessions of coconut palms will be planted, each in a distinct isolated site, such as islets near inhabited islands, isolated valleys, or large plantations of other tree crops. This geographical remoteness will ensure the reproductive isolation needed for true-to-type breeding of the crop varieties through natural and cheap open pollination.

A challenge being faced is that of gathering (in the same network and database) accessions held in international genebanks, as well as accessions conserved on islets owned by municipalities, islanders' families or tourism enterprises.

Between 1992 and 2003, in a step-by-step manner, a database called CGRD (Coconut Genetic Resources Database) has been developed to manage and describe the accessions conserved in the ex situ coconut field genebank. This database system will have to be updated in order to integrate further geographical, social and ethnological information. Data will include not only Bioversity standard descriptors, but also additional information regarding places where the germplasm is conserved, information about the owners of these places, and rules that regulate access to the germplasm.

The responsibility of funding such a networked/virtual collection could be shared by participants (who could provide part of the infrastructural costs) and by donors (through the funding of specific activities focussed on priority unique accessions). In order to improve the quality of conservation, funding could be allocated on an accession basis, according to evaluations conducted by the COGENT network. Criteria for the selection of an accession for conservation in the

virtual/networked collection include: the ability to reproduce true-to-type, genetic representativeness, uniqueness of the germplasm, and policy considerations. Database management will be essential for conducting such evaluations.

8. Accessing information on Agricultural genetic resources and crop wild relatives

8.1. ACCESSING INFORMATION ON PLANT GENETIC RESOURCES

Elizabeth Arnaud, Sónia Dias
Bioversity International

SINGER and EURISCO on-line catalogues provide inventories of worldwide conserved agricultural diversity with a primary access for the identification and localization of crop samples (= accessions). Both catalogues apply the standard developed for plant genetic resources (PGR), the Multi Crop Passport Data format, and the networks are now working at integrating the phenotypic data, in the framework of the new project on a Global Information Portal for PGR. Linking these two inventories with a wider range of inventories (e.g.: botanical gardens, herbarium specimens) is of great interest for the community working on Plant genetic resources. The Global Biodiversity Information Facility (GBIF) harvests the Passport data from these two catalogues on which BioCASE and Tapir protocols have been applied.

EURISCO - THE EUROPEAN PLANT GENETIC RESOURCES SEARCH CATALOGUE
http://eurisco.ecpgr.org/EURISCO
EURISCO is designed to serve as the European PGR information hub, with the European ex situ Inventories National Inventories (NIs)as its backbone, providing access to passport data; using international standards for information access and exchange, enabling users to search and access information on crops, forages, wild species, landraces and breeding lines for all crops. It assists policy makers with meeting their countries' commitments, national, regional and international obligations; especially those related to the FAO Global Plan of Action (GPA), the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA), contributing to the Multilateral System (MLS) and the Standard Material Transfer Agreement (SMTA) of the Treaty and the Convention on Biological Diversity (CBD), regarding the conservation and utilization of PGR materials. EURISCO is based on a European network of 42 National Focal points that makes European crop biodiversity data available worldwide, linking users to information on over 1 million accessions from 38 European countries, representing 5,394 genera and 34,469 taxa (including synonyms and spelling variants). These samples of crop diversity represent more than half of the ex situ accessions maintained in Europe and roughly 18% of total worldwide holdings. EURISCO is hosted at and maintained by Bioversity International on behalf of the Secretariat of the European Cooperative Programme for Plant Genetic Resources (ECPGR)

SINGER - THE SYSTEM-WIDE INFORMATION NETWORK FOR GENETIC RESOURCES
http://www.singer.cgiar.org/
This catalogue is the product of the germplasm information exchange network of the Consultative group on International Agricultural Research (CGIAR). These centers placed their collections under the inter-governmental authority of the Food and Agriculture Organization of the United Nations (FAO). This means that the conserved diversity is held and exchanged in the context of the Multilateral System (MLS) in accordance with the Treaty. The samples held by the CGIAR are conserved in the public domain and are available for distribution. The CGIAR centers also foster research and policy development to bring the benefits of agricultural diversity to subsistence farmers. SINGER provides a single entry point to the inventories of the 11 CGIAR genebanks and the Asian Vegetable Research and Development Centre (AVRDC). Information on over 696,500 accessions from 77 collections is available. SINGER makes information about the diversity of plants available to all. An online sample ordering gateway has been added enabling users to request samples from several CGIAR genebanks at the same time.

Support is acknowledged from: CGIAR System Wide Genetic Resources Programme, Global Public Goods Programme (World Bank), European Cooperative Programme for Plant Genetic Resources (ECPGR)

8.2. Integrating the monitoring of agricultural pests into biodiversity assessments

Gail E Kampmeier
University of Illinois at Urbana-Champaign

Traditional biodiversity assessments have ignored the contributions of agricultural studies, despite their rich heritage of purposefully constructed hypotheses tested in a variety of habitats with replicated experimental designs. That the data collected in this manner are often a combination of many observations chained by vouchered specimens deposited in museums for each identifier, makes these data different from those traditionally reported to biodiversity assessments (such as the Global Biodiversity Information Facility, GBIF) from most museum collections.

Conservation managers indicate that long term, purpose-driven data are more valuable to them in making their decisions than one-off records of species that are often documented in museum collections. Data from agricultural experiments and long-term monitoring of particular pest species can help fulfill this need, even though agroecosystems are usually highly structured and artificially manipulated. Agricultural experiments often need to be repeated to gain the necessary degree of confidence to make predictions about the behavior and presence of a specific set of organisms under changing environmental conditions. Agricultural scientists conducting field research realize their trap data usually vary, often significantly, from year to year. The reasons for these variations provide a rich backdrop of different abiotic conditions that may cascade into different planting or cultivation dates, a different ratio of the target fauna appearing and disappearing from trap samples, even different proportions appearing in one sampling device vs. another. All of these experiences may be integral to building an understanding of how a pest and its natural enemies behave in the agroecosystem, giving clues to their management. Plotting the phenology over time of target species in their environment may be indicators of the effects of changing climate and modifications to ecosystems.

So the challenge to collectors of biodiversity information lies in knowing how to document and preserve the raw data whose purpose and value are often discarded after the experiment is analyzed and the manuscript published. With the advent of the Darwin Core Standard, we can now more seriously examine how agricultural datasets might be integrated and used to augment our biodiversity heritage.

Support is acknowledged from: Illinois Natural History Survey, University of Illinois

8.3. Can Biodiversity Informatics (help to) save the AQUATIC genetic resources? ∞

Nicolas Bailly[1], Roger S. V. Pullin[2]
[1] WorldFish Center, [2] Consultant

The gathering of information on live stocks and crops at global level started several decades ago, which was subject to many research programmes, registers then databases, with more institutions involved. Strangely enough, fishes, and marine organisms in general have barely been considered as "animals" and "plants" by these programmes, although more than 150 species are used in aquaculture, and certainly much more if one adds the aquarium trade species. Ten years of efforts to push for this recognition and launch dedicated programmes did not generate actual actions by international genetic resources organisations, although several plans and resolutions were published.

For one decade now, it is evidenced and demonstrated that commercial fisheries on wild population are not sustainable in the current world context: search for profit, population increase, globalized market structure, hypertechnicity of boats and gears, etc. Drastic measures and mentality changes should be put in place very quickly in order to save many stocks, if not the aquatic environment itself down to 2,000 m depth.

A possible alternative is aquaculture, although with some restrictions to make it at the same time sustainable, "green", and profitable for poor populations. But to be able to develop and diversify commodities, wild stocks and populations must remain available as genetic resources reservoirs, as it was demonstrated with the creation of the GIFT tilapia. And information must be gathered at population level in the wild, not only at ecosystem and species levels.

Fortunately, FishBase and SeaLifeBase, Biodiversity Information Systems on all fishes and all marine organisms respectively, were developed to accommodate information at that level.

Unfortunately, no programme was funded to populate them at that level (about 70 stocks and aquaculture strains in FishBase over 31,100 fish species in the world!).

We firmly believe that Biodiversity Informatics should help to increase the visibility and concerns about aquatic populations and their inherent genetic resources; by consequence, it may help to save genetic diversity necessary to develop new commodities, and hopefully, release the pressure from the wild catches. Saving the present to insure a future. To avoid reinventing the wheel. information and informatics standards developed for terrestrial genetic resources will be adapted to aquatic species.

8.4. A global information portal to facilitate access and use of information on genebank accessions

Michael C. Mackay
Bioversity International

Plant genetic resources of food and agriculture (PGRFA) conserved in about 1500 ex situ genebanks are a critical resource of biodiversity, not only for general conservation and agricultural production, but also to provide the genetic variation required to address such global issues as food security and adaptation to climate change. The availability of passport and other information is a key to the effective and efficient utilization of these PGRFA in plant improvement and therefore meeting humanity's needs. There are many dispersed information systems for ex situ genebanks providing information about holdings. Most systems utilize accepted standards for passport data, but including other types of data in these systems, particularly phenotypic data, presents numerous challenges for information standards. A partnership between the Global Crop Diversity Trust (GCDT), the Secretariat of the International Treaty for Plant Genetic Resources for Food and Agriculture (Treaty) and Bioversity International (Bioversity) is investing in a project to move one step closer to a global information system which is required to ensure that genetically unique diversity is effectively and efficiently conserved and made available for use. SINGER (the System-wide Information Network for Genetic Resources), EURISCO (the European search catalog) and GRIN (the United States Department of Agriculture Agricultural Research Service's Genetic Resources Information Network) are also partners in this project. Progress towards bridging the gaps between existing genebank information systems and networks, facilitating the participation of new genebanks, and linking all these together to provide a single portal will be reported and discussed. For example, 1.2 million Treaty Annex 1 crop passport records from SINGER, EURISCO and GRIN have been migrated into this global portal's data warehouse. Additionally, a schema to completely manage phenotypic data has also been developed and currently holds around 3 million records for characterization and evaluation data for the same 22 crops, mostly sourced from GRIN. Functionality for searching accessions using any combination of passport, phenotypic and environmental data (the latter only for accessions with geo-references) has also been developed. The ongoing development of this global portal, its structure and functionality is an iterative process that necessitates the contributions of stakeholders.

Support is acknowledged from: Bioversity International

8.5. Accessing information on domestic animal diversity

Beate Scherf, Dafydd Pilling
Food and Agriculture Organization of the United Nations (FAO),Animal Production and Health Division

Terrestrial domesticated animals used in food production and agriculture – also known as livestock – are components in the livelihoods of hundreds of millions of people around the world. Globally, the livestock sector supplies one-third of the protein consumed by humans and a range of other products and services (fibre, fertilizer, transport, etc.). Genetic diversity is one of the keys to the sustainability of the sector; it underpins the development of animal populations that are adapted to the environments where they are raised and to the demands placed on them by humans. Effectively utilizing this diversity, and ensuring that the options it provides remain available for the future, requires good access to information.

The world's livestock production is dominated by relatively few species, the most significant being cattle, sheep, goats, pigs and chickens; another 30 or so species of birds and mammals are recorded in the Domestic Animal Diversity Information System (DAD-IS – www.fao.org/dad-is) maintained by the Food and Agriculture Organization of the United Nations (FAO). Most livestock species, however, comprise numerous distinct subpopulations, referred to as breeds. The management of livestock diversity largely involves decisions at the level of the breed rather than the species. Genetic diversity within breeds is also an important resource and has to be carefully managed. Information requirements reflect these decision-making levels.

Information on livestock breeds can be obtained from a range of sources, but the only one with global coverage is DAD-IS, which records more than 7000 breeds from 181 countries. It is a multilingual, dynamic database-driven information system – part of a network of such systems which includes a European regional and 16 national systems.

All breed-related data in DAD-IS are provided by individual countries, who are responsible for their completeness and accuracy. The structure of DAD-IS is, therefore, based on national breed populations rather than breeds per se. Breeds that are present in more than one country, so-called "transboundary breeds", are linked within the system and can treated as single populations for the purposes of global-scale assessments and planning. The range of breed-related data that can be entered into DAD-IS includes origin, morphology, uses, special qualities (e.g. disease resistance), performance and population size and structure. A module for recording details of breeds' production environments, which will also involve georeferencing breed distributions, is being developed.

The wide coverage offered by DAD-IS makes it the best source of information for global assessments of livestock diversity. Several such assessments have been published by FAO, the most comprehensive being The State of the World's Animal Genetic Resources for Food and Agriculture (2007). Status and trends reports will be published biennially. The approach taken is to assign breeds to risk-status classes based on the size and structure of their

populations. The proportion of breeds falling within each class can be calculated and compared the equivalent figures from previous years. While this breed-based approach is currently the only feasible means to provide a global assessment, changes in within-breed diversity and the effects of cross-breeding are not accounted for. Incomplete and out of date population datasets are another problem. No population data are available yet for about 35% of recorded breeds.

8.6. The Crop Wild Relatives Portal: Conservation and utilization of crop wild relatives through better use of information – a global project with Armenia, Bolivia, Madagascar, Sri Lanka and Uzbekistan

Imke Thormann[1], Dany Hunter[1], Armen Danielyan[2], Beatriz Zapata Ferrufino[3], Jeannot Ramelison[4], Anura Wijesekara[5], Sativaldi Djataev[6]

[1] Bioversity International, [2] Armenia, [3] VBRFMA/DGBAP - FUNDECO, Bolivia, [4] Centre National de la Recherche Appliquée au Développement Rural -FOFIFA, Madagascar, [5] Horticulture Crops Development & Research Institute (HORDI), Sri Lanka, [6] Institute of Genetics and Plant Experimental Biology, Academy of Sciences Republic of Uzbekistan

Crop Wild Relatives (CWR) can be generally defined as wild species that are more or less genetically related to crops, but unlike them, have not been domesticated. They are often found growing in disturbed habitats but also occur in protected areas, which to date have been the focus for their conservation in situ. CWRs are under threat as never before and continue to be seriously under conserved both in situ and ex situ. Predictions are that 16-22% of crop wild relative species studied could go extinct by 2055 under certain climate change scenarios. Paradoxically many CWRs harbor genetic traits that could hold the key for many crops to adapt to climate change in the future. The global UNEP-GEF CWR Project involving Armenia, Bolivia, Madagascar, Sri Lanka and Uzbekistan, and coordinated by Bioversity International, aims to enhance awareness, conservation and use of CWRs at the global level so they are safeguarded for the future. The project, which is coming to an end in early 2010, includes a considerable component on information management, an important aspect for enhanced decision-making and conservation. Earlier studies, as well as baseline studies for the UNEP-GEF project have shown that, although data on CWR were available, it was often scattered and hard to access, since it was not in digital format. All 5 countries have now set up national inventory databases on CWR, storing previously existing data from various sources, which in most cases have been digitized during the project life time, as well as many additional records gathered during recent field surveys. Given the different national and institutional contexts and varying levels of expertise and use of software programs, all five national inventories have been designed according to appropriate national preferences and settings. Armenia developed a web-based system with PHP and MySQL, which is used in the institutions that have CWR data. Data is sent through modem connection from the institutions to the central database, which now contains more than 30000 records for 104 species. The Uzbek national database was developed in Access while in Madagascar and Sri Lanka the newly digitized data was first entered into Excel worksheets. Bolivia compiled at least 3010 records for over 160 CWR species. The development of the national systems allowed countries to map distribution of wild relatives in their countries, identify areas for CWR conservation and prioritize protected areas for inclusion of CWR in the protected areas management plans. In addition to the national information systems, a global portal was developed to provide access to CWR information at the global level at www2.cropwildrelatives.org. The national CWR inventories are all searchable through the portal and are linked to it using TapirLink as publishing software, and DarwinCore 1.4 as the data standard. Further information and resources on CWR provided by the portal include publications, searches for projects and experts, news, images and other resources. The choice of freely available and easy to use tools as well as approved and widely used standards makes it easy to link additional national CWR inventories to the portal in the future and to provide a CWR-viewpoint on plant genetic resources data and distribution.

1 Bioversity International, Rome, Italy. i.thormann@cgiar.org

Support is acknowledged from: UNEP

9. Agriculture Information for development

9.1. ARCAD- Agropolis Resource Center for Crop Conservation, Adaptation and Diversity : a new open multi-function platform devoted to agrobiodiversity.

Objectives and challenges ∞

Jean-Louis Pham[1], Jean-Pierre Labouisse[2]

[1] Agropolis Fondation, [2] CIRAD

ARCAD, short for Agropolis Resource Center for Crop Conservation, Adaptation and Diversity, is an initiative supported by Agropolis Fondation, the Region Languedoc Roussillon (France) and several French research institutions.

ARCAD aims at setting up a new open multi-function (conservation, research and training) platform devoted to the assessment and better use of plant agro biodiversity in Mediterranean and tropical regions. ARCAD will sustain the conservation of collections of Mediteranean and tropical crop genetic resources which are currently maintained by local research institutions, and develop innovative research to analyze the diversity of these crops.
Understanding how genes, genomes and populations have been shaped by history, environment and societies is a key factor to enhance the quality and sustainability of germplasm conservation and use. The programme's scientific agenda will thus prioritize the study of history and patterns of crop domestication and adaptation as well as the analysis of key parameters underpinning adaptation and diversity structure, at various time scales, through studies of evolutionary genomics, population genetics and social sciences. The research will focus on Phylogenomics, Crop adaptation to climate change and Cereal crops in Africa.
These activities will be complemented with technological and methodological components for the conservation (DNA bank, cryopreservation) and analysis (bioinformatics, linkage disequilibrium) of crop diversity.
A major objective of the programme is also to set up a demand-oriented capacity building platform, based upon the educational facilities offered by universities in Montpellier and the development of specific training modules.
The programme will generate and use a large and extremely diverse array of datasets. Biological, technical and legal issues about the data generation and management processes are numerous.
The ARCAD programme was jointly developed by French research institutions and universities: CIRAD (Centre de coopération internationale en recherche agronomique pour le développement), INRA (Institut national de la recherche agronomique), IRD (Institut de recherche pour le développement), Montpellier SupAgro and University of Montpellier 2, in partnership with numerous South and international institutions.
As an open platform, ARCAD will continuously seek the involvement of interested partners that are able to add value to this new programme.

Support is acknowledged from: Agropolis Fondation, CIRAD, INRA, IRD, Montpellier SupAgro, Université Montpellier 2, Région Languedc-Roussillon

9.2. BIODIVERSITY NETWORKS IN AFRICA: FROM KNOWLEDGE MANAGEMENT TO TECHNICAL AND INSTITUTIONAL IMPLEMENTATION

Charles Kahindo[1], Franck Theeten[2], Patricia Mergen[3], Garin Cael[3], Michel Louette[3], Olivier Bakasanda[4], Motonobu Kasajima, Patricia Kelbert[5], Jorg Holetscheck[6], Elizabeth Arnaud[7], Dheda Djailo[8]

[1] Universite Officielle de Bukavu, [2] MRAC, Tervuren, [3] RMCA, Tervuren, [4] CEDESURK, Kinshasa, [5] BGMB, Berlin, [6] BGBM, Germany, [7] SINGER, [8] Kisangani University

Africa is one of the most biologically diverse continents on earth. Biological diversity is widespread across different types of habitats, including protected and non protected sites. In different countries, a number of institutions have been involved for some decades in data collection, in most cases in collaboration with international partners.

The CABIN (Central African Biodiversity Information Network) project, supported by the Belgian General Direction of Development Cooperation, has identified the lack of efficient networking as an impediment to the sustainable management of biodiversity data, with an impact on local development. Networking amongst research centres and higher learning institutions is needed to maximize the use of limited expertise, to maximize information sharing and to promote effective and cost efficient biodiversity informatics capacity building activities within the central African region and beyond.

Africa needs to benefit from recent advances in information technology. A survey showed that there is an increased willingness to share primary biodiversity data within Africa. For the method to work, several limitations must be overcome, including a lack of high-quality taxonomic determination, imprecise georeferencing of data, and the poor availability of high-quality, updated, taxonomic treatments.

Geographic scopes and topics for networking greatly vary and can be considered at country, regional and continental levels.

Agriculture is a key area where access to information about taxonomy and biodiversity is crucial. This sector plays a key role as a major economic activity intertwined with peoples' livelihoods in Africa.

CABIN is planning to collaborate with existing institutions and ongoing projects in Africa. As a pilot project, eight collaborators have been selected from Central Africa: four major agriculture research stations (Mvuazi, Mulungu, Yangambi, Nyoka) and four universities (Bukavu, Kisangani, Kinshasa, Yangambi). In collaboration with UniversiTic and CEPDEC (Capacity Enhancement Programme for Developing Countries) of GBIF (Global Biodiversity Information

Facility), CABIN will organize a training workshop in 2010 for scientists from those research centers and universities on standards for digitizing data for easy sharing. In future, this activity may be expanded to other parts of Africa.

Biodiversity Informatics techniques have the potential not only to support fundamental studies , but also to assist developing countries in tackling biodiversity management issues in practical ways.

Support is acknowledged from: Belgian Directorate General for Development Cooperation

9.3. Digitization and richness of type specimen collections at the Paris herbarium - spotlight on the Global Plants Initiative

Pascale Chesselet[1], Jean-Noël Labat[2]
[1] Museum national d'Histoire naturelle, [2] Muséum national d'Histoire naturelle

With over 11 million botanical specimens, the "Herbier National de Paris" (P & PC) features as one of the largest herbaria in the world. Its tremendous richness reflects botanical exploration and collecting over the past 450 years. Estimates for the number of type specimens contained within its Vascular Plant and Cryptogam collections exceed the 600 000s.

The Muséum National d'Histoire Naturelle has participated in an international, multi-institutional collaboration by contributing high resolution images and label (meta) data of type specimens to the African Plants Initiative (API - http://www.aluka.org). Access to African botanical type specimens is no longer an impediment to taxonomic studies of the African Flora and information has now been made available to the countries of origin. This highly successful project was expanded, initially to Latin America (Latin American Plants Initiative -LAPI) then to the Global Plants Initiative (GPI - http://plants.jstor.org).

The primary objective of GPI is to build a comprehensive research tool aggregating and linking scholarly botanical resources around the globe. This has been achieved through collaboration among more than 125 partners in 44 countries. To date, over 500 000 type specimens, 95 000 documents and supporting material are available on its web-based platform that provides powerful tools for research and discovery, collaboration, teaching and knowledge exchange. Through its emphasis on type specimens, the GPI makes primary species information accessible world-wide.

Support is acknowledged from: The Andrew W. Mellon Foundation

9.4. Plant Resources of Tropical Africa (PROTA): a tool for sustainable development

Michel Chauvet
Agropolis International

PROTA is an encyclopaedia of all the plant resources present in tropical Africa. Once completed, it will document about 8000 species, and includes botanical as well as agronomical or management data. The articles are available in book and CD-ROM form, and freely on the Internet. PROTA follows the same format as a previously published series, PROSEA (Plant Resources of South-East Asia).
For the sake of efficiency, a list of species to be included has been compiled, and each species has been attributed one primary use and several secondary uses, with some exceptions when several uses are of utmost importance. Sixteen classes of uses (or commodity groups) have been distinguished, and writing is scheduled according to the commodity groups, which allows to publish books according to the primary use of species. Up to now, Cereals and pulses, Vegetables, Dyes and tannins, Timbers-1 and Medicinals-1 have been published. As all uses of a particular species are documented, the database on the Internet allows queries regardless of commodity groups.
The option of producing readable texts instead of filling closed fields makes PROTA friendly for users, and avoids showing too many void fields. The drawback is that queries can be done only through a free text mode. No systematic effort of standardizing the use of terms has been made, although all the articles go through a strict editing process. Translation (everything being published in two languages, English and French) also allows to check the quality of terminology, in fields as different as the names of human diseases, plant pests and diseases, chemical nomenclature.
As such, PROTA allows teachers to develop courses and all users to prioritize their actions. After completion of each volume, an expert meeting produces a booklet in the series 'PROTA recommends', facilitating decisions by researchers, development agencies, conservationists or policy makers.
The task of writing such review articles is often not given a high profile in research curricula. It must be stressed that they are of great importance for common users, who have a limited access to primary literature, or no time to make a critical use of it. If we are really convinced that the sustainable use of plants has to be promoted globally, PROTA and PROSEA should be seen as first strong steps towards a comprehensive information system on the 'Plant Resources of the World'.

http://www.prota.org/
http://proseanet.org/prosea/

9.5. Sud-Expert-Plantes

Eric Chenin
Institut de Recherche pour le DÃ©veloppement

Sud Expert Plantes (SEP) is an initiative of the French Ministry of Foreign Affairs operated by IRD (Institut de Recherche pour le Développement). Its objective is to help the endeavour of numerous developing countries to know, preserve, and sustainably exploit their plants.

It was mainly designed and is largely lead by South countries scientists ; it aims at building upon and enhancing existing capacities and current programs ; and it endeavours to help network the tropical botany community.

The initiative, scheduled on 4 years (2007-2010), covers 22 countries in Africa, Indian Ocean and Asia ; and is structured in three complementary components:

1. Training & exchange seminars between scientists, policy makers and actors;
2. Support to institutions and networks;
3. Research projects.

SEP addresses the issues of information standards, management, sharing and use within its three components.

Several obstacles prevent South countries from efficiently managing, sharing and using biodiversity data: mainly the lack or scarcity of awareness, technical skills, know-how and Internet access.

SEP attempts to reduce these obstacles in various complementary ways:

• Technical regional training sessions for herbaria staff on collections management and digitisation, where commonly adopted data models are proposed (RIHA, Brahms), advice is given regarding the connection to GBIF, and examples of data use are provided;
• National and regional workshops on various aspects of digitised data sharing and use, including data intellectual property, data cleaning and reliability, data use techniques, and identification of scientific and socio-economic questions requiring occurrence or other biodiversity data;
• Within GBIF CEPDEC program: national meetings for raising awareness and establishing national biodiversity networks around GBIF nodes; and regional training workshops on biodiversity informatics and GBIF Integrated Provider Toolkit;
• Networking of the botany community and enhancing exchange and collaboration, including on data management, sharing and use;
• Support to herbaria and botanic gardens to improve the organisation of their collection, and the preparation and identification of their specimens; and to digitise their collection and connect their database to GBIF;
• Encouraging SEP funded research projects to provide herbaria with new specimens whenever relevant, to digitise and share any produced data, as well as to make use of occurrence data and web services as found on GBIF portal.

Good progress has already been obtained, especially through the technical regional training sessions for herbaria staff, the support to herbaria collections digitisation, and CEPDEC regional training workshops. New countries have joined GBIF, national nodes and biodiversity networks are being set up, and specimens have been digitised in thousands in several herbaria.

The sustainability of these accomplishments is of course a concern, but the level of awareness already reached on the benefits of occurrence and other biodiversity data sharing and use, should ensure the continuation of this endeavour. Also, a think tank has been set up to identify ways of sustaining SEP itself and possibly make it a permanent program with multiple public and private donors.

For more information on SEP, see http://www.sud-expert-plantes.ird.fr/

10. Wild Ideas!

10.1. Convergence of Ontology, Hypothesis, Workflow, and Query ∞

Mark D Wilkinson
Heart + Lung Institute at St. Paul's Hospital

The Semantic Automated Discovery and Integration (SADI) Semantic Web Services Famework was designed to more transparently integrate the output of Web Services, in particular those that expose the output of analytical tools, into the Semantic Web. Such services generate data that does not necessarily exist a priori and as a result cannot be warehoused. SADI web services have a distinct design requirement that is critical to the semantic behaviours I want to present and explore in this presentation - that is, SADI explicitly exposes the semantic relationship between the input and output of a Web Service in the form of one or more RDF Triples. To demonstrate the semantic behaviours enabled by this design requirement, we created SHARE (Semantic Health And Research Environment) - a SADI client application that exposes this dynamically-generated data as if it were a SPARQL endpoint. SHARE generates a transient warehouse of query-specific data through logically matching the triples referred-to in a SPARQL query to an index (registry) of triple-classes provided by all known Web Services. When phrased in this way - that Web Services generate triples of a certain class - a second interesting semantic behaviour of the system becomes apparent. SADI/SHARE can be used to dynamically generate instances of Web Ontology Language (OWL) Classes by first looking at the OWL Class property-restrictions, invoking Web Services generating data with those properties, and then reasoning over the OWL and resulting data warehouse to determine if any Class instances have been generated. Importantly, the OWL Classes may or may not represent a real or known category of data. Effectively these OWL Classes - these ontologies - are behaving both as hypotheses, and as workflows. Decomposition of the OWL Class definition results in a plan (workflow) for generating data that might be capable of fulfilling those class restrictions, and the data is then evaluated to determine if any such instances exist to confirm the "validity" of this hypothetical Class. We are intrigued by the convergence of ontology, hypothesis, workflow, and query exposed by the SADI/SHARE system, and are currently exploring the utility and limitations of this approach.

Support is acknowledged from: Heart and Stroke Foundation of BC and Yukon, Microsoft Research, Canadian Institutes for Health Research, NSERC

10.2. Have Standards Enhanced Biodiversity Data? Global correction and acquisition patterns

Javier Otegui, Arturo H. Ariño
University of Navarra

The Global Biodiversity Information Facility (GBIF) was developed with the prime idea of making biodiversity data freely available for everyone, responding to the increasing demand of basic information for addressing environmental challenges. This vision required that many differently built databases could abide to some standards, and the development of the latter was largely entrusted to TDWG.

Adhering to a standard should, in principle, contribute to data quality: the danger of misinterpretation by the user querying different databases could be removed, as the standard would guarantee at least the semantic contents of each data item. Furthermore, it could be argued that the need to map diverse databases to an agreed-upon standard might lead to discovering errors, or more often gaps, in data availability. On the other hand, however, if data are made available through some standard, failure to comply (e.g. by incorrect mapping) might lead to the injection of error: if not in the original databases, at least on the data as they are served to the user.

In previous contributions (Ariño & Otegui, 2008; Otegui et al., 2009), we made transversal assessments of the quality of some basic pieces of information (the “what, where, when” that form the primary biodiversity data) at the moment of retrieval. The assessment revealed a vast majority of apparently correct data, along with obvious issues that should be addressed. However, were these data correct from the beginning? Were there wrong data that someone corrected at some time, possibly as a result of the application of an exchange standard? Could wrong data have been actually used for research, results of which went uncorrected even after detection of the data errors? And, could correct data have turned wrong because of some standardization mistake?

Observing the evolution of the quality of standardized data over time should enable us to address a critical, overarching question: How far can we trust the world’s available biodiversity data?

Extending over time our previous, snapshot-type assessments we set to try to unveil patterns that may appear on acquisition of new data, on trends and rates of error correction, and on the likelihood of new errors resulting from the implementation of the standardization processes. Our results should help drawing the portrait of what was, what is, and

what is expected to be, about the availability of biodiversity information.

References
Ariño A.H. & Otegui J., 2008: Sampling Biodiversity Sampling. Proceedings of TDWG, 2008.
Otegui J., Robles E., & Ariño A.H., Noise in Biodiversity Data. Contribution to e-biosphere, London, 2009.

10.3. Moving biodiversity to the cloud

Javier de la Torre
Vizzuality

Everybody talks about the cloud nowadays. Cloud computing is a mix of dynamically scalable services for processing, storage and other tasks. Some people think it is just a new name for pre-existing terms, but what no one can deny is that it is changing the way most web applications are deployed, data is processed, served and archived.

Cloud computing is about rapidly and inexpensively re-provisioning technological infrastructure resources, offering cost-effectiveness, scalability and sustainability.

So considering that cloud computing is revolutionizing internet development, is biodiversity informatics taking advantage of it? Could we optimize the way we share, consume and process biodiversity data?

There are still many concerns regarding privacy when using cloud computing and political reasons for not using it, but what could we achieve if none of these issues exist; if we could just think from the most computer efficient point of view?

We think there is great potential for the use of this technology out in the field and we will present different scenarios where we have used it. From Amazon Public Datasets, to Hadoop processing, cheap storage, geospatial analysis, etc., there are many resources that could help us move to a different infrastructure level and that could dramatically change our community.

So if you want to be fashionable, break rules, become 20 years younger and save biodiversity on the way, this what you should be doing: computing in the cloud! Come to the session and see how you can make use of it for groups other than flying birds.

Support is acknowledged from: Vizzuality

11. Taking Data Integration Forward

11.1. Using Distributed Annotations for Continuous Quality Control of Biodiversity Data

Paul J Morris[1], James Macklin[1], Maureen Kelly, Robert A Morris[2], Zhimin Wang
[1] Harvard University, [2] Harvard University Herbaria; University of Massachusetts/Boston

Meaningful federation of scientific data is not attainable without the assessment of the quality and validity of the aggregated data in the context of particular research problems, i.e., its fitness for use.
The Filtered Push platform (http://etaxonomy.org/FilteredPush) implements a network that circulates annotations signaling the location and consequence of potential errors in data, and provides optional ability for corrections to be pushed back to the original data curator. In addition to the network cyberinfrastructure, we have prototyped a domain-independent XML Schema for annotations which has proved suitable for some of the needs we have identified for data quality control. Among these are:
• simple accuracy problems such as errors made during the capture of the data (e.g. spelling, numeric reversal, etc.),
• errors arising from representations and interpretation of the data
(e.g. inconsistencies in local to global concept mapping, unit
conversions), and
• timeliness issues arising from the currency of taxonomic identifications.
Our Java and Web Service APIs support collaborations with other platforms, such as the GBIF Integrated Publishing Toolkit (IPT), for which we have demonstrated the injection into the network of annotations from an IPT client; similarly, we have prototyped interfaces to the Specify6 collection management software for both the injection and acceptance of annotations about data in the local specimen database.
We refer to "Continuous Quality Control" because science, data, or data corrections that emerge after a scientific analysis based on a data set may change the conclusion of the analysis. This changing knowledge at any time, in any

place, is a variant of the Open World assumption and brings two consequences: (1) Any annotation schema or ontology must be able to transport any present or future domain concepts and (2) a notification mechanism such as a publication/subscription overlay is necessary to insure that network participants know when existing annotations (or un-annotated data) are the subject of new knowledge or have become inconsistent with new data.

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

12. Computer Demonstration

12.1. TaxoBrowser: a visual mashup for taxonomic browsing

Stéphane Azard, Julie Chabalier, Amandine Sahl, Olivier Rovellotti
Natural Solutions

In the last ten years, tremendous progress has been made by the Biodiversity Informatics community. Very large online datasets are now available through the Global Biodiversity Information Facility (GBIF) and other online efforts. This has been made possible by the use of the latest technological advances in Service Oriented Architecture. The idea of combining these online data sources into a single interface to provide one page per species was first coined by Roderick Page in his iSpecies.org mashup [1][2].

In order to assist ecologists in their online information collection tasks, we developed an application that weaves data from different sources into a new service. TaxoBrowser is a mashup that combines taxonomic classification, distribution maps, images, and species descriptions in a user-friendly Web site [3].

TaxoBrowser is developed using Flex, the latest RIA (Rich Internet Application) technology from Adobe. Flex is an open source framework for building and maintaining Web applications that deploy on all major browsers [4].

The current version of TaxoBrowser uses different GBIF online services. The first one performs a search from taxonomic classifications and the second embeds distribution maps in a Web page. The user friendly navigation component guides users through taxonomic hierarchies by using a graph where each node represents a taxon that can be expanded by double-clicking. The selection of a taxon triggers an image search through the Yahoo Search Web service and displays a taxon description from Wikipedia.

[1] Page R.D. (2008), Biodiversity informatics: the challenge of linking data and the role of shared identifiers, Brief Bioinform. 2008 Sept. 9(5):345-54. [http://bib.oxfordjournals.org/cgi/content/full/9/5/345].
[2] Butler D. (2006), Mashups mix data into global service, Nature, 439(7072): 6-7.
[3] TaxoBrowser [http://biodiversitydata.blogspot.com/2009/10/taxobrowser-beta.html]
[4] Flex [http://www.adobe.com/products/flex/]

12.2. The Moorea Biocode Project: Tracking Barcoded Specimens from Collecting Event to Sequence ∞

John Deck
University of California at Berkeley

The goal of the Moorea Biocode Project (funded by the Moore Foundation) is to inventory every species of multi-cellular organism on the island of Moorea (French Polynesia) including the surrounding lagoon. The inventory includes specimen identification based on morphology and DNA barcoding. The purpose of the IT component of the project is to build a system to track all information, including collecting event, specimen photos & metadata, tissue samples, through the Laboratory processes for sequencing. An essential component of the project is to interface with multiple partner institution databases and associated downstream databases such as BOLD (Barcode of Life Database) and GenBank.

Support is acknowledged from: The Gordon and Betty Moore Foundation

12.4. A graphical system for computer-assisted plant identification

Pierre Grard[1], Pierre Bonnet[2], Juliana Prosperi[1], Le Bourgeois Le bourgeois[1], Claude Edelin[3], Frédéric Theveny[1], Alain Carrara[1]
[1] CIRAD, [2] INRA (French Nat. Inst. for Agricultural Research), [3] CNRS

Species identification is a major constraint for biodiversity conservation. Conventional identification tools are usually difficult to use for non specialists, mainly because they require important botanical knowledge during the identification process. For this reason, we developed a graphical identification approach that resulted in the IDAO (IDentification Assistée par Ordinateur) software. Through simple clicks on vector drawings, the user selects morphological (shape, size, position, color and texture of organs) or ecological characters corresponding to the plant he/she wants to identify, thus building a sort of "identikit" for the species. The software compares this set to all those available in its database with a simple matching coefficient, and provides a probable identification. At any time during the process, the user may consult species description files. Missing information is tolerated, and users can thus access to an identification result without needing to use all characters in the set. Numerous illustrations are present in each species description file in order to facilitate identification.

This graphic multi-entry identification system has been adapted to various floras (weeds, trees, orchids) around the world (West Africa, India, Cambodia, etc.), for weed control or biodiversity conservation. It is accessible on-line on Internet (http://umramap.cirad.fr/amap2/logiciels_amap/index.php?page=idao), or available on CD-ROM. Current developments for the new version of this identification tool will include (i) a free open version, which will allow adaptation of the graphic interface by users according to their own flora, (ii) generalisation of the use of open drawing format (SVG: Scalable Vector Graphics), (iii) the extension of this approach to new characters (such as anatomical characters of the wood), and floras (such as paddy fields weeds).

Support is acknowledged from: EU, Agropolis-Fondation

12.5. eFlore: An Electronic Flora

GREGOIRE DUCHE

The electronic flora, eFlore http://www.tela-botanica.org/page:eflore, allows users to browse for information on various vascular plants or bryophytes selecting from families or genera or by searching for a scientific or common name. Searches are organized by various regions (France, Guadeloupe, Martinique, Reunion, and North Africa). Each web page includes links to useful information about each plant species such as nomenclatural status, geographical distribution, detailed descriptions, and pictures.

Data are frequently updated depending on user comments and feedback on the plant data checklist update project. Changes may be rolled back to a previous version.

The information, presented in French, is divided in various tabs :

Identité : gives a summary of the database content:
− classic description found in Coste's Flora
− provides information from Baseflore and Baseveg
− allows the user to switch from one description to another by the "Description disponible" drop down list.
− link to access contributed pictures and those found on the web about the taxon.
− the BDNFF (Base de Données Nomenclaturale de la Flore de France, the name and synonyms database for French flora) notes in the « Correspondances » paragraph
− a summary PDF file for the taxon including common names in Greek, Coste's Flora description, details from Baseflore and Baseveg, bibliography and other data sources

Bibliographie: compiles the references from the « Bibliobota » database about the taxon.

Répartition: shows a distribution map with presence/absence data by French administrative units (called « Départements »)

Carnet en ligne: the new online field note-book for botanists http://www.tela-botanica.org/page:menu_395 from Tela Botanica offers users the opportunity to document their observations in a simple and efficient way (with input assistance), and to search and sort the contents. Each observation may include illustrations, distribution, and phenology data. You may choose to publish your observations so that they appear on a map (in eFlore, in the section « Vos observations ») on the Tela Botanica's website. You can also import photos directly into the online field notebook, and associate them with an observation. You can create or use existing tags for the images, insert notes, date, rate your photo, and manage a photo's metadata (EXIF/IPTC) information. A Google Maps interface is planned so that users may display their observations precisely.

Support is acknowledged from: Tela Botanica

12.6. The Biofinity Project: An Extensible Semantic Bridge between Biodiversity and Genomics

Stephen D. Scott[1], Leen-Kiat Soh[1], Etsuko Moriyama[1], Federico Ocampo[2], Mary Liz Jameson[3], Steve Harris[1], Ian Cottingham[1], Shawn Baden[1], Adam Eck[1], Derrick Lam[1], Catherine Anderson[1], Yuji Mo[1], Dan Clark[3], Matt Moore[3]

[1] University of Nebraska, [2] Instituto de Investigaciones de las Zonas Aridas, [3] Wichita State University

Despite the enormous amounts of biological data available, researchers are often hard-pressed to fully exploit it due to heterogeneity, large scale, and decentralization. We present the first release of The Biofinity Project, part of the Semantic Cyberinfrastructure for Investigation and Discovery (SCID) Project, as our answer to research roadblocks raised by decentralized, large-scale, heterogeneous datasets.

Our current version of the Biofinity Project is at http://biofinity.unl.edu. It unifies genomics and biodiversity data, thereby empowering investigation and discovery in both fields. The current release of the Biofinity Project has the following features:

1. It unifies two biodiversity databases (on scarab beetles) and one fungal genomics database, all from the project's PIs. The databases are both browsable and searchable.

2. It offers a Google Maps-based mapping tool, which allows for study of species distribution of collected specimens, and the climate and topological conditions of the regions where these specimens were collected.

3. It offers BLAST (Basic Local Alignment Search Tool), a sequence similarity search tool, which runs in parallel on a 516-core computer cluster called PrairieFire.

4. It offers our new “My Lab” feature, which allows research groups to create an online laboratory database and set up its own user accounts. Using My Lab, a research group can manage its users' ability to access or upload data, to modify existing datasets, and to publish private lab data to the Biofinity Project ontologies—all from a single, easy-to-use web interface.

5. All the aforementioned tools and datasets are integrated under a single user interface.

All functionality is accessible via a web browser. An interface to search and browse the data is also available for the iPhone and iPod Touch.

In the next few months, we will add the following features:

1. Full access to the large, publicly-available datasets at NCBI (National Center for Biotechnology Information) and GBIF (Global Biodiversity Information Facility)

2. More data sets from other Biofinity Project users, including messenger RNA data.

3. Additional bioinformatics tools, such as DesktopGarp (www.nhm.ku.edu/desktopgarp) for ecological niche modeling and ClustalW (www.clustal.org) for multiple biological sequence alignment.

4. A users-only wiki for describing scientific results discovered by users. These descriptions would be summaries of the lab’s published results, and would be immediately disseminated to other Biofinity Project users to build upon in their own research.

In addition to adding even more bioinformatics databases and tools, in the long term we plan to add:

1. More intelligence in the Biofinity Project user interface to track each user's use of the Biofinity Project and to offer suggestions as to what other tools and data sets may be useful to that user's work.

2. Automated support for new users to quickly integrate their new data sets into the Biofinity Project and federate it with the existing data.

Support is acknowledged from: National Science Foundation

12.7. Maximising the potential of digitised literature-INOTAXA prototype and TDWG standards

Chris Lyal[1], Anna Weitzman[2]
[1] Natural History Museum, [2] Smithsonian Institution

Legacy taxonomic literature is rich in information in the form of taxon treatments, biological observations, biogeographic conclusions, as well as the data on which these were based. These data include specimen data, geographical distribution, host records, and collector names. INOTAXA (INtegrated Open TAXonomic Access) is a system, built primarily for taxonomists, which allows them to retrieve what they need from within a corpus of digitised publications. The INOTAXA interface includes simple or complex (Boolean) searches, the latter operating on more than 50 indexed fields including taxon names, geographical terms, associated taxa (e.g., hosts), nomenclatural details, or people such as collectors or authors. These searches return taxon treatments, keys, mentions of the search terms in other parts of the original text, and specimen data. Specimen data can be downloaded for analysis. INOTAXA equals in speed and greatly outperforms the accuracy of retrieval of searches of a pdf or other digitised versions of the original text. In addition to searches, the contents can be accessed by browsing taxonomic or geographic hierarchies, or sorting through a list of all people (e.g., authors, collectors, editors) mentioned in the texts. The output of this type of browsing leads to a range of results including taxon treatments, keys, specimens (collected or determined when associated with people), collecting sites, publications, and images.

The INOTAXA prototype was developed to evaluate, through user feedback, which functions are needed and how the interface might become more intuitive. INOTAXA is based on the highly atomised XML schema, taXMLit, one of two developed for taxonomic literature, which is being considered by the TDWG Literature group. The current focus is in developing a workflow and appropriate tools to facilitate text mark-up, data upload, and mark-up review. Adoption of TDWG standards for literature schemas will facilitate the mark-up and review process and enhance the ability of taxonomists and other users of biodiversity data to retrieve precisely the data they need. There may need to be more than one such standard to serve all of the identified user needs. In addition, literature citations are also being considered by TDWG's Literature Group as a means to compile resources such as standardised and synonymic lists of people (e.g., authors and collectors) and of journals. In these cases, the plan is to use and build on the work already done in the botanical community, which was published as standards earlier in TDWG's history, but which are not currently available in electonic form or easily used by computers. The group will work actively on these standards during this meeting. Responses to the INOTAXA interface are assisting us in refining both the schema and the features of the interface. The prototype being demonstrated may be found at http://www.inotaxa.org

Support is acknowledged from: Atherton Seidell Fund and the National Museum of Natural History of the Smithsonian Institution; the Natural History Museum, UK and TDWG. Protoype developed by Information International Associates.

12.8. GCP Crop Ontology Browser

Martin Senger[1], Rosemary Shrestha[2], Elizabeth Arnaud[3]
[1] IRRI, Philippines, [2] CIMMYT, [3] Bioversity

The GCP (Generation Challenge Programme) Crop Ontology Browser is a software tool that draws on a variety of localized ontological databases with a Web application that allows the user to browse these ontologies intelligently. It can be simultaneously installed (mirrored) on several servers, using the same ontologies maintained by their respective managers through a shared repository. An example installation can be seen at http://koios.generationcp.org/ontology-lookup/.

The GCP Crop Ontology Browser utilizes the files maintained in OBOEdit (http://oboedit.org) to provide updated synchronized ontologies, using the CVS (Code Versioning System) and/or SVN (Subversion - version control system) repositories. Ontology teams share and exchange their views and inputs in this centralized repository.

Updated local copies of OBO files (queried from the CVS/SVN repositories for latest versions) are loaded into the system, changing ontologies into a database. Or, on demand, whole ontologies can be reloaded. The database searches are enhanced by text searching based on a technology known as Lucene indexes (http://lucene.apache.org/).

A Java Web application, once deployed on a server, gets data from the database and serves them as Ontology browser web pages or as an Ontology web service, a traditional SOAP-based (Simple Object Access Protocol) web service that can be accessed by programs written in any programming language.

The resulting system is customized by configuration property files. The most important properties define what CVS and SVN repositories to access, what ontologies to load and browse, and who has the access rights to the ontology database.

The GCP Crop Ontology Browser was derived from the Ontology Lookup Service, developed at EBI (European Bioinformatics Institute) by Richard Cote (http://www.ebi.ac.uk/ontology-lookup). Both the original and our version are distributed under the Apache License (http://commons.apache.org/license.html).

Support is acknowledged from: GCP - Generation Challenge Programme (SP4; IRRI Philippines; Bioversity France

12.9. Using Citizen Science to Process Digital Herbarium Labels

Michael Giddens
SilverBiology

At SilverBiology, we are developing a software engine entitled "SilverArchive" to process typed and handwritten label data from digitized herbarium labels on specimens. Large digitized specimens are provided by collections for processing. These images are loaded into a queue for parallel processing using a website called http://www.helpingscience.org (in closed beta-testing at the time of writing). Citizen scientists sign in to the website to provide three different tasks. The first role is to identify all the label and determination locations on a given specimen sheet, the second is to identify the Darwin Core (DwC) fields within each label, and the third is to type in the text values of each field image.

Once labels have been identified on a specimen sheet, using a mouse to outline the borders, a label image is created and sent to Evernote (http://www.evernote.com) for optical character recognition (OCR). They return the position of every word, all the permutations of each word, and if the label is handwritten or typed. We use this information for making educated guesses and to help in expediting the field tagging process. We try to focus more on human input for accuracy and only use the OCR information as a secondary source.

Each part of the specimen label itself, whether it is the scientific name, date, country, etc., is parsed into associated DwC fields. These tags are assigned by a human using a simple click and drag interface. Once this is completed for a label, each marked field is created into individual images so they can be processed in parallel.

Each tagged field will be examined by three or more distinct users or citizen scientists who all input what they think the field image says. Typing the words is the most time consuming so we are trying different game style interfaces to see which type of game gives us the best response. No user will see the same field image twice. Each field image is circulated until enough people type in the same value, which gives a measure of accuracy. When the predetermined level of accuracy has been reached, the value for the field is accepted. Once all the field values are verified for a given label a DarwinCore record is created.

All processed data runs through a series of taxonomic and geographic validations. Any issues are reported to the collection manager for review. All data will be available in a variety of formats including DwC.

12.10. EURISCO - The European Plant Genetic Resources Search Catalogue

Milko A. Škofič, Sónia Dias
Bioversity International

The EURISCO intranet was developed for the launch of its search catalogue for plant genetic resources in Europe in 2003. This allows European country National Focal Points (NFPs) to upload and update respective National Inventories (NIs) of germplasm holdings. Only crop-independent data on germplasm holdings may currently be included within EURISCO http://eurisco.ecpgr.org/. This is done according to international standards, allowing for comprehensive automatic checks, providing top-quality reports to data-providers.

A facility to upload plant trait observation data into the EURISCO catalogue is being developed using community endorsed data exchange formats (e.g. Multi Crop Passport Descriptors). However characterisation & evaluation data will only be included in the EURISCO catalogue when linked to an existing accession.

The demonstration will illustrate:
- Current upload format. Datasets uploaded via a web-browser as TAB-delimited text files.
- Error checking. Hourly parses of uploaded datasets flag warnings and errors in records containing erroneous data that does not conform to agreed standards. Depending on the selected error-checking profile, one of three actions is taken:
1. The whole upload transaction is rejected; the user must correct errors and re-submit.
2. Records containing errors are rejected.
3. Fields containing errors are cleared; if a required field has an error, the record is rejected.
- Data quality checking. Along with the error checking, two data-quality reports are generated:

1. Taxonomy. A report covering each NI is generated and sent periodically to NFPs. The taxonomy is matched with standard taxonomic catalogues (currently GRIN Taxonomy for Plants). For each taxon, the report indicates whether there is a full match, or up to which rank the taxon matches.
2. Coordinates. Records with collecting site coordinates are matched with the country of origin; a report is generated and periodically sent to NFPs indicating which accession's collecting site is outside the declared country.
- Review. NFPs review the reports and provide feedback to participating gene banks which review, correct and re-submit the dataset. Once the national coordinator is satisfied with data quality, (s)he flags the dataset to be imported into the public EURISCO catalogue.

These procedures have been implemented since 2003 with minimal human intervention. Workflows and technologies were selected to accommodate less technologically advanced inventories. Now the situation has evolved, and considering accumulated experience, some updates are planned:
- Current upload format. TAB-delimited files are still a solid exchange medium, but push or pull web services may help automate the process.
- Standards. Currently, only multi-crop passport descriptors are provided to the catalogue. However, it is planned to add characterization & evaluation data, and material transfer information, so the catalogue structure and upload mechanism will need revision.
- Scope. Currently, the whole inventory is sent, but an incremental uploading workflow will be used in the future.
- Users. Currently, the national coordinator collects data from the participating gene banks, collates it and sends it to EURISCO. It may be possible to allow gene banks to directly provide data to EURISCO and let the system collate the data for the national coordinator.
- Data curation. Data quality curation will be developed further, including enhancing the existing taxonomy, coordinating reports and adding tools to crosscheck passport information among accessions from the same origin.

Support is acknowledged from: Bioversity International

12.11. The Atrium Biodiversity Information System: sharing, managing, analyzing, and disseminating biodiversity data

Mathias W Tobler, John P Janovec, Jason H Best, Amanda K Neill, Anton Webber
Botanical Research Institute of Texas

The Atrium Biodiversity Information System (www.atrium-biodiversity.org) is a web-based software platform developed by the Botanical Research Institute of Texas for managing a wide range of biodiversity data. Atrium can be used to manage institutional data such as herbarium collections or Geographical Information System (GIS) data, or to integrate biodiversity data at a regional scale, where various datasets can be made available through a single portal.

Atrium has a modular architecture allowing for varied configurations specific to the needs of each institution or project. Current modules include a virtual herbarium, a GIS data repository, a bibliographic reference management system, a meteorological data module, and a module for managing and analyzing vegetation survey data. The system supports many popular data and metadata standards including Distributed Generic Information Retrieval (DiGIR) and Darwin Core for collection data, International Organization for Standardization (ISO) 19115 for GIS metadata, OpenURL for bibliographic data, and the International Press Telecommunications Council (IPTC) standard for image metadata.

Atrium provides advanced tools for integrating, managing, sharing, and publishing biodiversity and allied data via the Internet. This includes dynamic distribution maps, zoomable high-resolution images, online specimen annotation, specimen and annotation label printing, dynamic creation of color field guides and checklists, online statistical data analysis, and a wide range of data import and management tools.

Support is acknowledged from: Gordon & Betty Moore Foundation, Beneficia Foundation, Conservation International, BRIT, National Science Foundation

12.12. Become an e-Taxonomist with Xper²

Visotheary Ung[1], Régine Vignes-Lebbe[2]
[1] CNRS, [2] UPMC

Systematists inventory, study and structure biological diversity as precisely as possible. All this information may be summarized into knowledge bases to perform identification of specimens for new inventories and monitoring surveys, phylogenetic analyses, as well as for biogeographic and ecological studies. A significant increase of this kind of digitized information should be expected in a near future due to the deep social and scientific impact of web 2.0 and the multiple "cybertaxonomy" projects to study global biodiversity and climate change. Nevertheless, taxonomic descriptions need to

be readily accessible, comparable, and as unambiguous as possible to be helpful for scientists as well as to be useful for analysis with computers. Computer Aided Identification systems provide users with the resources to relate morpho-anatomical observations with taxon names and to subsequently access other knowledge about the organisms. They are essential for both generators and consumers of biodiversity information.

Xper² offers a complete environment dedicated to taxonomic descriptions management. It assists taxonomists with knowledge acquisition for identification keys and with publication of descriptive data by providing a large panel of tools. For example a set of taxa can be quickly compared in a taxa X characters matrix with colors distinguishing equal, overlapping and discriminating character states. Each piece of content including for example taxon names, characters, and attributes can be documented using texts and images and associated with additional references or external links to keep the source and history of the information. Xper² provides excellent support for automatic on-line publication of descriptive data and free access keys, as well as for exporting of datasets for phylogenetic and systematic research. Xper² version 2.0 focuses on interoperability between systems and can import and export into structured descriptive data format (TDWG-SDD), and export to HTML and Nexus formats.

It is the role of taxonomists to use their expertise to structure taxonomic knowledge and Xper² is a perfect tool to assist them in their quest. The benefits in terms of access to new treatments are endless. Xper² includes several outputs, i.e. interactive free-access key (accessible on-line and/or locally), and it is associated with additional external modules able to generate printed identification keys, diagnoses, descriptions in natural language, or to compute numerical similarities between descriptions.

Xper² is a powerful tool for editing and managing taxonomic descriptions. Users may freely download a Windows™, Mac™, or Linux version in French, English, or Spanish at: http://lis-upmc.snv.jussieu.fr/lis/?q=en/resources/softwares/xper2. Our mailing-list provides users with full support. Users can publish and distribute their work on CD or on-line. With its user-friendly and intuitive interface, Xper² provides an easy way to become an e-taxonomist!

Support is acknowledged from: CNRS, UPMC, MNHN

12.13. Plazi: Building Communities and Software for Increasing the Utility of Digitized Biodiversity Publications

Guido Sautter[1], Donat Agosti[2], Terry Catapano, Robert A. Morris[3]
[1] Plazi, Universität Karlsruhe (TH), [2] Plazi, [3] Plazi, University of Massachusetts

Taxonomic literature includes a body of several hundred million printed and thus hardly accessible pages of highly structured data-rich descriptions. Digitization and semantic markup enables data mining and extraction, such as demonstrated by Plazi.

Plazi's document collection comprises over 500 taxonomic publications, including all literature on ants in Madagascar, all publications on ants worldwide published since 2007, and all Zootaxa papers on ants, fish, and platygasteroid wasps; in all, over 12,000 treatments on over 10,000 different taxa.

Plazi's main markup tool, the GoldenGATE Document Editor, is now in version 3. The new version improves usability, performance, and adaptability. Built on top of it is the GoldenGATE Markup Wizard, an easy-to-use highly automated tool with advanced user guidance through the document markup process. It is most efficient if adapted to a specific type of document, e.g. a specific journal. It allows users to create comprehensive markup in less than a minute per document page.

A GoldenGATE Server hosts Plazi's document collection and treatments. Its Tomcat-based web front-end provides multiple interfaces for accessing the treatments and their details. Through these and remote interfaces, Plazi collaborates with many other institutions and initiatives, both as a donor and a consumer of data. Document markup includes adding LSIDs (Life Science Identifiers) from HNS (Hymenoptera Name Server) and/or Zoobank to the taxonomic names, then uploading previously unknown taxa to both providers in the process. A generic XML interface providing raw treatments is the basis for most of the other services. An HTML-based search portal allows human users to browse the treatment collection, linking to specimen images on Antweb and Morphbank, and visualizing georeferenced occurrence records in GoogleMaps. GBIF (Global Biodiversity Information Facility) harvests occurrence records from a TAPIR (TDWG Access Protocol for Information Retrieval) provider. EOL (Encyclopedia Of Life) harvests treatments from an eXist-based SPM (Species Profile Model) interface.

Upcoming collaborations will further enhance Plazi's document collection. FishBase will join the line of LSID providers

and have new taxa uploaded to their database. Bibliographic meta data will be synchronized with Zoobank, GNUB (Global Name Usage Bank), and BHL's (Biodiversity Heritage Library) CiteBank. Original description treatments will be exported to Wikipedia.

In an upcoming project, Plazi will widen its scope to ecological publications. This includes the tools for marking up such publications as well as the facilities to host them and make them available on the web. As a side effect, occurrence records from taxonomy and ecology will become available as one larger dataset.

Furthermore, plans exist to mark up portions of BHL's vast and steadily growing data set by assembling individual pages to documents, marking up the documents, and exposing the contained treatments through Plazi's interfaces. To handle this huge amount of data, the markup will be handed over to a community of volunteer users. The web front-end of GoldenGATE Server will be extended for community functions, and the interactive document markup will be handled in small web-based dialogs. This alleviates the need for client-side software, and the community members can contribute in small time slices as it suits them, as the dialogs take at most a minute to answer. A voting mechanism ensures data quality.

Support is acknowledged from: Universität Karlsruhe (TH), University of Massachusetts, GBIF

12.14. CleanTax: An Integrated Framework for Mapping Biological Taxonomies and Merging Taxonomically Organized Presence/Absence Data Sets

David Thau[1], Shawn Bowers[2], Bertram Ludäscher[3]
[1] University of California at Davis, [2] Gonzaga University, [3] University of California, Davis

CleanTax is an integrated environment designed to support metadata curators and data integrators who need to compare related taxonomies, or merge data sets that have been collected using terms from different taxonomies.

CleanTax allows metadata curators to create mappings between taxonomic concepts and to check those mappings for logical consistency. Given a small set of relationships between taxonomic concepts, CleanTax can infer new, logically implied, relationships. It provides flexible mechanisms for defining taxonomies, and accepts input in a number of formats, including the Taxonomic Databases Working Group's Taxon Concept Schema (TCS).

In addition, CleanTax provides support for merging presence/absence data sets using mapped taxonomies. In some instances, two data sets that use different taxonomies may be merged into a single data set that precisely combines all the data. However, taxonomy mappings may be uncertain, or may introduce uncertainty when merging data sets. In these cases, CleanTax quantifies the uncertainty and presents the integrator with a set of possible merged data sets. Future work involves presenting integrators with a single best merge, which may involve using terms from the taxonomies that are more general than those that appear in the data sets.

This computer demonstration will use real-world data to show how CleanTax performs its reasoning about taxonomic relationships, and how it merges data sets. It will discuss ways of defining taxonomies to maximize the number of relationships that may be inferred, and/or minimize the amount of time required for inferring new relations. It will also discuss the role of uncertainty in taxonomies, taxonomic mappings, and data set merging.

By demonstrating CleanTax's functionality to the audience for which it was designed, we hope to provide a sense of its abilities and the constraints under which taxonomy mapping and data merging occur. We also hope to gain insight into ways CleanTax may be improved, and to learn which additional features the community deems critical.

Support is acknowledged from: NSF awards IIS-0630033, DBI-0743429 and DBI-0753144

12.15. A biodiversity cartography portal for nature conservationists, scientists, and naturalists

Tania Walisch[1], Guy Colling[1], John van Breda[2]
[1] Musée National d'Histoire Naturelle, [2] Biodiverse IT

The Luxembourg Museum of Natural History is the national node for biodiversity data. The Museum uses Recorder 6 (http://www.recordersoftware.org/), extended with a collection and thesaurus module, to manage and collate its data. A number of local associations and scientific collaborators also use local instances of Recorder 6 and contribute species observation data of plants, animals and fungi as well as biotope observations data to the node database at the Museum. A cache database is generated for web publication at regular intervals and holds non-confidential simplified records linked to biodiversity portals including Biological Collections Access Service for Europe (BioCASE) and Global Biodiversity

Information facility (GBIF) for the international scientific community. These use an internationally recognised data schema (ABCDEFG, Access to Biological Collection Database Extended for Geosciences) for the querying of bio- and geoscientific databases.

Full details of the records held by the Museum, for example the exact spatial references, are of particular interest to those engaged in nature conservation, environmental impact studies, and scientific work at a local scale. In the past 10 years, the Museum has received an increasing number of requests from non-governmental organisations, consultancies, and public bodies active in the domain of nature conservation for downloads of detailed data.

In 2009, a team at the Museum worked to build a geographical biodiversity web portal in collaboration with the Ministry of the Environment and the Administration of Topography of Luxembourg, and with the support of the government's eLuxembourg initiative. The application is Open Geospatial Consortium (OGC) compliant and uses open source software tools for the front end application, for securing web services for providers, and for content management.

The geographic portal is targeted at professional users who get controlled access to the data via secure logins. The application makes use of external map web services like the topography or orthophotography maps from the Administration of Topography. Users can select to display map layers including background topographic maps or polygon maps, for example Natura 2000 zones (http://en.wikipedia.org/wiki/Natura_2000). Quick search facilities allow the user to show all occurrences of a taxon or a biotope on the map or in a named place. Grid and point localisations of occurrences are shown in their original geometries. Advanced search facilities involve multiple filtering by geographical object, taxonomic level, legal protection or threat status of a taxon or a biotope, survey name, determiner, date, and the precision of the spatial references. Geographical selection of data can be done by choosing an existing polygon on the map, uploading a polygon (*.shp) file (http://en.wikipedia.org/wiki/SHP_file), or drawing a polygon on the map. Filter results are shown on the map and in a data grid below the map which displays a user configurable set of attributes for each occurrence. The report can be downloaded in pdf, csv (comma separated values), or rtf (rich text) file formats and the distribution map can be saved as a pdf. The portal is still under development and will be publicly accessible in winter 2009-2010.

Support is acknowledged from: eLuxembourg (Centre des Technologies et de l'Information de l'Etat), Ministère de l'Environnement du Luxembourg, Administration du Cadastre et de la Topographie du Luxembourg

12.16. A Framework and Workflow for Extraction and Parsing of Herbarium Specimen Data

Jason H Best[1], William E Moen[2], Amanda K Neill[1]
[1] Botanical Research Institute of Texas, [2] University of North Texas

Millions of specimens in museums and herbaria worldwide need to be digitized to be accessible to scientists. The volume and heterogeneity of the data are challenging, to say the least, for the digitization effort. A key challenge faced by all biodiversity collections is determining a transformation process that yields high-quality results in a cost- and time-efficient manner. The University of North Texas's Texas Center for Digital Knowledge (TxCDK) and the Botanical Research Institute of Texas (BRIT) are developing a workflow for combining human and machine processes to facilitate the transformation of herbarium label data into machine-processable parsed data. The workflow and framework, called the Apiary Project (www.apiaryproject.org), are made possible through integration of a variety of existing technologies and the application of standards developed by TDWG and the Dublin Core Metadata Initiative. The technology used is entirely composed of open-source components and upon completion, the workflow and framework will be released as an open-source project.

The workflow will be presented to human participants through a web-based application with interfaces focusing on four primary phases: layout analysis, text extraction, text parsing, and quality control. Within the layout analysis phase, individuals are provided with an interface that allows them to identify and delineate regions of interest within a digital herbarium sheet image, which will be analyzed for text extraction. In the text extraction phase, the workflow will take advantage of optical character recognition (OCR) to attempt to recognize text within the region of interest. In cases where OCR is not successful, individuals are presented with an image from a region of interest containing text and are provided with an interface that allows them to transcribe the verbatim text represented on the image. The parsing phase provides users with an interface that allows them to view the verbatim extracted text and to indicate by various methods which parts of the extracted text correspond to the standard data attributes for that region of interest. The quality control phase brings together various interfaces that allow the data curator to confirm, enhance, or correct the data. If available project resources allow, we will also examine how the workflow can integrate with the Natural Language Processing functionality in HERBIS (Herbis is the Erudite Recorded Botanical Information Synthesizer - www.herbis.org), which would allow for automated parsing of the text extracted from the regions of interest.

Fedora Repository (www.fedora-commons.org) is at the core of the workflow and provides an architecture for storing, accessing, and managing the digital objects that are ingested into and generated by the workflow processes. Drupal (drupal.org), an open-source content management system, serves as the framework for the human interfaces into the workflow and purpose-built modules provide the workflow business logic. Islandora (http://vre.upei.ca/dev/islandora) provides the integration between Drupal and Fedora Repository. Various machine processes are integrated into the workflow through the Fedora Repository service architectures. OCR services are provided by OCRopus (www.ocropus.org). The open-source image server, djatoka (http://sourceforge.net/projects/djatoka/), allows for large, high-resolution images to be scaled, displayed, and easily navigated in the user interface.

Support is acknowledged from: Institute of Museum and Library Services, BRIT, University of North Texas

12.17. The Encyclopedia of Life: Pathways to contribution

Cynthia Parr[1], Patrick Leary[2]
[1] Smithsonian Institution, [2] Marine Biological Laboratory

Present and future growth of the Encyclopedia of Life (www.eol.org) is possible only through the combined efforts of its contributors. This demonstration will inform existing and potential contributors of the many pathways for contribution to EOL and how these pathways have been implemented. It will contain walkthroughs starting with newly collected information and culminating with the content visible in EOL.

Much of EOL's recent growth can be attributed to data flow from existing online content management systems. EOL's Content Partner Registry http://www.eol.org/content_partner enables projects to establish and manage their partnership and data flow with EOL. A group created on the popular photo and video sharing website Flickr (http://www.flickr.com/groups/encyclopedia_of_life) has accumulated over 35,000 images and videos from 1,200 contributors over the last year. The vast majority of the media is tagged with at least one Flickr machine tag to identify the name of the featured organism. A script runs nightly to harvest images that have these machine tags and creative commons licenses. The Atlas of Living Australia also advocates contribution of biodiversity photos through this Flickr group (http://www.ala.org.au/news/sharing-images-through-flickr.html). The EOL has also started to index the Wikimedia Commons repository (commons.wikimedia.org), extracting species information and images. The Wikitext markup language provides special tags for taxonomic information, making Wikimedia repositories accessible to biological text-mining tools. Over 50,000 images have been identified in Wikimedia Commons and are currently accessible through the EOL website. EOL's LifeDesks (www.lifedesks.org) are collaborative content management systems that users can create and manage on their own. They provide tools for scientists and educators to manage their own biodiversity information. Each LifeDesk administrator can choose to have their content made public for incorporation by EOL or any other data consumer. All content, but especially the publicly contributed content, is subject to curation by credentialed scientists using a rich set of tools for trusting, untrusting, and rating content.

The EOL infrastructure uses and contributes to the development of TDWG standards and a global names architecture. In particular, LifeDesks and EOL promote the Taxon Concept Schema and Species Profile Model InfoItems. The Encyclopedia of Life has played a role in the development of Global Names Index (GNI, www.globalnames.org). The GNI is a first step in a global names architecture planned to connect biological nomenclators, taxonomists, and data collectors through their taxonomic information. Names information can be contributed to GNI in the form of Taxon Concept Schema (http://www.tdwg.org/standards/117/) XML or files in the Darwin Core archive (http://code.google.com/p/gbif-providertoolkit/wiki/DarwinCore#The_Darwin_Core_Archive) format. EOL content partners can opt to include their names in the GNI index automatically.

Support is acknowledged from: MacArthur Foundation; Sloan Foundation

13. Poster

13.1. Development of open source software for the management of observational data ∞

Michel Deshayes[1], Beatrice Carpy[2], Marie Demarchi[1], Sophie Gras[2], Isabelle Moins[1], Elise Mouysset[3], Aurelien Peironnet[3]
[1] Cemagref, [2] ATEN, [3] Tela-Botanica

In France, many biodiversity actors have recently been aware of the need to have their data more accessible, using efficient tools. In 2007, within the context of the French biodiversity information system (SINP), an inventory of software applications used to collect, store, manage and display observational biodiversity data has carried out by Cemagref. The survey has shown a diversity and heterogeneity of applications, and its results, in the form of a report with annexed forms describing each application and its characteristics, have been put online on the SINP website (www.naturefrance.fr).

The inventory had shown too that a number of new projects aiming at developing new software tools were about to be launched. Many similar tools would then be developed, and each of them would need to be accompanied by similar activities such as training and maintenance. Since almost all projects (if not all) were heavily subsidised, this meant a waste of public money. This analysis lead to the idea of making a step in mutualising developments, maintenance and training by a two-fold action: first by setting up a new website, specifically devoted to such applications (www.outils-naturalistes.fr), managed by ATEN, Cemagref et Tela Botanica; and secondly by creating a software forge for the development of a modular open source application, managed by ADULLACT (Association of developers and users of open source software for administration and local authorities). The project is supported by the Ministry of environment and to date, three software projects have already joined the forge, and a fourth one is about to do. One of the questions unanswered yet is whether similar open source mutualised projects have been launched in other countries to examine possible sources of cooperation.

Support is acknowledged from: Ministère de l'Ecologie, de l'Energie, du Développement durable et de la Mer (MEEDDM)

13.2. EBONE, a project to design and test a European biodiversity observing system, integrated in time and space ∞

Rob H. G. Jongman[1], France F.O. Gerard[2], Philip Roche[3], Michel Deshayes[3]
[1] Alterra, Wageningen UR, [2] Centre for Ecology and Hydrology Wallingford, [3] Cemagref

Measuring and reporting reliably trends and changes in biodiversity requires that data and indicators are collected and analysed in a standard and comparable way. This is valid for a national park, but also for larger areas such as the European Union. However at present all responsible authorities (over 100 national and regional authorities) have different and uncoordinated approaches. Worldwide the problem is even bigger as in different continents species and ecosystems do differ. There is a need to develop a system for a coherent system for data collection that can be used for international comparable assessments.
The EBONE project is developing a system of biodiversity observation at regional, national and European levels as a contribution to European reporting on biodiversity as well as to the GEOSS (Global Observation System of Systems) tasks on biodiversity and ecosystems. EBONE assesses existing approaches on validity and applicability starting in Europe, expanding to regions in Africa and seeking cooperation with projects in other continents.
The objective of EBONE is to deliver:
A sound scientific basis for the production of statistical estimates of stock and change of key indicators that can then be interpreted by policy makers responding to EU Directives regarding threatened ecosystems and species;
The development of a system for estimating past changes and forecasting and testing policy options and management strategies for threatened ecosystems and species.
A proposal for a cost-effective system;
The results contribute to the GEOSS 10 year implementation plan.
The system is making use of existing networks of site observations, wider countryside mapping and earth observation. Techniques are under development for upscaling and downscaling. The system and its representativeness are being tested and validated using existing and new data from Europe and Mediterranean regions outside Europe. Based on the validation we will propose refinements to the system (sites, protocols). A link will be made between the methods, data and observation sites available in different countries and regions as well as with various ongoing projects, available databases and observation and monitoring systems. Important steps are to carry out tests on the data from LTER (Long-term Ecosystem Research) sites in relation with data from nation-wide habitat monitoring programs, to test the representativeness of NATURA 2000 sites and test the usefulness of Earth observation information in relation to field data on habitats. Power analysis of the existing datasets at different levels (species, habitat, ecosystems) is carried out to test representativeness and usefulness of sampling schemes and data sets.
The main outcome will be an integrated monitoring system based on key biodiversity indicators and implementation within an institutional framework operating at the European level. This framework will provide continued access to indicator data for CBD reporting against the 2010 target and form the basis for the continued development of a European Biodiversity Observation system.

The project focuses on GEO task BI 07-01 to unify many of the disparate biodiversity observing systems and create a platform to integrate biodiversity data with other types of information and will deliver a European contribution to the development of a global biodiversity observation system that is spatially and topically prioritised. It also delivers to GEO task EC 06-02.
The beneficiaries are the agencies within and outside the European Union that have the task of biodiversity monitoring, the GEO Community and Biodiversity monitoring NGOs worldwide.

Support is acknowledged from: European Commission and European partners to the project

13.3. AfriBes-Towards a social network of scientific and technical information for Africa ∞

Maxime Thibon
fondation pour la recherche sur la BiodiversitÃ©

In 2005, the Millennium Ecosystem Assessment (MA) was the first global assessment tasked with measuring ecosystem services for human well-being worldwide. One project was the Southern Africa Sub-Global assessment (SAfMA), using a multi-scale approach to assess ecosystem services across three different spatial scales.
In 2006, an international consultation was launched to assess the need, scope and options for an International Mechanism of Scientific Expertise on Biodiversity (IMoSEB). The African consultation provided a set of needs and recommendations for how knowledge could be better harnessed to meet needs of African Biodiversity stakeholders:

- Develop a spirit of information sharing
- Consider a wiki type system
- Create synergy between possessors of traditional knowledge and scientists
- Promote South–South cooperation..

After completion of IMoSEB consultation and the MA Follow-up, UNEP (United nations Environment Program) took the lead to set-up an Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES). The preliminary Gap Analysis underlined the issue of information services and coordination for sharing knowledge and experience as ones of its preliminary findings.

An African social network could be seen as one of the means to create and strengthen social ties among African communities, researchers, and policymakers, and contribute to IPBES efforts. Such a social network could also bring real added-value to existing information and expertise, while fostering their dissemination and use in decision-making processes for sustainable development.

This social network on biodiversity and ecosystem, based on Web 2.0 technologies, and characterized by user participation, openness, interconnectivity and interactivity of web-delivered content will allow envisaging:

- Building up African collective and distributed intelligence
- Using peer-to peer networking
- Fostering dialogue
- Emancipating people and communities
- Creating a forum between information suppliers and producers
- Establishing an E-learning capacity building centre

Support is acknowledged from: FRB-CIRAD

13.4. TermForm : an online Web application to define and share concepts for the development of a trait-based ontology ∞

Marie-Angelique LAPORTE[1], Isabelle MOUGENOT[2], Eric GARNIER[1]
[1] CEFE, [2] LIRMM

Understanding the dynamics of biodiversity requires the integration of organismal responses to various changes in their environment at the level of populations, communities, ecosystems and landscapes. This can be achieved through a functional characterization of organisms using their “traits” (Chapin et al. 2000, Nature 405: 234-242; Lavorel & Garnier 2002, Functional Ecology 16: 545-556). A trait is defined as any morphological, physiological or phenological feature

measurable at the individual level, from the cellular level to the whole organism (Violle et al. 2007, Oikos 116: 882-892) .The number of data sources dedicated to ecology (including plant trait data), has rapidly increased during the last two decades, particularly due to Web development. However, the integration of trait data still remains a challenging issue, in part due to considerable semantic and structural heterogeneity. Ontologies can potentially aid to unify access to such disparate data sources. Consequently, a domain ontology designed for the description of plant traits as well as the plant community is currently under development. We are developing a semantic Web community portal to facilitate collaborative ontology building and ontology re-use. This portal is expected to guide experts to share the design of a knowledge hierarchy of functional traits. The underlying support is provided by Tomcat JSP technology to offer a dynamic Web application and the Java framework Jena to deal with the semantic aspects.
In our poster we are presenting the first layer of our tool, named TermForm. TermForm is dedicated to terminological aspects and is focusing on the identification and the meaning of the terms assigned to functional plant traits. Experts for functional traits will interact with and take advantage of TermForm to rigorously describe well identified traits using their definitions and synonyms, with the possibility of adding new traits if required. Traits are classified into general concepts, such as vegetative trait, litter trait and regenerative trait.
TermForm allows the import and the export of the trait concepts and their descriptions as OWL/RDF specifications. In the near future the trait domain thesaurus will be used as a foundation for the trait ontology building and development. The semantic relations between trait concepts and various instances and axioms will then be specified. A new layer of our tool, termed OntoForm, will support this approach.

Support is acknowledged from: CNRS

13.5. Domestic Animal Diversity Information System - A clearing house mechanism ∞

Beate Scherf
Food and Agriculture Organization

The Domestic Animal Diversity Information System (DAD-IS) developed by the Food and Agriculture Organization of the United Nations (FAO) is a multilingual, dynamic database-driven web-based communication and information tool for the management of animal genetic resources for animal genetic resources for food and agriculture (AnGR). It is recognized as a clearing house mechanism and early warning tool for AnGR by the Convention on Biological Diversity. For many years it has facilitated global and regional analysis of the status and trends of livestock diversity. The Global Plan of Action for Animal Genetic Resources, adopted in 2007 as the first agreed international framework for the management of AnGR, calls for DAD-IS to be strengthened and further developed. DAD-IS has a multilingual interface and content; it is currently available in English, French and Spanish (Arabic, Chinese and Russian are in preparation). It the centre of an expandable global network of information systems (FABISnet), which includes a European-regional and 16 national systems.

DAD-IS provides countries with means to manage and disseminate information on their livestock breeds. Data are entered into the system by officially nominated National Coordinators for the Management of AnGR, and countries take full responsibility for data quality and completeness. More than 14000 national breed populations representing 181 countries, 37 species and more than 7000 separate breeds are recorded.

Anyone with access to the internet can visit the DAD-IS web site at http://www.fao.org/dad-is and make use of a range of tools for browsing and analysing breed-related data. For each national breed population a data sheet can be displayed showing all the recorded information on its origin, morphology, performance, uses, special characteristics (e.g. disease resistance) and population size and structure. An “early warning tool” can be used to display graphically the risk status and direction of the population trend for any breed for which the relevant data are available. Another tool presents population structure and calculates the rate of inbreeding. A “cross table generator” can be used to produce easy-to-read tables showing the number of breeds corresponding to selected criteria (species, country, region, risk status category). Users are also able to browse more than 4000 images. A module for managing data on breeds’ production environments is being developed, which will also include georeferences of breed distributions.

Moreover, DAD-IS provides users with up-to-date news on AnGR management and an extensive library of full text publications and links to other Web resources. Contact details of National Coordinators are listed, so that users can seek further information about a particular breed.

13.6. Semantic data integration for the analysis of the genetic diversity of plants ∞

Wollbrett Julien
CIRAD

Biology has become a scientific area that produces a huge amount of information, and research depends on the availability of this information. The data are distributed across several resources, and the biologists need an access to all available resources to find out new knowledge. Integration systems provide a single centralized and homogeneous interface for biologists to query multiple resources simultaneously. However, biological data integration is a difficult task for two main reasons: (i) new data are being generated all the time; (ii) both resources and data are very heterogeneous.
Generation Challenge Program (GCP) Pantheon is a middleware-based data integration system for genetic and genomic data of plant. This Platform is developed in Java and includes a semantic part.
Our study focuses on the benefits of providing the integration of new data sources. To facilitate data sharing we wish to develop an automated tool that would enable semantic mapping between resources scheme and ontological concepts. This mapping will enable the development of semantic web services. Such a tool could increase the number of integrated data sources.
The semantic mapping between resources and ontological concepts is carrying through the enrichment of a Resource Description Framework (RDF) file. The RDF file was firstly automatically created with the D2RQ tool. This RDF file contains the structure of the selected resource. To enrich the RDF file, a Web form will allow the user to annotate a dataset from one resource with ontological concepts. This RDF file is used to store data about resources into a database (DB) of correspondences. Data stored into this correspondence DB are metadata for resources and ontologies, correspondences between resources and ontological concepts, and data localization into a resource.
We would like to use the correspondence DB to annotate web services with ontological concepts, to browse data associated to a single concept, to search available output concept of a web service for a selected input concept, or create a query builder.
To develop semantic web services, a Protégé plugin named BioMoby Converter was developed. It allows registering automatically an ontology (or part of an ontology) as BioMoby datatypes. These datatypes could be used to annotate semantically input/output of web services with ontological concepts.

13.7. EURISCO – The European Plant Genetic Resources Search Catalogue ∞

Sonia Dias
Bioversityt International

Since the 1970s, major efforts have been undertaken by European countries to collect and safeguard plant species which are either no longer cultivated or simply no longer exist. Genetic materials from these plants are housed in various ex situ collections, or genebanks, which play a primary role in guaranteeing the food security of future generations. Accurate scientific analyses, based on well-documented data retrieved directly from national collections (i.e. National Inventories), are vital to providing both policy makers and genebank managers with important information to develop strategies on the conservation and sustainable use of biodiversity.

To answer common questions relating to plant biodiversity, such as “Who holds what?” and “Where can I find these materials?”, the European region has established National Inventories (NIs) to compile all existing information on plant genetic resource (PGR) collections for each European country. This was the first step towards organizing and sharing biodiversity information in order to safeguard European PGR.

The next step involved the development of a web-based window: EURISCO, a ‘one-stop shop’ based on international standards for information access and exchange, enabling users to search and locate information on crops, forages, wild species, landraces and breeding lines maintained in European ex situ collections using a range of search criteria. EURISCO is now being further developed to include characterization and evaluation data (trait information) for future use.

Today, 40 European countries have established NIs on plant genetic resources, making this valuable data publically available to users through the EURISCO network. EURISCO presently contains passport data on more than one million accessions, representing approximately 5,500 genera and 34,500 species. These accessions represent over half of the ex situ accessions maintained in Europe (estimated at about two million) and roughly one-third of the holdings worldwide. As a key online information resource and European database for material maintained ex situ, EURISCO aims to support universal access to biodiversity information through a network of national PGR inventories, allowing countries to meet their national, regional and international obligations; especially those related to the UN Food and Agriculture Organization’s (FAO) Global Plan of Action (GPA), the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA), and the Convention on Biological Diversity (CBD).

13.8. Orylink: a personalized integrated system for functional genomic analysis ∞

Pierre Larmande

Plant functional genomics requires data integration from several sources. A classical example is the need for cross-references between gene location and the corresponding mutant lines, a feature already present in reverse genetic databases like OryGenesDB or T-DNA express. We recently developed three plant databases specifically designed for rice functional genomics: OryGenesDB, OryzaTagLine, and GreenPhylDB. OryGenesDB is a reverse genetics and genomic database and works together with OryzaTagLine, which contains the corresponding phenotypic description of the mutant lines. GreenPhylDB is designed for comparative functional genomics and links the two model plant species Oryza sativa and Arabidopsis thaliana through ortholog predictions. We developed Orylink to run web queries on remote databases. Using Orylink, biologists can speed up information retrieval across these three databases including FST, mutant phenotypes and Arabidopsis orthologs. The interface supports user logins and profiles. Any user can personalize the system using specific forms to display relevant information synthesized from many data sources. Furthermore, we developed and registered some Web services on the BioMOBY registry that can be used to retrieve genomic location, gene information, germplasm name, phenotype description, and information on Arabidopsis thaliana and rice gene orthologs independently of Orylink. The application is available with many other tools at http://orygenesdb.cirad.fr.

13.9. eRelevé: a suite of solutions for biodiversity data management

Amandine Sahl[1], Olivier Coullet[1], Jean-Francois Leger[2], Yves Hingrat[2], Julie Chabalier[1], Olivier Rovellotti[1], Eric Lenuz[2]

[1] Natural Solutions, [2] ECWP, Missour, Maroc

Ecologists have struggled with environmental data management issues for many years now. It has become urgent to focus on building the technological solutions required for a more modern and efficient process of data collection, exchange, and analysis.

To address this issue, we have developed a suite of innovative solutions for global biodiversity assessment. Designed for biologists involved in observation gathering, the eRelevé Desktop solution structures and organizes biodiversity data (i.e. taxonomy, geography, and protocols). A "relevé" is a set of ecological observations performed in a particular place at a particular time on one or many biological entities. After selecting a geographic location several protocols are promted to the user in order to perform data entry. A protocol is seen as a set of measurements (related or not to a taxon) that can be regrouped logicaly into a single entity. It is possible to add new protocols at a later stage, in order to improve code reuse and extensibility. It is a fully generic solution fitting many of the complex requirements of the ecological community. Data can be exported into different formats such as Excel, KML (Google Earth format) and Shapefiles.

A mobile solution, named Pocket eRelevé [3], was designed to replace the old pen and paper method on field data surveys. Observation data are automatically structured and associated with GPS points. Data can be exported to the eRelevé Desktop solution for analysis.

Desktop eRelevé and Pocket eRelevé are currently used to collect, structure, and record biodiversity data [1] by the Emirates Center for Wildlife Propagation (ECWP) a Research and Conservation project, created by His Highness Sheikh Zayed Bin Sultan Al Nahyan with the aim of ensuring the restoration of a sustainable wild houbara bustard (Chlamydotis undulate) population in Morocco, and in NARC (National Avian Research Center), the equivalent of ECWP in United Arab Emirates [2].

We plan to enrich these solutions with a decision support system making the eRelevé platform a fully integrated toolbox for conservation projects.

[1] Currently, Desktop eRelevé -ECWP records ecological data concerning 120 000 locations with around 15 protocol entries (botany, invertebrates, vertebrates, etc.)

[2] A demonstration video is available here: [http://biodiversitydata.blogspot.com/2009/05/ereleve-video-is-online.html]

[3] Download it here:
[http://ereleve.codeplex.com/]

Support is acknowledged from: Natural Solutions

13.10. STERNA Semantic Web-based integration of digital resources on birds

Stijn Cooleman[1], Guntram Geser[2], Michel Louette[1], Maarten Heerlien[3], Danny Meirte[1], Patricia Mergen[1], Andrea Mulrenin[2]

[1] Royal Museum for Central Africa (RMCA), Tervuren, Belgium, [2] Salzburg Research Forschungsgesellschaft m.b.H., Salzburg, Austria, [3] Trezorix, Delft, The Netherlands

STERNA (Semantic Web-based Thematic European Reference Network Application, http://www.sterna-net.eu) develops a distributed digital library solution that allows for semantic search of, and improves access to, bird-related content from the natural history and cultural domains.

The project is related to the European Digital Library initiative and demonstrates how integrated access to heterogeneous collections of many institutions can be realized based on Semantic Web technologies and standards such as RDF (Resource Description Framework) and SKOS (Simple Knowledge Organization System).

The selected content comprises natural history and biodiversity information, material on wildlife such as documentaries, as well as art and ethnographic objects. A variety of contributors provide services for users interested in birds, ranging from amateurs to land managers to scientists.

As a content provider, RMCA (Royal Museum for Central Africa) contributes information on ornithological publications and specimens as well as ethnographical objects composed of feathers or beaks. These data are tied to taxonomic names.

The RMCA works to enrich content semantically, and advises on compliance with biodiversity information standards, on interoperability and enhancement of tools, and on methods to extend the STERNA network.

The semantic integration of knowledge and content resources into one Web-accessible network is an ongoing challenge. On top of a basic semantic layer provided by interlinked thesauri, classification schemes, and other knowledge organisation systems, domain and core ontologies will be needed to allow for higher-level integration, reasoning, and other capabilities. The semantic search prototype will be available in December 2009.

With respect to natural history and biodiversity resources, the core ontology developed by the Taxonomic Database Working Group (Technical Architecture Subgroup) and/or simple classes from their Life Science Identifier (LSID) metadata vocabularies may allow for some ontological alignments.

STERNA is a showcase project for using semantic technologies and standards (like RDF and SKOS) and demonstrating the capability they provide to link, search, and access heterogeneous bird-related collections from the natural history and cultural domains. Project results will be of interest to digital library initiatives such as Europeana and the Biodiversity Heritage Library (BHL), and those providing the discovery and mobilisation of biodiversity data like the Global Biodiversity Information Facility (GBIF).

13.11. The Diversity Workbench Framework: Data retrieval with DiversityMobile and Dataflow from DiversityMobile to GBIF

Markus Weiss[1], Stefan Jablonski[2], Gerhard Rambold[2], Dagmar Triebel[1], Bernhard Volz[2]
[1] IT Center of the SNSB, [2] University of Bayreuth

The Diversity Workbench database framework consists of single application components collaborating through agreed software interfaces. Further information about the Diversity Workbench can be found in the contribution "The Diversity Workbench framework as data repository for biological data". The gathering of data in the field is organized in the component DiversityMobile which is set up to enter, modify, or – if necessary – delete ecological and biological monitoring data in the field via a mobile device (www.diversitymobile.net). The data model used in DiversityMobile is fully compatible with the data model of DiversityCollection which is the component for the storage of collection and observation data within the Diversity Workbench. DiversityCollection provides the possibility to export the data to GBIF via the BioCase-Wrapper in the TDWG standard ABCD.

The field data retrieved via the mobile device using the DiversityMobile client are stored in a database on this device. This local database contains a set of definitions, e.g., lists of taxonomic names and project-specific settings. Moreover, the user can choose to download field data from DiversityCollection, already stored there. Furthermore, it is planned to allow data import directly from web services e.g. for taxonomic or other biological backbone data. As soon as the user starts gathering data in the field, additional data are added to the mobile database. This includes the option that the mobile database contains subsets of data from the central DiversityCollection database. Hence, the upload of field data to the central database and the dataflow to and from DiversityMobile, e.g. to avoid duplicates, needs to be strictly organized by a complex synchronization process which is explicitly described in the poster.

The user can access the synchronization component via a special interface which is installed on his personal computer. This component connects the mobile database with a data repository e.g. via the internet. The synchronization data are stored in a separate database. This database is linking the data items on the client-side with data items on the server-side using Global Unique Identifiers (GUIDs). The data are stored in the local master database DiversityCollection. For a

further consistent data flow it is crucial that the database is set up at an institutional data repository which acts as GBIF data provider like the IT Center of the Staatliche Naturwissenschaftliche Sammlungen Bayerns (SNSB IT Center). There the GUID is connected with a persistent Hypertext Transfer Protocol (HTTP) Uniform Resource Identifier (URI). The latter is published and accessed by global networks like GBIF.

Support is acknowledged from: Deutsche Forschungsgemeinschaft: Förderbereich LIS – Informationsmanagement

13.12. The Diversity Workbench framework as data repository for biological data

Markus Weiss[1], Dagmar Triebel[1], Gregor Hagedorn[2], Stefan Jablonski[3], Gerhard Rambold[3], Tobias Schneider[3], Bernhard Volz[3]

[1] IT Center of the SNSB, [2] Julius Kühn-Institut Berlin, [3] University of Bayreuth

The database framework Diversity Workbench consists of databases and client components interacting via interfaces. Because of this, each component of the Workbench can be directly used as a stand-alone application but can also be called from other clients, without exhibiting details about the internal design and implementation (encapsulation principle). This results in an increased flexibility concerning the technical design, allowing a differentiated user administration and the rapid setup of user-adapted entry forms for specific projects. It facilitates the direct access to web services and external data resources. Among the more than ten components, one is providing GIS (Geographic Information System) functionality and another is set up for mobile platforms with synchronisation mechanisms. The framework is appropriate to store different kinds of data about interorganismal interactions and facilitates the management of institutional collections. The Diversity Workbench therefore represents a modularized system for multiple scientific purposes concerning analysis and management of biological data, e. g. collection data, observation data, descriptive and ecological data, taxonomy, taxon-mediated checklist data, ecological plots, images and literature data. It is also flexible enough to build the software backbone for long-term institutional data repositories for biological data.

A large number of biodiversity research projects depend on field mapping and ecological data of high quality. Therefore it is necessary to link data sets gathered in the field to a verified backbone information from major biological, taxonomic or environmental data sources, e.g., lists of taxonomic names. Further on, there is a need to link additional information like Global Positioning System (GPS) coordinates and multimedia information, such as images or sound, at the time of data gathering. A seamless and transparent flow of data from the field to a central data storage system which may simultaneously be used by several participants is a necessity. “Seamless” in this sense means that data are available as stored datasets shortly after gathering; and “transparent” in the sense that every operation applied to data can be traced back ("data provenance"). For these core requirements, the Diversity Workbench includes a mobile application used to gather biological research data in the field that enables the dataflow to the data repository. The application DiversityMobile is currently designed as a mobile system for entering ecological and biological monitoring data already in the field. It is capable of accomplishing the core requirements mentioned above. For reasons of model consistency the mobile client uses a subset of the database model of the central database of the Diversity Workbench framework. See http://www.diversityworkbench.net and http://www.diversitymobile.net and the contribution “The Diversity Workbench Framework: Data retrieval with DiversityMobile and Dataflow from DiversityMobile to GBIF” for further information.

Support is acknowledged from: Deutsche Forschungsgemeinschaft Förderbereich LIS – Informationsmanagement

13.13. Ca-SIF: A SHARED AND STANDARDIZED CATALOGUE TO PROVIDE INFORMATION ON FOREST ECOSYSTEMS

Wilfried Heintz, Guy Landmann, Damien Maurice

GIP Ecofor

The growing need for reliable information on forests (taxon assessments, conservation concerns, forest planning, assessments of the potential effects of climatic change, etc.) calls for a system to enhance data mobilization, structuring, and accessibility.

A good knowledge of the available data and data providers is a first step towards an advanced “Information System”. However, while numerous experimental and monitoring sites and networks exist, a general overview of the French forest sector does not yet exist.

This situation led the French public forest research platform, Ecosystèmes Forestiers (ECOFOR), to propose the construction of a shared resource, a Catalogue of information sources on forests – Ca-SIF. The objective is to provide, on a public website, standardized descriptions of information sources. Ca-SIF hopes to become a tool to share forest information and to improve networking among stakeholders using forest data.

We used an international standardized metadata format as well as standardized web services for spatial information (ISO 19115/139 [International Organization for Standardization] and Open Geospatial Consortium [OGC] specifications). Our tool is fully compliant with the European INSPIRE (Infrastructure for Spatial Information in Europe) directive. This directive aims to buid a spatial data infrastructure for the European Union, to enable the sharing of environmental spatial information among public sector organisations and better facilitate public access to spatial information across Europe.

Nevertheless, there is still a lack in this standardization effort to share our metadata in a proper way with other metadata formats. The next step consists in mapping these metadata with relevant formats of the biodiversity informatics community (ABCD, DarwinCore, etc.).

We will present the details about the Ca-SIF project including the underlying metadata model, the technical implementations, and an outline of some of its applications.

13.14. Metadata – a Core Component of the GBIF Network

Éamonn Ó Tuama, Tim Robertson
Global Biodiversity Information Facility

Tasked with developing a global informatics infrastructure to enable discovery and access to diverse biodiversity resources, the Global Biodiversity Information Facility (GBIF) can play a central role in such global spatial data infrastructure initiatives as the Global Earth Observation System of Systems (GEOSS), and in particular, its Biodiversity Observation Network (GEO BON), the goal of which is to enable data integration and interoperability across systems serving a wide variety of data (e.g., species occurrences, ecological, environmental, climatic or oceanographic data).

GBIF is designing its informatics infrastructure as a scalable, distributed architecture that adheres to international standards for data exchange formats and protocols thereby enabling the maximum degree of interoperability across heterogeneous, distributed data holdings and applications. Large distributed networks featuring numerous providers and consumers of data can be characterised as a Service Oriented Architecture (SOA) where consumers discover providers and loosely coupled system components can be orchestrated as required to serve particular use cases. In an SOA, the key activities of inventory, discovery and access must be well coordinated through provision of registries and metadata catalogues, and through generation of specific indexes.

Metadata is thus a central component in an expanded GBIF network that would offer many types of web services for delivering data to users, and client applications that can use the data in novel and specialised ways that supplement the mapping/visualisation and ecological niche modelling applications already available.

This poster will outline GBIF's plans for implementing a metadata framework for a decentralised GBIF network.

Support is acknowledged from: Global Biodiversity Information Facility

13.15. Enhancing the Access and Publication of Biodiversity Data in Central Africa: The CABIN Technical Infrastructure

Franck Theeten[1], Patricia Mergen[1], Olivier Bakasanda[2], Jörg Holetschek[3], Patricia Kelbert[3], Montonobu Kasajima[2], Garin Cael[1], Charles Kahindo[4]
[1] Royal Museum for Central Africa, [2] CEDESURK, [3] Botanischer Garten und Botanisches Museum Berlin-Dahlem, [4] UOB - Université Officielle de Bukavu

Since 2008, the Royal Museum of Central Africa has maintained CABIN (Central African Biodiversity Information Network) with the support of the Belgian Directorate General for Development Cooperation. CABIN promotes the ease of access to biodiversity data published on the Internet for researchers from Central Africa, as well as the publication of data from local datasets to networks such as GBIF (Global Biodiversity Information Facility).

In March 2009, a milestone was reached at the CEDESURK ("Centre de Documentation de l'Enseignement Supérieur Universitaire et de la Recherche de Kinshasa"), where the first part of the required technical infrastructure for accessing and publishing data was implemented.

In order to facilitate the maintenance of this infrastructure as well as the integration of local biologists and IT scientists into existing scientific networks and communities, tools and software were developed within the framework of existing projects and well-known standards such as ABCD and DarwinCore.

This poster gives an overview of the technical architecture of the installed components:

1) For accessing data, a biodiversity portal based on the BIOCASe portal (Biological Collection Access Service for Europe) has been installed at the CEDESURK facilities, with the help of the Eb@lé project, which supports the Internet backbone of academic institutions from the Congo Democratic Republic (see : http://cabin.ebale.cd). This application features an enriched and multilingual query interface that can be customized with presentation templates. The integration of the portal into the CEDESURK infrastructure was made with support from the BGBM (Botanischer Garten und Botanische Museum) which developed it.

2) The BioCASe Provider Software was installed to publish data; the integrated QueryTool is being used as an interface for the local scientists, again after having customized the original presentation templates in order to better visualize multimedia content. In 2010, CABIN will issue a call for local researchers who might be interested in publishing their data on the Internet through this platform. The project will also provide support and training for maintenance of the infrastructure by local institutions.

These two tools, developed by the BGBM, enable separate functional components to query data providers independently from the presentation templates. The CABIN project illustrates that this distinction allows the installation of the same application in different institutions and technical contexts while facilitating the sharing of common data and metadata among different projects.

Support is acknowledged from: DGDC (Belgian Directorate for Development Cooperation)

13.16. Canadensys : unlocking Canada's biological collection information

Desmet Peter, Anne Bruneau, Luc Brouillet
Institut de recherche en biologie végétale

Biological collections are replete with taxonomic, geographic, temporal, numerical, and historical information. This information is crucial for understanding and properly managing biodiversity and ecosystems, but is often difficult to access. Canadensys (www.canadensys.net), operated from the Montréal Biodiversity Centre, is a Canadian-wide effort to unlock the biodiversity information held in biological collections. In its initial phase, the network focuses on data from three of the most ecologically diverse and economically important groups of organisms: plants, insects, and fungi. Despite recent efforts, the number of fungal and insect species remains difficult to estimate and our knowledge of these groups is still relatively poor. At this point, the network includes 11 participating universities, five botanical gardens, and two museums, which collectively house over 13 million specimens. We aim to digitize and georeference 3 million specimens (20%) in the next five years, and share these data via a network of distributed databases, compatible with the Canadian Biodiversity Information Facility (CBIF - www.cbif.gc.ca) and the Global Biodiversity Information Facility (GBIF - www.gbif.org). Canadensys will use and promote biodiversity information standards (www.tdwg.org) like Darwin Core, the TDWG Access Protocol for Information Retrieval (TAPIR) and Globally Unique Identifiers (GUIDs). Collection managers will publish their data using the GBIF Integrated Publishing Toolkit (ipt.gbif.org). A central webportal will allow access to the network's specimen data (including images and geospatial information) in combination with other data such as names from the Database of Canadian Vascular Plants (VASCAN) and the Catalogue of Life (www.catalogueoflife.org). By enabling the sharing of these data, Canadensys will allow for their synergistic cross-analysis with geospatial and environmental models. This will enhance both our understanding of global environmental issues and the development of sound biodiversity policies across the country.

Support is acknowledged from: Canada Foundation for Innovation

13.17. The importance of a good balance between physical collection management and digitisation. How SYNTHESYS aims for a benchmark standard throughout Europe.

Garin Cael[1], Patricia Mergen[1], Robert Huxley[2], Simon Owens[3]
[1] Royal Museum for Central Africa, [2] Natural History Museum London, [3] Royal Botanical Garden of Kew

Where a good digitisation project proposal will have a decent chance to attract funding, it is nigh impossible to secure funds for the physical maintenance of the actual specimens. Funding agencies argue that physical collection care and management fall under the core business of a natural history museum or research institution, and thus warrant no additional external funding.

The danger is that digitisation and biodiversity informatics as a whole is an ever changing item that needs continuous updating and adapting. Indeed, one might not be too far off when claiming that new digitisation techniques are invented much faster than any institution could digitise their collections using the existing techniques. If this means that

collections are slowly deteriorating whilst there is a continuous flow of money towards digitisation, then there is an imbalance. Surely to ensure a good quality of digital information, there must be an equally solid base of well maintained, accessible specimens.

Within the larger project of SYNTHESYS (Synthesis of Systematic Resources), the Network Activity C focused on improving the collection management standards throughout Europe. This has been done by assessing the collections management in a large number of European natural history museums and evaluating them, using a standardized set of criteria and comparing them to a benchmark standard. These evaluations have in some instances been successfully used to obtain funding for better collection management.

The award-winning Daubenton Leonardo mobility project (http://www.aef-europe.be/index.php?Rub=leonardo) focuses on the transfer of expertise and knowledge of collection managers and technicians between European institutions. It allows for a collection manager or technician to incorporate himself in the collections management team of the target institution of his choice, in mutual agreement with the host institution. As the host doesn't need to pay wages for the person involved, this benefits both sides, without raising an extra cost. Quite a large number of institutions possess expertise or techniques that are often poorly known to their colleagues throughout Europe and in many European countries, technical staff only gets replaced after the retirement of their predecessors, thus causing the loss of valuable knowledge. With the mobility scheme, they have the chance to spread their knowledge to their colleagues abroad.

Support is acknowledged from: European Union, SYNTHESYS, Leonardo, AEF Europe

13.18. The Musa Germplasm Information System Enhances Knowledge of Banana Diversity

Max Ruas[1], Stéphanie Channelière[1], Guilhem Sempere[1], Manuel Ruiz[2], Elizabeth Arnaud[1], Ines Van den Houwe[1], Jean-Pierre Horry[1], Nicolas Roux[1]
[1] Bioversity International, [2] CIRAD

Bananas (Musa spp.) are a staple food and vital source of income for millions of people. These livelihoods in developing countries depend on over 1000 traditional varieties that are mostly consumed locally.

Because Musa cultivars are usually seedless, their genetic diversity must be conserved as full-size plants or plantlets, in field collections or in in vitro genebanks. More than 6000 accessions are conserved in about 60 Musa national collections. The Global Musa Germplasm Collection (ITC) in Belgium, managed by Bioversity International, stores more than 1400 Musa germplasm accessions in trust. The utilization of the germplasm in the collection depends on the availability of information relating to the characteristics of each germplasm accession. In 1997, the Musa Germplasm Information System (MGIS) was developed. It is a global exchange system and the most extensive source of data on Musa genetic resources. It contains information on 5522 accessions managed in 22 banana collections, including passport data (where and when the germplasm accession was collected, donated or developed), botanical classification, morpho-taxonomic descriptors, and evaluation data (characteristics such as agronomic traits, disease, and stress tolerance) as well as many different photographs. Each participating collection enters and manages its own accession data, which is centralized by Bioversity. Links have been created to external data sources such as the System-wide Information Network for Genetic Resources (SINGER), under which FAO in-trust accessions held by ITC are published. MGIS has been recognised by the Generation Challenge Programme as a model system for storing accession-level data. However, it represents an incomplete dataset due to either the lack of capacity or motivation by several collections to contribute to it. The database has undergone two upgrades (see new release http://www.crop-diversity.org/banana/) and new features should be made available in the coming months, such as links to a molecular database (TropGENE DB), Geographic Information System (GIS) information, data quality control and inter-collection data comparison.

Support is acknowledged from: Bioversity International

13.19. Biodiversity Heritage Library for Europe – Interoperability of European biodiversity digital libraries

Adrian Smales[1], Kai Stalmann[2], Henning Scholz[2]
[1] Natural History Museum, [2] Museum fur Naturkunde

A serious barrier preventing the implementation of the Convention on Biological Diversity (CBD) of the United Nations is the lack of access to essential information on animals and plants. Much of this important information can be found in the scientific books and journals of the past centuries. At the moment, the only way to access this knowledge is to visit a number of different, geographically dispersed, specialist libraries. This complicates much of the fundamental research in

biological science. Since 2007, the Biodiversity Heritage Library has been systematically removing this fundamental barrier by making access to this literature easier via the Web. Managed by the Museum für Naturkunde Berlin, the Biodiversity Heritage Library for Europe (BHL-Europe) started on 1 May 2009 within the framework of the EU program eContentplus. BHL-Europe will now further develop, expand, and enhance the Biodiversity Heritage Library by bringing together the extensive collections of biodiversity literature held in major European natural history, botanical, and research libraries.

BHL-Europe is currently setting up a prototype portal to test the metadata harmonisation and content ingestion of the 15 content providers of the project consortium. All partners have metadata and content generated from their respective workflows, in different formats, and from various Information Technology (IT) environments. The functionality of the test portal will be continuously improved as the ability to handle these differences is resolved. The first user requirement survey is underway to take user needs into account and refine our approach.

The interface of the Portal will be multilingual, enabling users to search in their native language. In the future, the Portal will be modular to allow users and interested parties the ability to extend its capability by adding functions and services. This will help BHL-Europe to address the needs of the various users of the Portal. It will also facilitate the exploitation of our results by other projects and the integration of BHL-Europe services into related initiatives.

In addition to the BHL Portal, all the literature will be accessible through Europeana, the portal of the European Digital Library (www.europeana.eu). This first significant contribution of science material to Europeana will be available in the summer of 2010. For the first time, the wider public, citizen scientists, and decision makers will have unlimited access to this important source of information.

One component of BHL-Europe is the building of a large data centre to ensure long-term preservation and curation of the digital content provided by the partners. This data centre will also serve as the main European mirror of the content for BHL on a global scale. It will be complemented by additional local or regional data centres and access nodes.

The main goals of BHL-Europe are the interoperability of existing repositories and the implementation of technological solutions for search and retrieval and for long-term sustainability of the digitized objects. It is designed as a best practices network, not focused on the task of digitization, which is within the competence of each content provider, but supporting the strategic aims of implementing digitizing programmes.

Support is acknowledged from: European Union, Biodiversity Heritage Library

13.20. Pl@ntWood: A computer-assisted identification tool for 110 species of Amazon trees based on wood anatomy

Carolina Sarmiento[1], Christine Heinz[2], Pierre Détienne[3], Pierre Bonnet[4]
[1] CIRAD, UMR AMAP, [2] UM 2, UMR AMAP, [3] CIRAD, UPR Bois Tropicaux, [4] INRA, UMR AMAP

World interest in conservation of tropical forests has increased due to elevated rates of deforestation and climate change issues. Tropical forests are threatened by extensive agriculture and timber trade among other factors; thus, sustainable management and conservation of tropical tree species require reliable and user accessible identification tools. Although wood anatomical features provide a considerable amount of information, only a handful of experts are able to use it for plant species identification. Here, we propose an interactive tool, based on vector graphics, illustrating 96 states of 22 wood anatomical characters from 110 Amazonian tree species belonging to 34 families. Pl@ntWood is a graphical identification tool based on the IDAO (Identification des plantes Assistée par Ordinateur) system, a multimedia approach to plant identification. This system allows non-specialists to identify plants in a user-friendly interface while stimulating self-training in wood anatomy of tropical species.

Support is acknowledged from: Agropolis Fondation

13.21. EDIT Community Single Sign-On

Lutz Suhrbier[1], Andreas Kohlbecker[2], Andreas Müller[2]
[1] Freie Universität Berlin, [2] Freie Universität Berlin, Botanic Garden and Botanical Museum Berlin-Dahlem (BGBM)

The European Distributed Institute of Taxonomy (EDIT) platform, as well as biodiversity providers in general, provide a multitude of web-based taxonomic applications and services. Also, the diversity of service providers reflects the highly distributed, cross-national organisational infrastructure of taxonomic institutions and collections. This results in a problem of identity management. While the provider's system administrators have to register users and maintain individual access control lists for each offered service, users have to remember a variety of login/password combinations

to use all these different services.

Therefore, EDIT implemented the Community Single Sign-On (CSSO) security infrastructure. CSSO protects and provides access to all EDIT platform components. While service providers keep the sovereign power on the provided collection data and information infrastructures, users now get access to any attached EDIT platform component with a single user account. These fundamental enhancements have been achieved through the adoption of the Security Assertion Markup Language (SAML), a standard protocol, which addresses the specific requirements of the EDIT platform.

Since, organisational and informational infrastructures of EDIT project partners are quite similar to those in the general biodiversity community, our approach may motivate other providers to join or extend our upcoming EDIT federation.

Support is acknowledged from: European Distributed Institute of Taxonomy

13.22. Capabilities and interfaces of a prototype Filtered Push network

Zhimin Wang[1], Maureen A Kelly[2], David B Lowery[3], James A Macklin[2], Paul J Morris[4], Robert A Morris[3], Donna Tremonte[2]

[1] University of Massachusetts, Boston, [2] Harvard University Herbaria, [3] University of Massachusetts, Boston; Harvard University Herbaria, [4] Harvard University Herbaria/Museum of Comparative Zoology

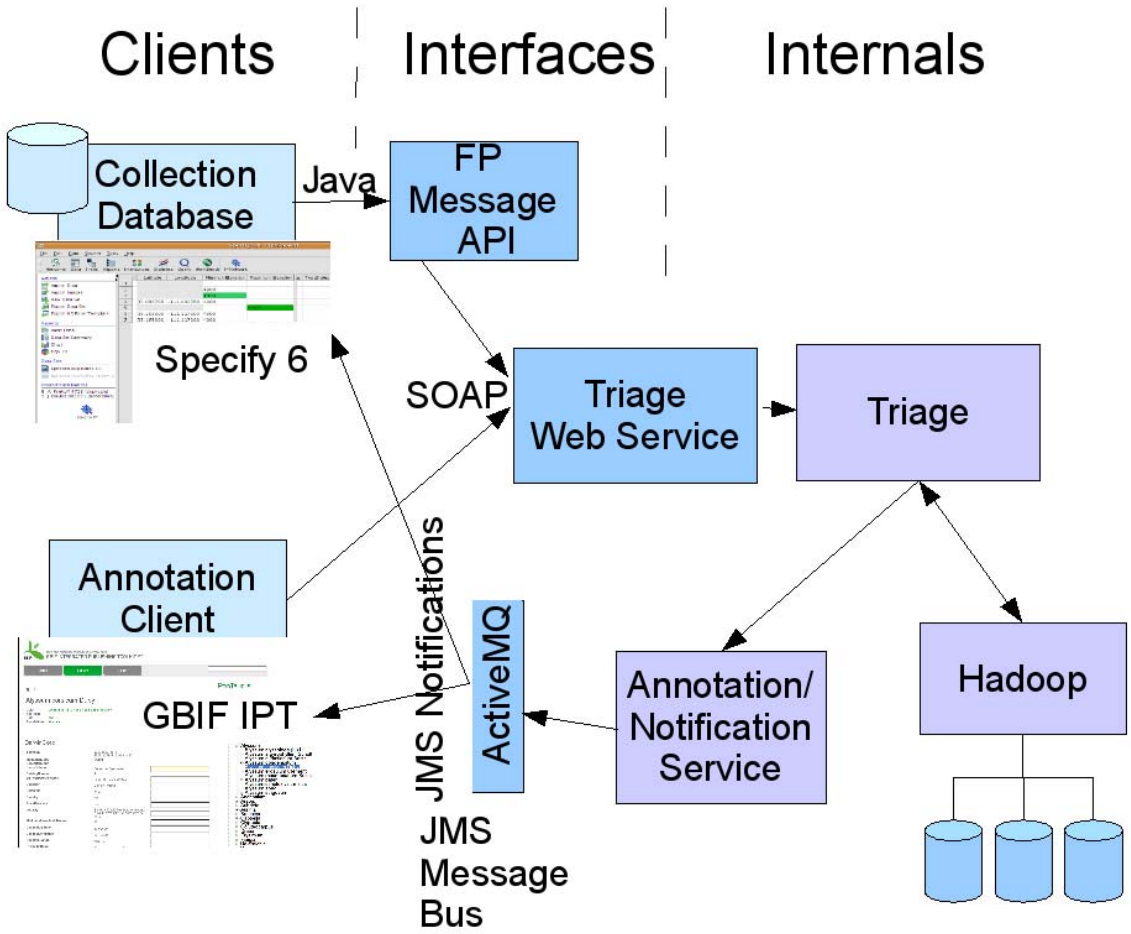
A natural history collection curator or researcher is not much concerned with the implementations of transport and distribution of annotations of distributed data (i.e. distributed specimen databases) by software over a network. Rather, the concerns are, more fundamentally, how to communicate annotations (e.g. new determinations) to the community, how to receive annotations from the community, and perhaps how to receive only annotations pertinent to the curator or researcher. Motivated by these concerns, we have designed and implemented a prototype Filtered Push (FP) network in the domain of botanical collections built over the open source Map-Reduce platform Hadoop and the Java Messaging Service (JMS) ActiveMQ.

The FP software is a domain-neutral framework for originating, distributing, and analyzing record-level annotations. Questions of how to communicate and discover annotations are answered by the FP client applications. However, as the FP applications are written to interact over a network of clients using Java application programming interfaces (APIs) and web service interfaces, issues of transport and distribution are paramount to the software developer who develops or maintains the client software for the curators and researchers who use it. We present our prototype FP clients as a particular example of the general nature and purpose of those interfaces.

In our prototype demonstration, we have developed two user-level implementations of FP clients. First, we modified the Specify 6 natural history collections management software to send and receive annotations concerning herbarium sheet duplicates. (The discovery of duplicates is the original use case for the FP architecture, as data capture from only one of a set of duplicate sheets is one means for efficient data capture in herbaria). Next, we modified the GBIF Integrated Publishing Toolkit to send annotations to the community of clients on a FP network. These user interfaces allow a specialist to annotate specimen data with information that is semantically targeted to concepts represented in the specimen record. The FP network makes use of the semantic nature of the annotations to direct them to interested parties.

The attached figure broadly illustrates how the FP architecture addresses the questions of transport and distribution of annotations over a network of participating clients. The FP “Triage” module is the heart of an FP network's dispatch of incoming annotations. It may invoke analysis tools (e.g. fuzzy matches on collector names to propose putative duplicate specimen sheets) after which it may place the annotation and results of the analysis into a global annotation store, which we presently implement on top of the Apache Hadoop Map-Reduce framework. In addition to (or instead of) that, it may invoke the Annotation Service module which, among other things, is responsible for publishing notices about the annotation onto message queues using the Active MQ open source implementation of the Java Messaging Service standard interfaces. Finally, using JMS service calls, client-side code can receive current messages on JMS message queues to which they subscribe. Clients may have subscribed (or recorded a desire to subscribe) to queues of interest to them (e.g. any queue of annotations about the Orchid family), or this subscription may have happened automatically if the client has launched a message whose type in our annotation Schema indicates it expects a reply.

Support is acknowledged from: National Science Foundation DBI:0646266



Overview of Clients, Interfaces, and Internals of a Filtered Push network.

13.23. Crop Ontology: a controlled vocabulary for trait descriptions for maize, wheat, chickpea, sorghum, banana and plantain, potato and rice

Rosemary Shrestha[1], Ramil Mauleon[2], Reinhard Simon[3], Jayashree Balaji[4], Stephanie Channelière[5], Martin Senger[2], Kevin Manansala[2], Thomas Metz[2], Guy Davenport[6], Richard Bruskiewich[2], Elizabeth Arnaud[5]

[1] CIMMYT, [2] IRRI, Philippines, [3] CIP, Peru, [4] ICRISAT, India, [5] Bioversity International, [6] CIMMYT, Mexico

In order to support and encourage researchers and breeders to use and share information among agricultural databases, the Generation Challenge Programme (GCP) has emphasized the need to build a crop ontology. At present, the Crop Ontology (CO) contains controlled vocabularies that relate to traits for chickpea, maize, banana and plantain, potato, rice, sorghum, and wheat. Moreover, vocabularies related to germplasm and multi-crop passport data (provenance data for accessions) are also included in CO so that researchers and end users may query the keywords of geographic information, traits, plant structure and growth stages, and facilitate retrieving phenotype and/or genotype data e.g., germplasm, genes, and QTL (quantitative trait loci), which already exist in the GCP database. The GCP Crop Ontology browser is now available at http://koios.generationcp.org/ontology-lookup/ for searching ontology terms or a specific ontology hierarchy present in CO. With the inception of the Crop Ontology project, the rice mutant ontology has been integrated as a GCP crop ontology resource, and the controlled vocabulary of each mutant phenotype is now an ontology term in the GCP rice mutant ontology. Implementation of a standardized or controlled vocabulary in existing databases has been a challenge for the present work. However, integration of the International Crop Information System (ICIS) model and other crop databases with the application of the CO has begun, which will enable researchers to query phenotypic data using the CO terms in these databases. In addition, CO has started using the text mining tool, Terminizer (http://terminizer.org/) to capture ontology terms in documents and publications. The submission of CO terms as new

terms to the Gramene Trait Ontology (TO) and Plant Ontology (PO) is ongoing. Further collaborations are being organized with teams in the Food and Agriculture Organization (FAO), who are developing the AGROVOC (Agricultural Thesaurus) related ontology, the Thai Rice Ontology, the Plant Ontology Consortium, Solanaceae Genomic Network (SGN) (for potato trait ontology), and MaizeGDB (Genetics and Genomics Database).

Support is acknowledged from: Generation Challenge Programme, Bioversity International, CIMMYT, IRRI, ICRISAT, CIP,

13.24. Using the Electronic Field Guide Project's software to make digital guides ∞

Robert Stevenson[1], Robert Morris[2], Robert Sheldon[3]
[1] Biology Department, UMass Boston, [2] Computer Science Department, UMass Boston, [3] UMass Boston

The Electronic Field Guide Project (EFG) (http://efg.cs.umb.edu) was conceived with the idea that digital technologies would make it easier for authors to create their own field guides and that electronic guides could bypass the inherent multimedia limitations of printed guides. Here we address two questions 1) Do the current EFG products (see http://efg.cs.umb.edu/EFGsoftware/) allow authors to take full control of the field guide production process? 2) Do the guides that authors produce take advantage of the electronic medium?

To make guides, authors organize their data using spreadsheets for visual keys and taxonomic descriptions (http://efg.cs.umb.edu/efg/). Spreadsheet entries contain links (folder plus file name) to images or other multimedia files. Based on the experience of 11 authors, the input process works well, once they understand the format and have seen several examples. Eleven picture keys and 24 guides have been produced on the EFG website, with four more in production. Subjects have included taxonomy (dragonflies, tree families), life form (trees, shrubs), ecology (Mimics of Clearwing Butterflies), and location (Flora and Fauna of Sailor's Home Pond) based groups. The number of taxonomic descriptions in these guides range from 16 to over 200.

While the authors can readily control the kind of data they want to display, the software that serves and displays data does not entirely meet the goal implied by the first question. After importing and configuring the data, authors select from a variety of templates for displaying the information. An option to make templates using eXtensible Stylesheet Language Transformation exists, but none of the authors have attempted this. Once installed, the EFG package is relatively easy to use, but installation can be tricky mainly because the EFG software, Tomcat, and MySQL do not configure themselves as easily as biologists have come to expect from commercial products.

Keys and guides are viewed with a browser. Users can print keys and their own picture guides in standard paper sizes for easy lamination. On Windows machines, entire guides can also be put on USB thumb drives and executed without requiring any installation. The same thumb drive installation also supports the EFG authoring software. There is no option in the EFG software for authors to make a stand alone website as some have requested, but we publish a documented API (Application Programming Interface) that will allow web designers to call the EFG taxon page generator from any web page.

As to the second question, authors have been taking advantage of the options offered by the electronic medium. All have used color images, often in multiple sizes, allowing many details to be seen. Authors are taking advantage of the opportunity to include several forms or life stages of featured species and to provide details of life history and ecology, which would be cost-prohibitive using a printed guide. Often guides have just two images per description, but several have six or more. This allows authors to illustrate variability. One of the EFG software features not found in standard field guides is the option to include links to similar species. About half of the guides use this option, which requires careful thinking by the authors. The EFG software has the core elements that authors have requested, but to achieve wider use, it needs to reach the standards of installation and user interfaces expected in commercial products.

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

13.25. Making the TDWG Ontology understandable for experts of other scientific and scholarly domains

Karl-Heinz Lampe, Mark Fichtner
Zoologisches Forschungsmuseum Alexander Koenig

Knowledge representation in terms of information integration and cross-domain collaboration among various scientific and scholarly disciplines becomes more and more relevant for natural history museums and other so-called "memory institutions" such as libraries and archives. In this respect ontologies (formal representations of sets of concepts and the

relations between these concepts) play a key role. While top level core ontologies are basic ontologies describing very general concepts across all domains, domain ontologies such as the TDWG ontology are restricted to elementary parts in terms of their respective domain (here biodiversity).

Without general and commonly understandable concepts, the knowledge represented by domain ontologies usually is only understandable for experts of the respective domain, or even worse, just for the people who created the ontology. Domain ontologies can be mapped to top level core ontologies to make scientific and scholarly concepts of the respective domain commonly understandable.

The TDWG Ontology can be used to express dependencies and relations between certain concepts in the domain of biodiversity. The current draft of the ontology was presented at TDWG 2006 and includes an implementation in the Web Ontology Language (OWL). The TDWG Ontology is based on the four existing TDWG XML schemas: Access to Biological Collections Data (ABCD), Darwin Core (DC), Structure of Descriptive Data (SDD), and Taxon Concept (Transfer) Schema (TCS).

The International Committee for Documentation [in museums] - Conceptual Reference Model (CIDOC-CRM; ISO 21127) is used as a top level core ontology. The CIDOC-CRM (v. 5.0.1) is a compact object-oriented model consisting of 84 named classes or entities (E1, E2, etc.), which are interlinked by 137 named properties (P1, P2, etc.). With these entities and properties everything up to the whole world can be modeled. In this context, the CIDOC-CRM is a lingua franca of a transdisciplinary research approach. The identification and communication of common concepts (a central idea of transdisciplinarity, as opposed to interdisciplinarity) enabled by the CIDOC-CRM standard due to the definition of common upper level abstractions and relations is a generalized scientific approach.

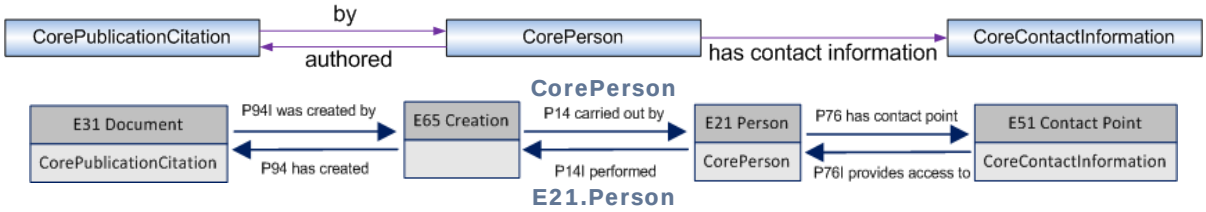
Three central concepts of the TDWG Ontology are exemplarily mapped to the concepts of the CIDOC-CRM. These concepts are CorePerson for the representation of living actors, CoreBioSpecimen for biological objects (voucher specimens etc.) and finally CoreTaxonConcept as the respective concepts identified by scientific names.

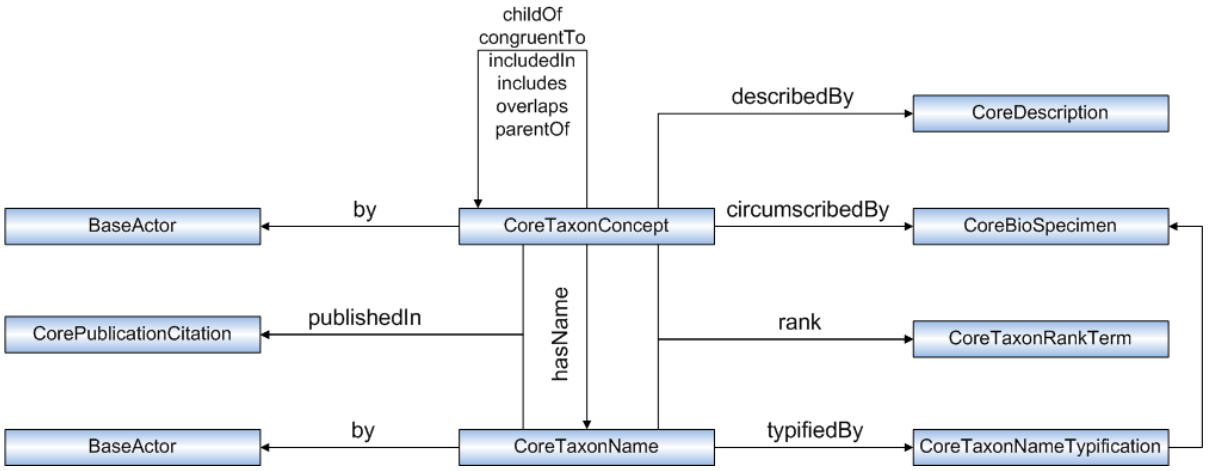
In the context of a top level core ontology domain specific concepts become commonly understandable in a domain neutral form (even without knowing the respective terminology). This is a requirement for transdisciplinary research approaches.

At the implementation level the CIDOC-CRM uses OWL-Description Logics (OWL-DL) enabling more complex semantic constructions while still being processable by machines in contrast to the TDWG Ontology which uses OWL Lite.

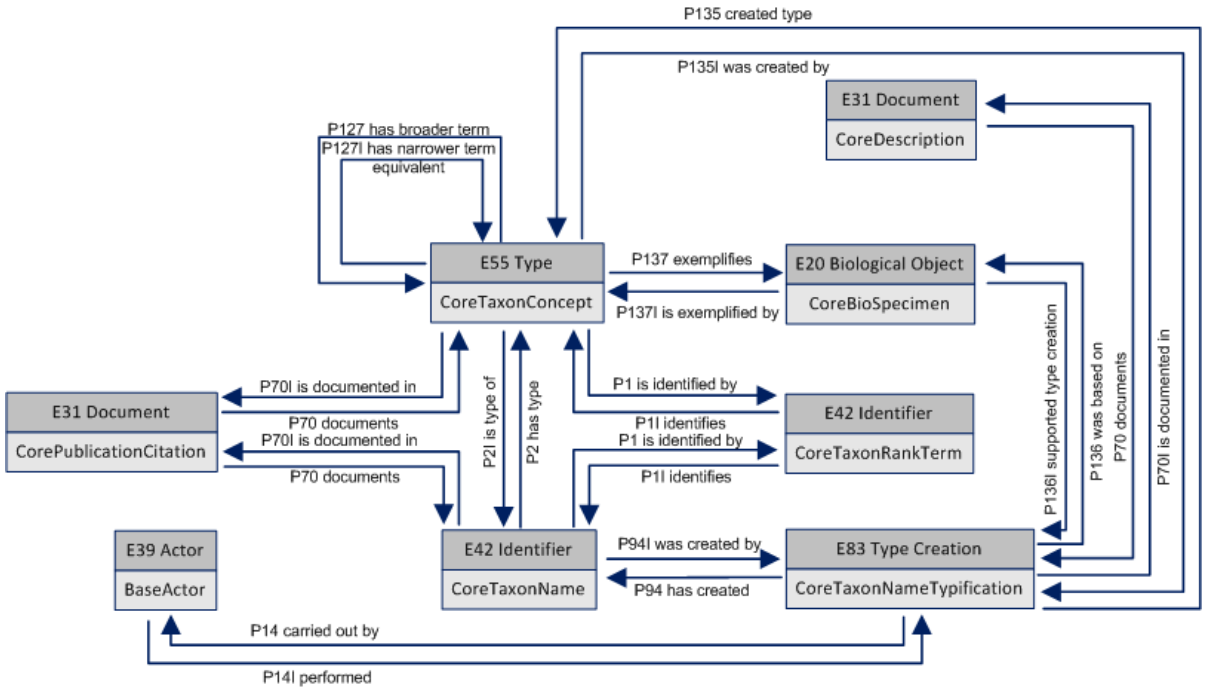
Acknowledgements: This study is part of the research project “WissKI — Wissenschaftliche KommunikationsInfrastruktur” (scientific communication infrastructure) and funded by the German Research Council (DFG).

Support is acknowledged from: German Research Council (DFG)

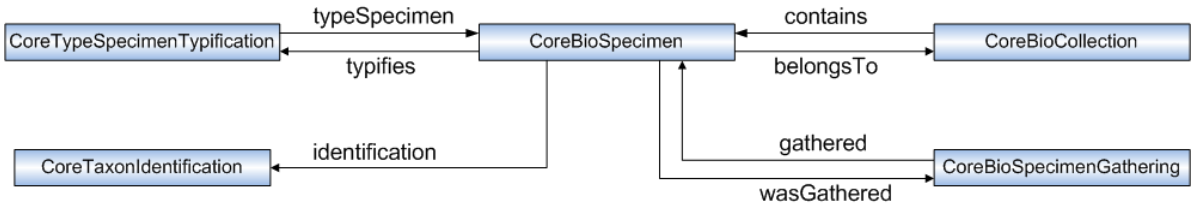




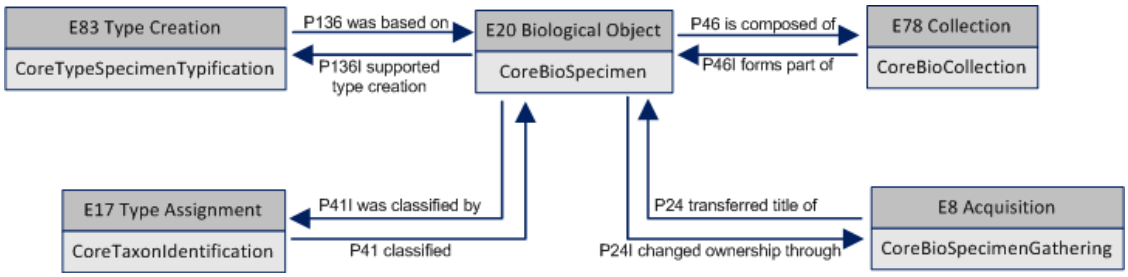
CoreTaxonConcept and CoreTaxonName



E42.Identifier



CoreBioSpecimen



E20.Biological_Object

13.26. Siregal: the INRA Plant Genetic Resources Information System ∞

Sophie Durand[1], Erik Kimmel[1], Cyril Pommier[1], Jean-Marie Prosperi[2], Delphine Steinbach[1], Hadi Quesneville[1]

[1] INRA-URGI, Route de Saint Cyr, 78000 Versailles, [2] INRA-DiAPC, Domaine de Melgueil, 34130 Mauguio

The French National Institute for Agricultural Research (INRA) manages genetic resource collections for more than 50 species (model species and crops). These genetic resources are regularly used for research programs of INRA or its partners, and are also widely distributed to the scientific community. The objectives of the Genetic Resources Centers (GRC) are to gather, conserve, and provide high quality materials for the scientific community. In order to do that, the GRCs have to be able to trace their actions and to assure the community of a high degree of quality.

The INRA Plant Genetic Resource Information System (Siregal) (http://urgi.versailles.inra.fr/siregal) fulfills the essential need to manage the collections and associated data following the recommendations of the Biological Resource Centers (OECD). Siregal is suitable for all plant species and is used by INRA staff and its partners. It respects community standards, and it is possible to integrate existing and future genetic data.

The web interface of Siregal may be used by the general public to discover biodiversity through genetic resource collections, and researchers can choose varieties based on selected criteria and can order plant material.

The stored data are of two types: (1) generic multicrop passport descriptors, including taxonomy, country of origin, biological status (wild, mutant, hybrid, etc.), and pedigree; and (2) a detailed characterization, specific for each taxon, including: morphology, agronomy, resistance to diseases, etc.

The current version of Siregal is hosted on the URGI's server (Unité de Recherche en Génomique-Info INRA), and includes the model's core data: accession (=germplasm, genotype), taxonomy, provenance, pedigree, and simple phenotypic descriptors. It currently contains the main French National Collections hosted at INRA, including grapevine, wheat, cherry, pea, and chestnut. Other plant species will follow. Other national or international projects studying biodiversity can also post their data, and restrict access to it.

In the next two years, Siregal will strengthen its integration into the URGI's global information system (GnpIS http://urgi.versailles.inra.fr/gnpis/). Accessions are already connected to genotyping data of molecular polymorphisms. But they will also be connected to phenotypic and environmental data through the Ephesis project (http://urgi.versailles.inra.fr/projects/Ephesis/), to physical and genetic map data, and to gene expression data, all stored at URGI. The strong data integration will provide significant value to genetic resources and facilitate access to all experimental data that characterize them. Links to external databases (GBIF, for example) will also enrich the information.

The data contained in GnpIS (and Siregal) will be able to be exported for analysis in many scientific applications: for example, variety identification aids, extraction of core collections (subset of a large germplasm collection, maximizing the genetic variability), diversity analysis, computing a coefficient of parentage, studies of phenotype as an interaction between genotype and environment, and population genetics.

Dedicated software is currently being evaluated to be installed locally for the daily management of the Resource Centers. This software will address specifically the management of plant material using barcodes, and orders for plant material.

Support is acknowledged from: INRA for funding, URGI Bioinformaticians and Siregal User Scientific Committee for their work, Siregal-Ephesis Steering Committee

13.27. LSIDs for managing biological names in data integration

Nina Laurenne, Mikko Koho, Arto Mertaniemi, Hannu Saarenmaa
Finnish Museum of Natural History

Biological nomenclature is a hierarchical system and change is its essential nature as taxon names reflect advances in systematics and taxonomy. A taxon may have multiple names or the same name might refer to different taxa. The species content of higher taxa can change radically due to splitting and lumping of taxa. The name itself does not necessarily carry useful information, but the information content increases remarkably if the context of the name is included. This applies specially to taxonomically difficult groups that have undergone several revisions. A taxonomic concept defines the meaning of a taxon name, and in its simplest form covers the biological name and the author. In cases where opinions differ on the limits of a taxon, the taxonomic concept clarifies the usage of a name.

Most taxonomic name servers (TNS) are based on only one preferred taxonomy and nomenclature, thereby making the integration of large data sets problematic. In addition to misspellings and synonyms, conflicting taxonomic concepts have led to redundant misinformation in databases. To overcome the problem, we have developed a taxonomic database with Life Science Identifiers (LSIDs) applied for each taxonomic concept (http://www.luomus.fi/taxondev/?lang=en_US). Thus one concept binds together all alternative names that are used for a given taxon.

The LSIDs representing these non-congruent concepts (i.e. alternative taxonomy and nomenclature) can be linked according to the relations of taxa. Three kinds of relations can be set to link taxon names: 1) synonyms, 2) being part of or belonging to another taxon, and 3) overlapping taxa. Names are uploaded into the system in the form of checklists and then linked to concepts by authorised experts. A software service or an individual user can send a query about taxon names or LSIDs to the taxonomic database, which returns an appropriate result (ie. a valid name or a LSID from a relevant publication) in a RDF (Resource Description Framework) format.
To guarantee the stability of LSIDs, they are generated by the Finnish Museum of Natural History. Also, different versions of LSIDs provided by Species 2000 and the Catalogue of Life will be resolved upon request. The system allows adding and applying identifiers generated by other providers.

Biological identifiers are required to promote efficiency in large scale data sharing, and therefore, LSIDs of the concepts need to be incorporated into databases. In an ongoing project in the Nordic region, multiple Lepidoptera databases will be incorporated with LSIDs. These datasets will in the near future form one large integrated dataset of millions of records. This pilot project manages two butterfly superfamilies (Papilionoidea and Hesperioidea) occurring in Scandinavia, Estonia and NW Russia. Altogether, there are nearly 50 000 zoological names available in the database. The implementation of LSIDs is a user friendly solution as the end user does not have to be aware of changes in classification and can follow the nomenclature in regionally published checklists.

This e-infrastructure project is regionally focused to Scandinavia including neighbouring countries mentioned with their fauna and flora, and it is funded by NordForsk. The taxonomic database project is carried out in collaboration with the FinnONTO project (Helsinki University of Technology), which develops technology for ontologies and semantic computing.

Support is acknowledged from: Nordforsk Foundation

13.28. 4D4Life Integrated Development and Documentation Infrastructure for Sustainable Software Production

Núria Torrescasana Aloy, Dennis Seijts, Gideon Gijswijt, Wouter Addink, Peter Schalk
ETI BioInformatics

The Species 2000 & ITIS Catalogue of Life (CoL)
The Catalogue of Life (http://www.catalogueoflife.org/info_about_col.php) is planned to become a comprehensive digital catalogue of all known species of organisms on Earth. Rapid progress has been made recently and the 2009 Annual Checklist edition contains over 1.1 million species compiled from 66 taxonomic databases from around the world. It contains contributions from more than 3,000 specialists. 4D4Life (Distributed Dynamic Diversity Databases for Life), a Scientific Data Infrastructure project within the European Union's Seventh Framework Programme, will enhance the service-based distributed architecture of the Catalogue of Life, making it a state of the art e-science facility.

Towards sustainable quality software
ETI BioInformatics (http://www.eti.uva.nl/) focuses on sustainable biodiversity software production. It has, in collaboration with the 4D4Life Systems Design Team, created an integrated software development infrastructure to support production-quality software for deployment across the Catalogue of Life networks. The infrastructure is available for all participants that develop software in the 4D4Life project.

Software development infrastructure
A development server has been installed at http://dev.4d4life.eu with three main components:

1. Version Control System. Versions of existing software used for the Catalogue of Life have been collected and stored in a central repository for development versions. Code in the repository is automatically checked for syntax errors.

2. Issue Tracking System. All tasks for software development, bug reports, user stories, acceptance tests, and suggestions for improvement are stored in a customizable issue tracking system that caters to both technical and non-technical users and supports team development. Issues are directly linked to software code in the repository. A developer can only submit code if he links it to the related issue assigned to him in the issue tracking system.

3. Continuous Integration System. Continuous Integration (CI) is one of the pillars of modern programming techniques. CI involves automatically building and testing an application at frequent intervals. This reduces the risk of integration issues appearing late in development and also reduces the time needed to prepare a release. Development tools for automated testing and automated code documenting have been integrated.

Support is acknowledged from: European Union (e-infrastructure)

13.29. The GBIF Data Portal

José Cuadra, Samy Gaiji, Andrea Hahn, Tim Robertson
Global Biodiversity Information Facility (GBIF)

The Data Portal of the Global Biodiversity Information Facility (GBIF), first launched in 2007, was a proof of the concept that a world-wide distributed network of biodiversity data publishers could be linked together and made searchable from a single point of access. The GBIF Data Portal allows complex searches on any taxon, country or dataset, or on a combination of a variety of parameters.

Since 2008, a series of critical improvements to the existing portal were made in response to user needs. For example, searches by altitude, depth and images are now integrated within the existing search filter. Other improvements like the availability of Open Geospatial Consortium (OGC) services, such as the Web Map Service (WMS) and Web Feature Service (WFS), have been incorporated into the GBIF Data Portal.

In 2009, the GBIF infrastructure is being upgraded so that two fully dedicated servers will be assigned as database backend to the Data Portal and Web services. One additional powerful server will handle the geospatial services to meet the exponential demand from the user community. Finally, with the release of the Harvesting and Indexing Toolkit (HIT) a more robust infrastructure is now in place to enable fast and easy indexing of data publishers in a more frequent and distributed manner.

Data Portal and Web Services usage statistics showed a rapid growth in terms of visits (+135% since January 2009). An important percentage of visitors are using the GBIF Data Portal on a more regular basis, such as returning for 10–50 visits/year, which indicates that the level of user loyalty is also increasing for a large portion of the GBIF Data Portal.

New Features

- Upgraded the map user interface to GeoServer.
- Several graphical user interface enhancements made based on user feedback.
- New statistics about data publishers included.
- Inclusion of new content from the World Database on Protected Areas (WDPA).
- Web Services performance improvement with faster response times.
- for more information, see: http://data.gbif.org/version.htm

Resources
http://data.gbif.org GBIF Data Portal

http://code.google.com/p/gbif-dataportal/ Project Home Site: documentation, downloads, source code, bug reporting, etc.

13.30. The GBIF Harvesting and Indexing Toolkit (HIT)

Kyle Braak, Andrea Hahn, Samy Gaiji, Tim Robertson, Markus Döring
GBIF

The Global Biodiversity Information Facility (GBIF) aims to be the preferred gateway, worldwide, to a comprehensive, distributed array of primary species-occurrence data. In moving 'towards full operation' of a fully distributed network architecture, the key focus in portal design was to enable customisation by GBIF Participant Nodes to their local needs, through appropriate and user-friendly tools. In 2009, particular focus was given on simplifying the process of publishing data as well as to improve the frequency of data indexing. The GBIF Harvesting and Indexing Toolkit (HIT) is a software platform developed by GBIF (http://www.gbif.org/) to manage biodiversity data harvesting and quickly build indexes of the harvested data.

The HIT is capable of harvesting data from data publishers exposing their data through three protocols: Distributed Generic Information Retrieval (DiGIR) [1] , Biological Collection Access Service (BioCASe) [2], and TDWG Access Protocol for Information Retrieval (TAPIR) [3]. It can also harvest data directly from a single export, created in accordance with the new Darwin Core terms [4] as a dump in Archive format [5] using the Integrated Publishing Toolkit (IPT) [6] . By accessing all data publishers through a single tool, regardless of the protocol used, the HIT provides a convenient mechanism to coordinate and manage indexing operations and scheduling. Anybody wanting to mobilise data from several data publishers will find this increasingly beneficial as their list of publishers continues to grow.

HIT is an open source (Apache 2.0 license) Java based, customisable, multilingual web application that:
- Synchronises with the GBIF Universal Description, Discovery and Integration UDDI registry
- Harvests three types of protocols: DiGIR, BioCASe, and TAPIR; extensible to others
- Harvests from the Darwin Core Archive format
- Tracks activity with output log messages, filterable by provider or dataset
- Displays the complete list of data publishers and their datasets, filterable by provider name, dataset name, and country name, displaying statistics, etc.
- Displays the complete list of operations currently scheduled
- Allows the in-browser viewing of each individual XML request or response sent as part of the various operations
- Synchronises with one or more external databases
- Generates an index of the harvested data.

Additional features planned include:
- Role-based user management, which will allow distributed indexing management and will permit data publishers or network nodes to handle re-indexing of datasets under their domain
- Automatic scheduling of operations
- A names indexing plugin (handling checklist and other names-related data).

Resources
http://code.google.com/p/gbif-indexingtoolkit/
Source for GBIF HIT documentation, downloads, source code, bug reporting, etc.

References
[1] http://digir.net/
[2] http://www.biocase.org/products/protocols/
[3] http://www.tdwg.org/activities/tapir/
[4] http://rs.tdwg.org/dwc/index.htm
[5] http://rs.tdwg.org/dwc/terms/guides/text/index.htm
[6] http://code.google.com/p/gbif-providertoolkit/

13.31. Global Biodiversity Resources Discovery System

Vishwas Chavan, Eamonn O'Tuama, Tim Robertson, Jose Cuadra, Samy Gaiji
Global Biodiversity Information Facility

One of the major challenges for existing biodiversity informatics infrastructure is to provide users with ways that substantially increase their ability to discover and access relevant biodiversity information and data resources. Our current ability to discover distributed, isolated data and information resources is limited. This challenge is being

addressed by the Global Biodiversity Information Facility (GBIF) through the development of a Global Biodiversity Resources Discovery System (GBRDS) for registration and discovery of biodiversity information and data resources and services, as set out in its Work Programme 2009-2010.

A GBRDS should ideally be a combination of 1) a Registry of resources, their relationships and services and 2) a set of discovery services interacting with existing infrastructure such as GBIF to facilitate the discovery of biodiversity information. The most important component, the Registry would facilitate the inventory of information resources by creating a single annotated index of publishers, institutions, networks, collections (datasets), schemas and services.

GBIF hosted the GBRDS Stakeholders planning workshop in September 2009. Workshop participants envisaged GBRDS as a widely shared global system facilitating discovery of all biodiversity information resources, digital and non-digital. The GBRDS will accommodate all levels of biodiversity from genes to species to ecosystem. It is expected that the GBRDS will eventually provide means to act as a decision support tool for investment in biodiversity informatics and facilitate demand-driven, deterministic data discovery and mobilisation. It is envisaged that the GBRDS will form the core of the next generation of the biodiversity informatics infrastructure, built on the principles of distributed architecture and decentralised implementation.

13.32. SINGER, THE SYSTEM WIDE INFORMATION NETWORK ON GENETIC RESOURCES OF CGIAR

Rajesh Sood[1], Kiran Viparthi, Gautier Sarah[1], Milko Skofic[2], Elizabeth Arnaud[1]
[1] Bioversity International, [2] Bioversity international

The System-Wide Information System for Genetic Resources (SINGER) (http://www.singer.cgiar.org/) is an online catalogue of crop collections that provides inventories of conserved agricultural diversity and offers primary access for identifying and locate where samples are conserved. SINGER provides a single entry point into the crop collections' inventories of 11 centres of the Consultative Group on International Agricultural Research (CGIAR) and the Asian Vegetable Research and Development Centre (AVRDC).

SINGER enables users to access the description of conserved samples, the distribution of samples worldwide, the location of the collected sample, and information on the site where the original sample was collected. It is possible to find a set of plants meeting selected criteria by choosing a geographical area on Google Maps, acquiring climate data from WorldClim's set of global climate layers (http://www.worldclim.org/), and using LocClim, a local monthly climate estimator from the Food and Agriculture Organization (FAO) (http://www.fao.org/sd/locclim/srv/locclim.home). An online sample-ordering gateway has been added to SINGER, so that anyone can now access material via a 'shopping-cart' function and send a request to the appropriate germplasm providers.

SINGER is based on the international standard called the MultiCrop Passport data (FAO/Bioversity) which is the 'Identity card' of the sample conserved: vernacular name, taxonomy, donor name, site of collect, georeferences, etc., and the crop descriptors' lists (Bioversity International and partners). In the future, the newly developed Crop ontology will be embedded into SINGER's metadata.

SINGER is a product of the CGIAR System-wide Genetic Resources Programme (SGRP). This programme unites several CGIAR research centres in a common effort to sustain biodiversity for current and future generations. Much of SGRP's efforts to date have focused on plant genetic resources; however attention is also being given to forest, animal, and aquatic genetic resources, given the interdependence of all components of agricultural biodiversity. The CGIAR is committed to helping build a global information portal of genetic resources, and SGRP serves to bring together the CGIAR Centres in this common mission. By linking SINGER with other types of data, such as information from breeders, the future global germplasm information portal will strengthen and facilitate SINGER's role as a gateway to the world's agricultural biodiversity.

Support is acknowledged from: SGRP, World Bank,Bioversity International

13.33. Ephesis : Environment and Phenotypes Information System, a GnpIS module

Cyril Pommier, Erik Kimmel, Pierre Roumet, Delphine Steinbach, Hadi Quesneville
INRA

The investigation of the relationship between genotypes, environment, and phenotypes is one of the main long term goals of the INRA (French National Institute for Agronomical Research). To fulfil this objective, researchers need a massive amount of data and tools to create, handle, and finally analyse it. This observation has led the IGEC (Interaction Génotype Environnement Cultural) scientific network to set up three axes of development: (i) new tools and technologies

for high throughput phenotyping, (ii) analysis tools (e.g., statistical, mathematical) to extract knowledge from the data, and (iii) information systems to store and prepare data for analysis. The Ephesis project was initiated to fulfil this last need.

INRA's plant phenotyping data produced by low throughput experiments are mainly stored in flat or Excel™ files, while the huge amount of raw data produced by high throughput phenotyping platforms is inevitably stored in databases. These platforms can either be growth chambers with automated measurement of environment and phenotypes at a very fast pace, or whole field phenotyping based on aerial multispectral imaging.

The Ephesis project will produce a public information system centred on the study of genotype x environment interactions. It will increase the visibility of experiments conducted at the INRA, give value to experimental data acquired during specific temporal or spatial series, and allow the analysis of genotypes under environmental constraints. The system will ease data access and structuring, and the set up of meta-analyses. Thanks to the confidentiality built in the system, it will be possible to use Ephesis as a data exchange system within scientific networks and projects. Finally, this information system will allow a greater durability and accessibility of the data currently stored in files.

Ephesis is expected to be able to store either raw experimental data or enhanced data produced by the analysis of raw data. Most information from experimental trials will be stored and will include the details of the experimental plan, environmental data, and phenotypic characterisation at different scales (plant part, plant, microplot, parcel). Environmental information will include cultural practices (e.g., treatment, inputs) and pedological and meteorological data obtained either locally or through relevant information systems like Agroclim or Infosol.

Furthermore, the genetic resources used in the experiment will be precisely traced thus allowing fine grained data integration and interoperability with the other modules of the GnpIS[1] information system. Indeed, Ephesis stores the lot identification coming from either experimental collections or Genetic Resources Center. Furthermore, this information is directly integrated with Siregal, the INRA Plant Genetic Resource Information System. Thus, it will be possible to cross all phenotypic data stored in Ephesis with genotype information linked to Siregal's genetic resources, e.g., molecular polymorphism, markers or expression data.

To achieve its goal, the Ephesis project relies on a User Scientific Committee representing the diversity of the INRA plant community. The information system is currently under development and a first private release, reserved for the project partners, is planned for January 2010.

[1]: Plant and pest genomic and genetic information system: http://urgi.versailles.inra.fr

13.34. An Application Profile Using Darwin Core Rendered in the New Dublin Core Application Profile Framework

William E. Moen[1], Amanda K. Neill[2], Jason Best[2]
[1] College of Information, University of North Texas, [2] Botanical Research Institute of Texas

The metadata landscape for digital resources is complex and evolving. In the biological diversity arena, the Darwin Core (DwC) comprises a "body of standards… meant to provide a stable standard reference for sharing information on biological diversity" (http://rs.tdwg.org/dwc/index.htm). The Biodiversity Information Standards (TDWG) ratified the Darwin Core Standard as a TDWG standard in October 2009. DwC is based on standards developed by the Dublin Core Metadata Initiative (DCMI) (http://dublincore.org); DwC, by providing a vocabulary of terms (or elements) for biodiversity information, can be considered an extension to Dublin Core (DC). Since the inception of the DC, its flexibility and adaptability (e.g., optionality and repeatability of elements) presented challenges to interoperability. Application profiles have been developed to address these challenges by specifying the use of, and constraints on, metadata elements in specific applications. This poster describes application profile work in the Apiary Project (http://www.apiaryproject.org/) using DwC as the base schema in the context of the new DCMI application profile framework.

The concept of application profiles has evolved in the past 10 years. Heery and Patel (http://www.ariadne.ac.uk/issue25/app-profiles/) proposed profiles as a method for documenting the use, in a single application, of metadata elements from various namespaces. In 2003, a European Committee for Standardization Workshop resulted in the Dublin Core Application Profile Guidelines (ftp://ftp.cenorm.be/PUBLIC/CWAs/e-Europe/MMI-DC/cwa14855-00-2003-Nov.pdf). More recently, the DCMI proposed a Dublin Core Application Profile (DCAP) framework "for maximum interoperability and for documenting such applications for maximum reusability" (http://dublincore.org/documents/singapore-framework/). DCAPs developed using the "Singapore Framework" are intended to support metadata applications that are in "conformance with Web-architectural principles," and in particular,

serve the needs of the Semantic Web.

The Apiary Project is a collaboration of the Botanical Research Institute of Texas and the Texas Center for Digital Knowledge at the University of North Texas, funded by a National Leadership Grant from the U.S Federal Institute of Museum and Library Services. One of the project's deliverables will be a systematic digital workflow that supports the transformation of written or printed specimen label data into high-quality machine-processable format. Another will be an Apiary application profile.

In the Apiary Project, we are defining an application profile that uses DwC terms as its base schema. We are also exploring how we can render this profile in the format of the new DCAP framework (http://dublincore.org/documents/profile-guidelines/). Three mandatory component parts of a DCAP include:

- Description of the functional requirements for the application
- A domain model that identifies the entities addressed by the profile
- A Description Set Profile that provides a simple constraint language for the metadata, based on the Dublin Core Abstract Model.

DCAPs are in early stages of development in the Dublin Core community, and the Apiary Project's profile work can provide the Darwin Core community with an example of the challenges and opportunities for developing profiles aligned with the new DCAP framework. This poster will describe the work to date, issues, and potential benefits of a DCAP-based Apiary Profile.

Support is acknowledged from: U.S. Federal Institute of Museum and Library Services, National Leadership Grant LG-06-08-0079

13.35. Scanning and Passport Data Extraction from Bioversity - Collecting Mission Reports and Related Documents

Hannes Gaisberger, Imke Thormann, Milko Skofic, Lorenzo Stabile, Tom Hazekamp, Aixa del Greco, Elizabeth Arnaud
Bioversity International

Bioversity International is one of 15 centres supported by the Consultative Group on International Agricultural Research (CGIAR) alliance of independent international agricultural research centers holding important collections of plant genetic resources. Bioversity and other CGIAR centres are currently implementing the project, "Collective Action for the Rehabilitation of Global Public Goods in the CGIAR Genetic Resources System" (GPG), funded by the World Bank. The goal of the second and ongoing phase of this project (GPG2) is to ensure the quality, security, accessibility, and sustainability of the public crop collections, including the information system. These crop collections have been established thanks to donations from regional and national genebanks and also through important collections from farmers' fields and from the wild, in their native area of diversity.

Bioversity supported numerous missions from 1976-1996, which harvested a total of 221.077 samples, gathered in 560 collecting missions, documenting approximately 4300 species in 137 countries. Most of them took place in the 1970's and 1980's. These collection missions, organized in collaboration with national and international partners, targeted germplasm that was subject to severe and acute threats of genetic erosion. At the end of every mission, samples were prepared for storage and shipment to genebanks. One sub-sample was to be kept by the originating country while other sub-samples would be stored in appropriate genebanks belonging to the Network of Base Collections. This would ensure that collected germplasm was safely stored and would remain available to users worldwide. The sub-samples sent to these genebanks were routinely accompanied by descriptive elements (like taxon name, collecting numbers, provenance, etc.), the so called "Multi-crop passport descriptors".

The missions carried out by collectors generated reports where all data related to the samples were consigned. So, this wealth of information is now regarded as being a global public good.

The current information on Bioversity collecting missions is partly available in electronic form in a database, although mainly still in paper form. This collecting mission database, which is also accessible through the System-wide Information Network for Genetic Resources (SINGER, http://www.singer.cgiar.org/), contains summary information on the species collected per mission and on where the samples have been sent for long-term conservation.

Valuable additional information, such as mission reports and collecting forms, in many cases unique copies, is available in paper format, which restricts access to these sources.

Ongoing activities of the GPG2 project focus now on the completion of passport descriptors and assessment of gaps in diversity conservation due to loss of samples. Activities include:
- Scanning of Bioversity's collecting mission reports, collecting forms, and related documents
- Storing of scanned reports into a safe document repository with metadata
- Extraction of passport information
- Expansion of the Collecting Mission Database
- Integration/linkage of corresponding records/documents in the Collecting Mission Database/SINGER and
- Extraction of information and data sets for the sample loss assessment

The poster will illustrate the progress on these activities based on examples of scanned document types, their information content, database templates, and explanations of the methods and technologies used.

Support is acknowledged from: The World Bank, Bioversity

13.36. TaxoBrowser: a visual mashup for taxonomic browsing ∞

Stéphane Azard, Julie Chabalier, Amandine Sahl, Olivier Rovellotti
Natural Solutions

In the last ten years, tremendous progress has been made by the Biodiversity Informatics community. Very large online datasets are now available through the Global Biodiversity Information Facility (GBIF) and other online efforts. This has been made possible by the use of the latest technological advances in Service Oriented Architecture. The idea of combining these online data sources into a single interface to provide one page per species was first coined by Roderick Page in his iSpecies.org mashup [1][2].

In order to assist ecologists in their online information collection tasks, we developed an application that weaves data from different sources into a new service. TaxoBrowser is a mashup that combines taxonomic classification, distribution maps, images, and species descriptions in a user-friendly Web site [3].

TaxoBrowser is developed using Flex, the latest RIA (Rich Internet Application) technology from Adobe. Flex is an open source framework for building and maintaining Web applications that deploy on all major browsers [4].

The current version of TaxoBrowser uses different GBIF online services. The first one performs a search from taxonomic classifications and the second embeds distribution maps in a Web page. The user friendly navigation component guides users through taxonomic hierarchies by using a graph where each node represents a taxon that can be expanded by double-clicking. The selection of a taxon triggers an image search through the Yahoo Search Web service and displays a taxon description from Wikipedia.

[1] Page R.D. (2008), Biodiversity informatics: the challenge of linking data and the role of shared identifiers, Brief Bioinform. 2008 Sept. 9(5):345-54. [http://bib.oxfordjournals.org/cgi/content/full/9/5/345].
[2] Butler D. (2006), Mashups mix data into global service, Nature, 439(7072): 6-7.
[3] TaxoBrowser [http://biodiversitydata.blogspot.com/2009/10/taxobrowser-beta.html]
[4] Flex [http://www.adobe.com/products/flex/]

Support is acknowledged from: Natural Solutions

13.37. Composition Assistance for Multiple Existing Scientific Workflow Systems

Russell McIver, Andrew Clifford Jones, Richard White
Cardiff University

Work in the field of biodiversity informatics involves using data and resources that are often complex, of a large scale, and highly distributed in nature. Scientific Workflow Systems are tools that have been developed to support researchers' tasks of incorporating such data into complex, multi-stage experiments to answer specific research questions. For example, to perform ecological niche modeling, an electronic species checklist can provide synonyms for use in searching for observations in a species distribution database. These observations and information from a climate database might be fed into a bioclimatic modeling service, and the results fed into some visualization software. We are working to assist users in composing a coordinated workflow from selected tasks.

Although beneficial in supporting composition of such experiments, existing workflow systems have significant limitations in their support for scientists. In providing a set of resources that users can manually compose, these systems

require considerable detailed knowledge from their users, including knowledge of the specific resources required, the structure and interoperation of those components, and the way a composition must be defined within the workflow system.

To overcome this knowledge barrier, we have developed a framework that extends existing workflow systems' capabilities to assist users with workflow composition. Our framework includes an ontology, which holds information about available resources, their relationships with one another, and their suitability for use as components of tasks that might be performed within a given scientific domain. Information held about workflow resources includes basic properties such as operation names and input/output requirements. More semantically rich information is also included, e.g. the producer of the resource, the domain(s) for which that resource has been produced, specific file types and properties of the data produced and consumed by resources, and the past history of connections that have been made between resources. This metadata is used to provide users with suggestions for alterations and additions that they may wish to make to their composition.

To help users to think at an appropriate level of abstraction, components are represented within the ontology as part of a "task hierarchy." This relates composable resources to the high level goals that they perform. Hence a user can compose a workflow of abstract goals, and the system can provide suggestions for translating this into an executable workflow.

In order to operate with multiple existing scientific workflow systems, our framework includes an API (Application Programming Interface), which acts as a standard interface for our software to communicate with each of them. In this way our software can interact with the differing implementations of underlying workflow systems in a uniform manner, enabling us to provide the same level of functionality across all supported systems. This API is made feasible by the conceptual similarity of operations supported by various workflow systems. As no single existing system provides the resources that one might wish to use, the API is of particular benefit in enabling us potentially to support a wider range of workflow compositions than would be possible if limited to a single workflow system.

Further detail regarding our framework, ontology, and API can be found here:
http://biodiversity.cs.cf.ac.uk/wiki/research:mciver

Support is acknowledged from: Microsoft Research Europe

13.38. Automatic Biodiversity Literature Enhancement

David King[1], Alistair Willis[1], Anton Dil[1], David Morse[1], Chris Lyal[2], Dave Roberts[2]
[1] The Open University, [2] The Natural History Museum

The ABLE (Automatic Biodiversity Literature Enhancement) project aims to enhance access to collections of scanned, historic taxonomic documents by developing robust fuzzy matching of search terms.

Many historic taxonomic publications, such as proceedings of learned societies and institutional annual reports, are held in a few libraries only and are difficult to access. This difficulty hinders research and delivery of biological taxonomy's benefits[1].

Digitisation can improve access to the publications, but it can also introduce errors because OCR (Optical Character Recognition) technology is not perfect. The errors may mean words are not recognised by standard search techniques. At the current rate of scanning it is not practical to check the output manually. For example, two biologists took most of one year to mark up 2,500 pages[2] while the average scanning rate of the Biodiversity Heritage Library is 600,000 pages a month[3].

The first aspect of our work is to identify some of the possible errors introduced by OCR. We assume the terms that are difficult for an OCR package to recognise are those most likely to be interpreted differently by different packages. Mismatches arise between packages due to the dictionaries and font recognition training used.

By comparing the output of two OCR packages using the Needleman Wunsch algorithm[4] we often find the taxon names we want to locate. Our ongoing work is to see how far the OCR mismatches can be used to recognise taxon names in the absence of a taxon dictionary to verify them and whether it is possible to interpret the spelling differences systematically. This understanding can be used to tag the OCR text and to enhance fuzzy searching of the text so that variants are identified.

A second aspect of our work exploits the stylised form, generally using typographical cues, in which biological data is often written. We have developed pilot scripts to analyse text for these cues, such as italicised text or Latin language.

The cues complement the OCR mismatched terms so that we can tag content by type. This will support searches for data such as: taxon, location, personal names and observation date.

A third aspect of our work is to provide article level access to scanned volumes. The large scale of digitisation projects means that scanning takes place by volume. Yet scientific tradition uses the article as the basic reference unit. Using typographical cues we hope to extend the work of Lu et al[5] to detect article boundaries within volumes.

ABLE is wholly funded by the UK's Joint Information Systems Committee (JISC).

1. Godfray HCJ, 2002: Challenges for taxonomy. Nature 417:17–19
2. Sautter G, Böhm K, Agosti D, Klingenberg C, 2009: Creating digital resources from legacy documents: An experience report from the biosystematics domain. In 6th European Semantic Web Conference, LNCS Springer-Verlag, Berlin, 738-752
3. Freeland C, 2008: An evaluation of taxonomic name finding and next steps in BHL developments. P-TDWG 2008. http://www.tdwg.org/fileadmin/2008conference/slides/Freeland_05_04_BHL.ppt.
4. Needleman SB, Wunsch CD, 1970: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453
5. Lu X, Kahle B, Wang J, Giles L, 2008: A metadata generation system for scanned scientific volumes. In 8th ACM/IEEE Joint Conference on Digital Libraries, IEEE press, New York, 167–176

Support is acknowledged from: Joint Information Systems Committee

13.39. Exchanging specimen interaction data using Darwin Core

Antonio Mauro Saraiva[1], Etienne Américo Cartolano Júnior[1], Renato De Giovanni[2], Tereza Cristina Giannini[3], Pedro Luiz Pizzigatti Correa[1]

[1] Escola Politecnica da Universidade de Sao Paulo, [2] Centro de Referencia em Informacao Ambiental - CRIA, [3] Instituto de Biociencias da Universidade de Sao Paulo

Pollinators provide an important service in agriculture and conservation of ecosystems. It is estimated that the values generated by their services reach 200 billion dollars a year. However, the Food and Agriculture Organization (FAO) of the United Nations indicates a significant decline of pollinators, pointing to a "pollination crisis." The actions for conservation and sustainable use of pollinators in response to this crisis require significant support from information technology, in particular regarding the integration of diverse data sources.

Since 2007, FAO and the Organization of American States (OAS) have been supporting the development of data standards and tools to facilitate data exchange on plant-pollinator interactions. The first data standards were designed as extensions of Darwin Core (v1.4) and then published on the Darwin Core web site for broader discussion with the TDWG community. They included: 1) A generic Interaction Extension intended to represent any observed interaction between two specimens – not just between pollinators and plants, 2) A Pollination Extension including additional data specific to the pollination process, for example to report evidence of pollen or nectar removal, and 3) An Environment Measurements Extension to include environmental conditions during the observation or collecting event. The focus of these standards was only on primary data, including situations when two interacting specimens are collected and deposited in different biological collections. This initial set of data standards evolved over the last two years within the Inter-American Biodiversity Information Network – Pollinators Thematic Network (IABIN-PTN) and Webbee projects (on the biodiversity of Brazilian bees http://pollinators.iabin.net/consortium.html), resulting in a single schema still based on Darwin Core v1.4.

The latest version of the Interaction Schema treats each specimen as an independent record, so that the interaction just references the specimens by means of globally unique identifiers. Besides referencing the two specimens, the primary interaction record includes data about the observer, location, and date/time of the interaction. This approach allows interaction data to be easily exchanged, studied, and modeled, and also allows multiple interactions associated with the same specimen. A prototype of the Biodiversity Data Digitizer (BDD), a tool created to facilitate the digitization, management, and publication of data using TDWG standards, is currently using this version of the Interaction Schema to share interaction data. Next steps include making a new version of the schema based on the latest and official Darwin Core standard recently published by TDWG. This presentation will show how different types of interaction data can be represented by the Interaction Schema, including more details about the tools being developed for the IABIN-PTN network. The latest version of the Interaction Schema is available at http://groselha.pcs.usp.br/schemas/tdwg_dw_interaction.xml (best viewed in Firefox).

Support is acknowledged from: Food and Agriculture Organization of the United Nations (FAO/UN), Pollinator

Thematic Network of the Inter-American Biodiversity Information Network (IABIN-PTN), Brazilian Minister of Environment (MMA), The State of São Paulo Research Foundation (FAPESP)

13.40. The Biodiversity Data Digitizer (BDD) tool

Etienne Americo Cartolano Junior, Antonio Mauro Saraiva, Jorge Augusto Teles, Diogo Borges Krobath, Pedro Luiz Pizzigatti Correa
Escola Politecnica da Universidade de Sao Paulo

The Biodiversity Data Digitizer (BDD) is a tool designed for easy digitization, manipulation, and publication of biodiversity data. It stands out by allowing the user to manipulate its data simply and objectively, especially the data from field observations and small collections, which do not justify or demand the use of collection management software. It is based on the Darwin Core (DwC) 1.4 schema and its extensions, which were draft standards under discussion at TDWG, and which were used by the Global Biodiversity Information Facility (GBIF) at the time of BDD's initial development.

Because interactions between specimens are key to understanding important biological processes such as pollination, the BDD also allows the digitization, manipulation, and publication of specimen interaction data based on the Interaction Extension schema, discussed on the TDWG Wiki (http://wiki.tdwg.org/twiki/bin/view/DarwinCore/InteractionExtension). The BDD is a browser-based system that can be accessed remotely from a server, or locally, when installed on a personal computer. Among its main features are the registration and handling (update, delete, and search) of species occurrences (specimens) following the Darwin Core schema v1.4, and of specimen interaction data, following the Interaction Extension. The data can be displayed on maps or table records, and can be published to other systems with the TDWG Access Protocol for Information Retrieval (TAPIR) using a Tapirlink provider software. The BDD helps users improve and maintain data quality. Where relevant, users are prompted by lists of suggested entries based on authoritative databases, such as the one obtained from the Integrated Taxonomic Information System (ITIS) for taxonomic names. When the user fills in a scientific name in the BDD, and this name is in the reference list, or has already been registered, all other fields linked to it (kingdom, phylum, class, etc.) will be automatically filled in, enhancing and completing the data record and decreasing the chance of entry errors. New features, always keeping in mind data quality, are being developed, such as user access control, validation of new records by key users, upload and publication of images and their metadata, a database of bibliographic references, and the ability to load data from spreadsheets.

The BDD was an outgrowth of the Pollinator Data Digitizer (PDD), which was developed within the scope of the Pollinator Thematic Network of the Inter-American Biodiversity Information Network (IABIN-PTN). It is based on open source software, including: PHP scripts, MySQL database, and the Yii framework. For the future, it can evolve to accommodate the recent changes in DwC, now a TDWG standard. Upon public release, the BDD can be accessed at http://groselha.pcs.usp.br/bdd, and its development will be open to participation by the scientific community, whose collaboration will help achieve a more effective tool.

Support is acknowledged from: The State of São Paulo Research Foundation (FAPESP) - BioAbelha

13.41. The GBIF Integrated Publishing Toolkit

Markus Döring, Tim Robertson
GBIF

The new Java-based Global Biodiversity Information Facility (GBIF) Integrated Publishing Toolkit (IPT) http://ipt.gbif.org/ was made available for early adopter use in March 2009. This tool has features that allow for efficient and easy hosting and sharing of organism occurrence data, taxonomic and nomenclatural information, and general dataset metadata.

By focusing on these specific biodiversity data types instead of providing a generic wrapper solution as standard DiGIR (Distributed Generic Information Retrieval) or TAPIR (TDWG Access Protocol for Information Retrieval) implementations do, the software provides a rich environment, e.g., statistical summaries and specialised interfaces. Similarly, the IPT addresses the needs of data holders to serve custom, extended data schemas that go beyond simple one-to-one relationships as provided by the Simple Darwin Core (DwC) http://rs.tdwg.org/dwc/terms/simple/index.htm. For example, it allows multiple taxon identifications per occurrence record. Additionally, the toolkit provides a simple web portal for data holders, which permits easy exploration of the data and offers access to simple and easy-to-use tools for data reporting and cleansing.

The IPT continues to support the protocols and standards defined by Biodiversity Information Standards (TDWG), such

as TAPIR (ratification anticipated end 2009), the Darwin Core (updates begun to match this recently ratified standard), and the Taxon Concept Schema (TCS). Support for Access to Biological Collections Data (ABCD) is planned. The IPT offers additional interfaces such as Open Geospatial Consortium (OGC) Web Feature & Web Mapping Services (WFS & WMS). Entire datasets are available in simple text files as Darwin Core Archives to allow frequent and efficient indexing by aggregators like the GBIF Data Portal or the Ocean Biogeographic Information System (OBIS).

The poster will provide a simple-to-understand graphical illustration of the internal components of the IPT, along with clear explanations of the interfaces available, and will highlight its strategic role within the GBIF network.

Support is acknowledged from: Global Biodiversity Information Facility (GBIF)

13.42. 4D4Life Work Package 7: Serving user needs with a new system architecture for the Catalogue of Life

Richard J White, Andrew C Jones, Alex R Hardisty
Cardiff University

The objective of work package (WP) 7 of the 7th EU Framework Programme Infrastructures project "4D4Life" is to build a new, state-of-the-art, sustainable and distributed "e-infrastructure" for the Catalogue of Life. Meanwhile, in WP6, ETI Bioinformatics, Amsterdam, will maintain and improve the current system until they deploy the new infrastructure at the end of the 4D4Life project.

The new architecture will address issues of maintainability, manageability, and extensibility of the entire data handling and management system for the Catalogue of Life, and use community standards for receiving data and providing services to users. It will facilitate structured information exchange within the project networks, synthesise a globally significant resource for science, and disseminate this in an array of modern Web services and products.

In WP7, the Cardiff University team will undertake the following tasks:

7.1: Gather user and system requirements and specify the functionality required.
7.2: Design a high-level system architecture able to meet the specification.
7.3: Survey the data and metadata formats and the exchange protocols currently used or being developed in the bioinformatics community.
7.4: Create a "proof of concept" demonstration of the system design.
7.5: Implement a test-bed system with enough functionality for users to evaluate and provide feedback.
7.6: Enhance this prototype into a functional although still experimental system, and hand this system to ETI to raise to production quality in WP6 for deployment at the end of the 4D4Life project.

A critical part of this process, in collaboration with the other work packages, is to seek input from project partners, users, data providers, regional hubs, and the wider biodiversity informatics community including TDWG. We shall seek information on user scenarios, use cases, users, needs for secure access, ease of use, and the need for specific features such as alternative classifications.

We shall pay particular attention to the standards and systems already available within the community, both to avoid "re-inventing the wheel" and to ensure maximum compatibility and interoperability. We shall evaluate the potential for using or interfacing with initiatives such as the facilities and services of the Global Biodiversity Information Facility (GBIF), including its Integrated Publishing Toolkit (IPT), and tools from the European Distributed Institute of Taxonomy (EDIT) such as Scratchpads and the Cyberinfrastructure for Taxonomy. We expect that the emerging consensus surrounding the use of persistent identifiers for biodiversity data elements will encourage the emergence of a pool of compatible data sources and services, and provide the basis for mechanisms for sharing data and synchronising data sets with other data aggregators such as GBIF, the Encyclopedia of Life, and the Global Names Architecture. The Catalogue of Life is ready to play a major role in this community.

Support is acknowledged from: EC Framework 7

13.43. The Future of EOL: Phase II Implementation

Cynthia Parr
Smithsonian Institution

In Phase I (July 2007- July 2009), the Encyclopedia of Life (EOL, http://www.eol.org) built an infrastructure for aggregating content from partners with existing databases, and developed tools for generating new content. EOL now

partners with database projects, scientists building new content via LifeDesks (http://lifedesks.org), and curators who review and approve (vet) content submitted by the general public. For Phase II (Aug 2009 - Aug 2012), EOL has received renewed funding from the MacArthur and Sloan Foundations. In this phase, EOL will take two complementary organizational approaches towards growth and internationalization: content themes and regional Encyclopedia of Life efforts. In addition, EOL is pursuing several technical solutions that foster integration and sharing.

Our Marine Theme is the first of a number of themes that will guide our efforts to build and foster engagement with the Encyclopedia of Life. Between 2009 and 2013, we aim to provide information on 90% of the world's known marine species, over 200,000 vetted pages. We partner with organizations like the Census of Marine Life and the World Register of Marine Species to reach this goal.

EOL is also working with international institutions to establish region-specific Encyclopedias. These can use EOL branding, serve content in their own languages, re-purpose EOL software for their communities, and share local content with the rest of the world through the EOL main site. Several regional EOLs are in active development, including Australia (http://www.ala.org.au/), Netherlands (http://www.nederlandsesoorten.nl), China, and South Africa. Several others are ready to begin.

With respect to technical solutions, EOL is designing and developing the Global Names Architecture (http://www.globalnames.org) in collaboration with Global Biodiversity Information Facility (GBIF), ZooBank, Pan-European Species directories Infrastructure (PESI), and several nomenclators. Also, technical teams have added a species pages module to GBIF's Integrated Publishing Toolkit (http://code.google.com/p/gbif-providertoolkit/). GBIF and EOL partners will be able to use the Integrated Publishing Toolkit to establish a data flow with EOL. Finally, EOL is leading refinement, implementation strategies, and ratification of the Species Profile Model TDWG standard.

Support is acknowledged from: MacArthur Foundation; Sloan Foundation

13.44. Building a scalable, open source storage solution for biodiversity data

Phil Cryer, Anthony Goddard
Biodiversity Heritage Library (BHL)

The Biodiversity Heritage Library (BHL), like many other projects within biodiversity informatics, maintains terabytes of data that must be safeguarded against loss. Further, a scalable and resilient infrastructure is required to enable continuous data interoperability, as BHL provides unique services to its community of users. This volume of data and associated availability requirements present significant challenges to a distributed organization like BHL, not only in funding capital equipment purchases, but also in ongoing system administration and maintenance. A new standardized system is required to bring new opportunities to collaborate on distributed services and processing across what will be geographically dispersed nodes. Such services and processing include taxon name finding, indexes, Globally Unique Identifier (GUID) and The Life Sciences Identifier (LSID) services, distributed text mining, names reconciliation and other computationally intensive tasks, or tasks with high availability requirements.

After reviewing proprietary enterprise offerings, BHL engineers identified a solution made from open source, standards-based, scalable technologies capable of storing 50 terabytes of mirrored data at launch, with dramatic future growth anticipated. Together they supply an open repository that provides redundancy, accessibility, security, and a future-proof archival path. Redundancy is provided via GlusterFS, which is an open source, cluster file system that allows disparate servers to pool their storage across many systems, thus becoming capable of scaling to several petabytes. The architecture of GlusterFS makes it far simpler to deploy, maintain, and expand than other options. By using a common infrastructure, projects can more easily create multiple instances of their data stores, even sharing them among multiple projects. With these redundant clusters in place, BHL intends on investigating load balancing by geolocation as a method to improve the site's performance. Previously, lack of storage space required BHL to serve content on demand, calling data housed on servers outside of its control, which in turn made it difficult to ensure the integrity of the data, from both a delivery and security point of view. With a clustered solution, BHL has full control over its data store, mitigating these issues. As a layer above the GlusterFS filesystem, the open source Fedora Commons digital object repository system provides standards-based archival data management and storage. This solution provides a service layer enabling applications to search content, harvest metadata and files via The Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH), and view data in the Resource Description Framework (RDF) triple stores format. These added services increase access options to BHL data, and define a community-supported digital preservation methodology.

This solution has been designed to be a blueprint for other biodiversity projects. By using easily obtainable hardware and software, it presents a low cost option for institutions and provides them with easy access to clustered processing of their data. The scalable nature of the solution makes its use applicable across all areas of biodiversity informatics and provides

a roadmap for future storage of biodiversity data.

Support is acknowledged from: Biodiversity Heritage Library (BHL)

13.45. The Global Names Architecture: an integrated and federated approach to enabling discovery and access to biodiversity information.

David Remsen
GBIF

Biodiversity informatics focuses on mobilising, accessing, and synthesising information about individual species of organisms and their temporal and spatial relationships with other species, and within human and natural environments. All of this information, past and present, is tied to our understanding of biodiversity through the use of a scientific name. Biodiversity science, however, lacks a complete list of scientific names of organisms or a comprehensive framework for describing the complex ways these names are used to represent the relationships among taxa. This is problematic because names of species are neither stable nor unique and a single species concept may be known by more than one name while the same name may refer to more than one taxon or concept. Taxonomic revisions may lump previously distinct species into one or split a species with the result that a name alone is insufficient to convey these different senses. This may severely impact the ability to search, access, and effectively synthesise biodiversity data.

The Global Biodiversity Information Facility (GBIF) is contributing to the development of a Global Names Architecture (GNA) that provides a comprehensive framework for discovering and resolving a complete list of all names of organisms. It provides the means to create a virtual and dynamic catalogue of all scientific and common names. It enables these names to be reconciled to a distributed array of nomenclatural databases that tie all names to their originating publications. It enables the discovery of an array of taxonomic databases that collectively tie these names to our current and past understanding of species definitions, their classification and synonymy.

Most importantly, the GNA provides a framework for collating and coordinating the use of these data as informatics tools and services. These tools and services are critical for ensuring effective access to any collection of biodiversity data. They ensure that a search of information on a species retrieves all relevant data, no matter if it is labeled with a name that is current, invalid or even misspelled. It supports the use of taxonomic identifiers that unambiguously tie a particular data resource to a specific taxonomic concept, enabling a more accurate linkage than a name alone can provide.

In 2009-2010 GBIF is contributing architectural components to the GNA that include a refined taxonomic exchange schema and a simplified data exchange framework. It includes the Integrated Publishing Toolkit that supports the publishing of taxonomic and nomenclatural data and the Global Biodiversity Resources Discovery System that will enable a global registration mechanism for multiple taxonomic databases. GBIF will utilise this architecture to catalog an array or taxonomic resources and provide a common discovery framework for enabling users to collectively access and utilize them. In addition, GBIF is developing tools and services that use the collective names catalogue and services to develop new and novel data discovery and access applications. This includes services for mapping content to published taxon concepts in a generalized manner that can operate across multiple published resources. It includes services for integrating search, retrieval, and data browsing with published taxonomic resources to increase the precision and recall of data access methods in any collection of biodiversity information.

13.46. Computer Tools for Descriptive Data Management for the Taxonomist

Ôna Maiocco[1], Régine Vignes-Lebbe[2]
[1] EDIT, [2] Université Pierre et marie Curie

A significant number of tools for managing biological descriptions and creating identification keys are available: DELTA (Descriptive Language for Taxonomy, http://delta-intkey.com/), Lucid (http://www.lucidcentral.com/), and Xper² (http://lis-upmc.snv.jussieu.fr/lis/?q=en/resources/softwares/xper2) software are good examples of what exists today. These tools are designed to create standardized descriptions, and provide an easy way of comparing and processing data, e.g. construction of diagnoses and keys in a classical dichotomous format or in interactive identification systems, production of taxonomic descriptions in structured or natural language, and import/export for phylogenetic and other studies.

One of the EDIT (European Distributed Institute of Taxonomy) project achievements is a cyber-platform (http://dev.e-taxonomy.eu/platform/) providing a large range of computer tools for taxonomists (http://www.bdtracker.net/). These tools are designed to assist the taxonomist from fieldwork to publication of results, including the management of descriptive data, which plays a key role in the taxonomic revision process. However, a common observation is that the

existing software tools, particularly those dedicated to generating or analyzing descriptive data, are often not used or well known by taxonomists.

This poster summarizes the capabilities of the main existing tools with the aim of helping the taxonomist to determine which one best fits his needs. Moreover, we propose concise guidelines based on several use-cases, showing how these tools can be widely integrated into the daily workflow of taxonomists, thanks to current standards, including the TDWG (Taxonomic Database Working Group) SDD (Structured Descriptive Data).

13.47. PESI: A European web portal for all species in Europe

Ward Appeltans[1], Bart Vanhoorne[1], Marie-Line Villers[1], Leen Vandepitte[1], Wim Decock[1], Francisco Hernandez[1], Marc Geoffroy[2], Anton Guentsch[2], Walter Berendsohn[2], Mark Costello[3], Yde de Jong[4]
[1] Flanders Marine Institute, [2] Free University of Berlin; Botanischer Garten und Botanisches Museum Berlin-Dahlem, [3] University of Auckland; Leigh Marine Laboratory, [4] Universiteit van Amsterdam; Faculteit der Natuurwetenschappen, Wiskunde en Informatica; Zoologisch Museum Amsterdam

PESI (Pan-European Species-directories Infrastructure), a 3-year project started in May 2008, is funded by the European Union under the Framework 7 Capacities Work Programme: Research Infrastructure.

The goal of PESI is to build a single web portal for all species in Europe. The portal will integrate the taxonomic information from the three major European check lists: Fauna Europaea (FaEu), Euro+Med plantbase (E+M) and the European Register of Marine Species (ERMS). Technically, database integration will be achieved using software developed for the European Distributed Institute of Taxonomy (EDIT) Platform for Cybertaxonomy. The web portal will also create links with several nomenclators and taxonomic databases like AlgaeBase, the International Plant Names Index (IPNI), Index Fungorum, and the numerous national species checklists, which will help in making the European catalogue complete and highly scrutinized.

The goal of PESI is not only to bring together the taxonomic information about species, but also to gather images and information on occurrences of terrestrial and marine species at regional and national levels, and to document local common names, and red listed species, with their national protection status. The PESI web portal will also make available literature references of the main national publications (e.g. local field guides and monographs) and will provide information on local species experts.

PESI will form a European hub within global species initiatives, such as the Global Names Architecture (GNA) and Catalog of Life (CoL), and it will also help other EU member states by providing a standardized and authoritative reference of European species names, which can serve as a taxonomic backbone for national biodiversity initiatives.

Support is acknowledged from: European Union under the Framework 7 Capacities Work Programme: Research Infrastructure

13.48. The Biofinity Project: Transforming Biodiversity Research

Mary Liz Jameson[1], Federico Ocampo[2], Daniel R. Clark[1], Matthew R. Moore[1]
[1] Wichita State University, [2] Instituto de Investigaciones de Zonas Aridas, Mendoza

The Biofinity Project (http://biofinity.unl.edu) is a free web-based repository for biodiversity data and tools designed to support research in the biological sciences. The Biofinity Project federates genomics and biodiversity information and provides support for inclusion of external data regardless of format. The repository allows scientists to access, analyze, share, and publish on biological data from a myriad of available resources. Tools provided by The Biofinity Project, such as mobile iPhone™ data-integration and geo-tagging, RSS (Really Simple Syndication) feeds for specimen identification and verification, and web applications for niche modeling, BLAST (Basic Local Alignment Search Tool), and phylogenetic analyses will further advance biodiversity research.

The Biofinity Project unifies genomics and biodiversity data, thereby empowering investigation of patterns that can lead to a greater understanding of broad-scale, widely applicable, and emergent biological properties. The Biofinity Project provides full access to enormous, publicly available biodiversity data at the Global Biodiversity Information Facility (GBIF) and genomics data at the National Center for Biotechnology Information (NCBI). In addition, it provides upload and unification of independent databases that are in different formats and based on different software programs. Searching, browsing, and uploading of data to an external database is possible via a web browser or a mobile interface such as the iPhone™ or iPod™ Touch. The in-field application allows instant specimen mapping, geo-tagging, video capture, and direct upload to an external database by using The Biofinity Project's API (Application Programmer Interface).

The Biofinity Project provides web access to bioinformatics tools such as: 1) GoogleMaps-based mapping tool, which allows for study of specimens distribution including climate, topology, and precipitation overlays; 2) ecological niche modeling using DesktopGarp; 3) CLUSTALW for multiple sequence alignment of DNA or proteins; and 4) "My Lab", a feature that that allows research groups to create an online laboratory database and establish their own user accounts for collaborative research. We showcase our on-going research on scarab beetle biodiversity that uses some of these tools.

Support is acknowledged from: NSF DBI-0743783 to Scott, Henninger, Jameson, Moriyama, and Soh; NSF-DEB 0716899 to Ratcliffe and Cave; and IADIZA - CONICET to Ocampo.

13.49. Crop science and breeding: contributions in the field of bioinformatics by the Generation Challenge Programme, an international crop breeding research programme

E van Strien
CGN WUR

The Generation Challenge Programme (GCP) was initiated in 2004, as a ten year programme by the Consultative Group on International Agricultural Research (CGIAR). The GCP mission is to use plant genetic diversity, advanced genomic science, and comparative biology to develop tools and technologies that help plant breeders in the developing world produce better crop varieties for resource-poor farmers. Its focus is on drought tolerance using modern breeding techniques that make use of molecular markers to incorporate these traits in local varieties.

This research, performed in many places, generates enormous amounts of information that needs to be shared amongst partners and with the researchers. The way this information is managed, analysed, and made accessible, determines to a large extent the information's value to the recipient.

A large component of the GCP subprogram 4 (SP4), focuses on bioinformatics and biometrics, and aims to create a network, intended for crop researchers and breeders, which integrates information on genetic resources, genomics, and crop improvement. To this end, SP4 is improving already existing and newly developing analytical tools. SP4 addresses methodologies for linking gene discovery with genetic resource characterisation and crop evaluation data.

The SP4 products encompass both software and services, http://www.generationcp.org/bioinformatics.php. Examples of their freely accessible products are Dayhoff, a comparative stress gene catalogue, and DARwin, used for dissimilarity analysis and representation. The GCP High Performance Computing (HPC) grid offers computing facilities with programmes and services for genetic and sequence analysis, and statistical data analysis.

SP4 products also feature access to high quality data, analytical tools and facilities, and support to scientists.

The Central Registry provides access to data generated by GCP plant and breeding research. High data quality is established by quality assurrance, management and standarization procedures, such as Laboratory Management Information Systems and GCP-approved data templates. The analytical tools and facilities are present for the fields of genomics, breeding and statistical analysis, and genetic diversity studies, as well as for fields more interesting to informatics developers: domain modeling, software architecture, and crop database architecture.

Finally, support to GCP scientists is provided by helpdesks, courses and training materials, and information from workshops and meetings organized by GCP over the past five years. All is being made publically available.

Support is acknowledged from: Generation Challenge Program (GCP)

13.50. Specify Collections Software: Cross-Platform, Open-Source, Collaborative, Robust, Localizable, Multidisciplinary, Mature, Portable, and Free.

Jim Beach
Biodiversity Institute, Univ of Kansas

Specify is a biological collection management platform designed to process specimen voucher and species observational data. Released in April 2009, with more than 12 software developer years of design and implementation, and over $2 million of investment, Specify 6 is an open software platform for international collaboration.

Specify is designed for integration with internet services and it is extendible through plug-ins. Written in Java, it runs

identically on the three common desktop environments: Windows, Mac OS X, and Linux, and it is open sourced and free. The Specify application is designed to maximize user interface flexibility for customization and localization. A single schema accommodates specimen data for all biological collection disciplines and the user interface can be localized into any Unicode language. Specify integrates with several web services, including the Global Biodiversity Information Facility's Internet Publishing Toolkit, and with the Specify WorkBench, a portable application on a USB flash drive, designed for offline and field data entry. Downloadable installation packages for all three desktop flavors are available from the web site (www.specifysoftware.org).

Specify 6 has an intuitive user interface aimed at streamlining routine collections data tasks, while preparing and validating collection information for biodiversity community networks. It includes capabilities for using record sets as subsets of the complete catalog for various types of processing, such as: online georeferencing with Tulane University's GEOLocate, locality mapping with NASA's World Wind, and importing and exporting records. Specify 6's scope has been enlarged to provide support for paleontological data, field notebooks, attachments, GUIDs (Global Unique IDentifiers), hierarchical storage locations, data entry and uploads through the Specify WorkBench and Excel, collecting trip data, repository agreements, accessions, collection object conservation treatments and containers, and additional data types. Specify's data entry forms are customizable to match institutional preferences, and it can print custom labels and report to exacting requirements.
Specify is the most widely-used computing platform in U.S. herbaria and museums. It is the primary production database system for 189 U.S. collections across 90 institutions. Outside of the U.S., 69 collections at 41 institutions in 17 countries use Specify to manage their specimen holdings information.

Innovation and new development is ongoing; new capability objectives for 2010-2014 include:

--Providing network-based collection management tools for cross-institutional workflows in support of regional, national, and international survey and inventory campaigns.

--Increasing the connectedness and vitality of cross-institutional collaborative work, through fine-grain messaging of current awareness alerts for collection-related events, project status, and computerization project milestone completion.

--Promoting focus, feedback, and incentives for collection researchers to address computerization progress, and to stimulate alert-driven actions for annotation, geo-referencing, and duplicate specimen discovery.

We look forward to initiating new software development collaborations to extend Specify and to bring specimen and observational data to broader computational research and networking initiatives in the environmental sciences.

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

13.51. At the frontline of publishing in systematic zoology: ZooKeys ∞

Lyubomir Dimitrov Penev[1], Terry Erwin[2], Jeremy Miller[3]
[1] Pensoft/ZooKeys, [2] Smithsonian Institution, [3] Naturalis

ZooKeys is open-access, peer-reviewed, rapidly ZooKeys is open-access, peer-reviewed, rapidly disseminated, online and print journal launched in July 2008 by Pensoft Publishers to accelerate research and free information exchange in taxonomy, phylogeny, and biogeography (www.pensoftonline.net/zookeys). The journal's objective is to respond to the major challenges in publishing and dissemination of scientific information at the beginning of 21st Century.

A universal electronic register of animal names, ZooBank, was proposed in systematic zoology. New amendments to the forthcoming 5th Edition of the International Code of Zoological Nomenclature towards recognition of electronic publications are currently in discussion. Open access publishing is well-established amongst scientists, institutions and funding agencies and government. In less than a year, Zookeys was accepted for ISI coverage and became official partner of Encyclopedia of Life (EOL) and Global Biodiversity Information Facilities (GBIF).

ZooKeys has published 26 issues with more than 4000 pages of valuable taxonomic information. The journal accepts taxonomic revisions of extant (or "recent") and fossil animal groups; checklists and catalogues; phylogenetic and evolutionary analyses; papers in descriptive and/or historical biogeography; methodology papers; data mining and literature surveys; monographs, conspecti, atlases; collections of papers, Festschrift volumes, and conference proceedings. ZooKeys is committed to develop and implement innovative publishing methods and semantic enhancements to its papers, such as: (1) All primary biodiversity data underlying a taxonomic monograph can be published as a dataset under a separate DOI within the paper; (2) The occurrence dataset can be uploaded to GBIF

simultaneously with the publication; (3) The occurrence dataset can be a published also as a KML file to provide an interactive experience in Google Earth; (4) Data matrices and primary data files for interactive keys (e.g., Lucid, Intkey, MX, and others) can be published as supplementary files to facilitate future use and reuse of the data; (5) All new taxa are being registered at ZooBank during the publication process (mandatory); (6) All new taxa are provided to EOL through XML markup on the day of publication (mandatory); (7) All new taxa and images are submitted to Wikispecies and Wikimedia Commons on the day of publication.

In the near future, ZooKeys plans to (1) implement a pre-submission XML mark up; (2) provide automated generation of manuscripts from Scratchpads; (3) extend and continuously improve the range of semantic enhancements to taxonomic papers; (4) create tools to maximum automated dissemination of contents to aggregators, indexing and bibliographic services and archives; (5) transfer the whole set of innovative publishing methods into the domain of plant diversity science through launching a new counter journal, PhytoKeys.

13.52. Repatriating the Botany of Tropical Africa

Henry R. Engledow, Quentin J. Groom, Piet Stoffelen, Alain Empain, Sofie De Smedt, Steven Dessein, Petra De Block, Elmar Robbrecht
National Botanic Garden of Belgium

The National Botanic Garden of Belgium has launched its online Virtual Herbarium (www.br.fgov.be). This is consistent with our policy of repatriating biodiversity knowledge to their countries of origin. Conservationists, policy-makers, and scientists worldwide need this information for their work. The project is of particular interest for central Africa, as the Garden has been focusing on the Democratic Republic of the Congo, Rwanda, and Burundi for more than a century.

The herbarium is being catalogued using BG-BASE software (www.bg-base.com), an information management system, specifically designed for use in botanical gardens. At present, the data is exported from BG-BASE and imported into a PostgreSQL database on a monthly basis, from where the data are accessible via dedicated web pages.

Images of herbarium specimens, are scanned at high resolution and saved as lossless TIFF (Tagged Image File Format) files. The original images are archived for future reference, while a copy is converted into a "zoomable" user interface (ZUI) to scale the image for more or less detail, using the open source conversion program, ZoomifyImage (sourceforge.net/projects/zoomifyimage/). These converted images are then presented online using a Flash-based viewer (www.zoomify.com).

Currently, researchers can browse the herbarium website with access to over 34,000 zoomable images of type specimens and label data for approximately 300,000 specimens.

Support is acknowledged from: National Botanic Garden of Belgium

13.53. Lifemapper: Finding the Good Life

Aimee M. Stewart, James H. Beach, C.J. Grady, David A. Vieglais
University of Kansas

Lifemapper2 (www.lifemapper.org) is an archive of biodiversity geospatial data and a set of cluster-based computational tools provided to users through web services. Lifemapper synthesizes the known distribution information of terrestrial plants and animals, and predicts future species distributions based on various climate scenarios. It is built around the openModeller ecological niche modeling (ENM) platform and uses a Service Oriented Architecture (SOA) to provide data and analysis. The Global Biodiversity Information Facility (GBIF) provides a monthly cache of their specimen occurrence database, which contains specimen records from over 283 data providers. With each new GBIF cache, models are recalculated for species with new occurence data.
Data products offered through web services include specimen records compiled from GBIF, models and projections created with the openModeller niche modeling library, and select environmental data. Predicted climate data from the the International Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) is now being converted for Lifemapper data services and ENM analysis, as well as for use in openModeller or other modeling applications directly.

Computational tools include ENM services, which allow users to request ecological niche modeling experiments through the Lifemapper web application or to access the web services programmatically. The user may define modeling parameters and either upload point data or select specimen points from the Lifemapper copy of the GBIF cache . For environmental layer inputs, Lifemapper offers observed climate and multiple IPCC scenarios of predicted future climate data, and soon will allow user upload of specialized environmental layers.

Additional tools to compute landscape ecological measurements are in the final stages of development. These tools calculate landscape metrics such as patch, core, edge, and shape measurements. Others quantify and describe dispersal models and timeseries changes. All landscape ecological functions are built to measure individual species and landscapes or summarize ensembles.

A related project integrates Lifemapper and the widely used desktop application SAM, Spatial Analysis in Macroecology (www.ecoevol.ufg.br/sam/), designed for spatial statistical inference and applications in surface pattern spatial analysis. The Lifemapper-SAM (LM-SAM) project has been funded by the National Science Foundation and will enable creation and analysis of multispecies macroecological grids, informed by data available in the ENM community.

As part of the LM-SAM project, a Lifemapper-Specify Plug-in will support species occurrence data from any data source, and will be able to follow specific work flows for any Lifemapper tools from data query and assembly to analysis, then catalog the results. The Specify client will hide the complexity and simplify data management associated with macroecological grid construction and analysis, change and dispersal analysis, landscape metrics, and ecological niche modeling workflows.

A collaborative project with the University of Michigan, ChangeThinking, adapts the Lifemapper portal to bring ecological niche modeling and climate change impact to science students in grades 9-12. When it debuts, ChangeThinking will provide interactive modeling and link food web models to ENM results to convey the complexity of food web changes as a result of species' reactions to climate change.

Support is acknowledged from: National Science Foundation

15. Contributed Abstracts and Papers

15.1. Herbarium Digitisation at Royal Botanic Gardens, Kew

Kathryn Beck, Sarah Phillips
RBG Kew

RBG Kew is working on a digitisation project, in collaboration with over 50 Partner Institutions worldwide. It is funded by the Andrew Mellon Foundation and which aims to produce high quality images and data from herbarium type specimens. To date we have digitised 70,000 African types and 66,000 Latin American types, over a period of six years. As well as digitising our own type specimens, the team at Kew also trains staff from institutions from Latin America and Europe. We also produce and supply the scanning equipment (Herb Scans) for institutions worldwide.

Managing such a high output of digitised herbarium specimens, demands a careful balance between individuals working to specific standards and achieving certain levels of output balanced with the team's communication and co-operation. The project must be aware of technological and operational advances that related to issues such as data and image capture, storage and management. Some recent developments that we have made in the digitisation team at Kew include the use of a digital camera for image capture of delicate and bulky specimens, and the development and use of a scratchpad for the exchange of knowledge and expertise both within the team and between other teams.

We will discuss the digitisation workflow at Kew, including the quality control process that we have in place, which ensures that we produce images and data to a high standard.

Support is acknowledged from: The Andrew Mellon Foundation

15.2. A data standard to integrate Farmers Knowledge and Science ∞

Adriana Alercia, Frederick van Oudenhoven, Pablo Eyzaguirre
Bioversity International

Bioversity International and its partners have developed farmers' descriptors to provide a standard format for the gathering, storage and exchange of farmers' knowledge of plants. It aims to capture key characteristics, uses and values of cultivated and wild plants as described by farmers and other people in farming communities. Wild and weedy plants are covered by this standard since they play a significant role in farming communities, being useful from a socio-economic and ecological standpoint.
This standard is a first attempt to document farmers' community descriptions about planting material, which complement the agro-morphological and agronomic crop descriptors such as those developed by Bioversity and/or UPOV (International Union for the Protection of new Varieties of Plants). This tool is tailored to farmers' knowledge and

integrates traditional crop descriptors and farmers' knowledge about plants and their cultural aspects.
This effort combines a documentation system as used in controlled environments (genebanks, breeding institutes) with an approach that involves people and their knowledge 'in the field'. It is the result of many years of review of fieldwork by scientists and field practitioners, and constitutes an important tool for integrating biology and traditional knowledge. Although the list is primarily targeted at the plant genetic resources community in order to increase the range of knowledge recorded during plant collection, its widespread use by others, including farming communities and organizations, is encouraged.
The goal of this initiative is to create a lingua franca to capture and share information amongst farmers and scientists and to integrate biology and traditional knowledge. Some of the expected benefits derived from the use of this standard are uniformity and consistency of documentation, increased visibility of farmers' role in crop diversity and the validity of farmers' knowledge; ability to work across geographic and knowledge boundaries; development of databases and networks, which in turn will contribute to create a platform to equitably share knowledge about plant diversity.

Support is acknowledged from: The Christensen Fund

15.3. DarwinCore Germplasm Extension and deployment in the GBIF infrastructure

Dag Endresen[1], Samy Gaiji[2], Tim Robertson[2]
[1] Nordic Genetic Resources Center (NordGen), [2] Global Biodiversity Information Facility (GBIF)

DarwinCore is designed around a set of general terms applicable for most biodiversity datasets. DarwinCore also implements a model of extensions to the core terms, designed to include terms of more specific utility in thematic domains. The DarwinCore Germplasm Extension has been developed to include the additional terms required to describe germplasm samples maintained by genebanks worldwide. The most widely used terms to describe germplasm samples are included in the Multi-Crop Passport Descriptors (MCPD) developed and published in December 2001 by Bioversity International (formerly IPGRI) and the Food and Agriculture Organization of the United Nations (FAO). In 2005 the MCPD terms were integrated to the ABCD standard (Access to Biological Collections Data). This work paved the way for the implementation of the BioCASE data publishing toolkit in the plant genetic resources community, and the improved sharing of germplasm datasets within the community as well as with the GBIF Network. The DarwinCore Germplasm Extension includes in a similar manner the missing terms from the MCPD standard. A few additional terms were included for description of the germplasm in relation to the new international Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA) and other regulatory mechanisms. You will also find included the new terms for exchange of germplasm trait measurements developed in Europe for the ECPGR network (European Cooperative Programme for Plant Genetic Resources).

The GBIF Integrated Publishing Toolkit (IPT) was released in March 2009. In a similar manner as for the DarwinCore extensions, the IPT has the great advantage to be extended with more domain specific application schemas or extensions. When publishing a dataset with the IPT, the publisher can select terms from the general DarwinCore as well as from the extensions. The GBIF Harvesting and Indexing Toolkit (HIT) to be released in October 2009 will further make the indexing of distributed and heterogeneous datasets easier for Network managers. IPT also introduces the new DarwinCore archive. The DarwinCore archive will significantly speed up the indexing of datasets by a central portal like the GBIF portal or a thematic portal like for example the new global germplasm portal (Global-ALIS). Both the IPT and the HIT are able to synchronize with the GBIF Global Biodiversity Resources Discovery System (GBRDS). Thus datasets, data sharing protocols or extensions registered at the GBRDS can be easily discovered and accessed by a distributed thematic or regional biodiversity information network or for example by a more specific data analysis tool.

DarwinCore: http://rs.tdwg.org/dwc/terms/index.htm
Germplasm Extension: http://rs.nordgen.org/dwc/
GBIF IPT: http://code.google.com/p/gbif-providertoolkit/
GBIF HIT: http://code.google.com/p/gbif-indexingtoolkit/
GBIF Portal: http://data.gbif.org/ and http://www.gbif.org
Prototype Global ALIS: http://www.global-alis.org/

15.4. INSPIREing Europe

Kathi Schleidt
umweltbundesamt

INSPIRE, the Spatial Data Infrastructure for the Environment in Europe, is a European legal directive requiring all data relevant to the environment and held by public bodies to be made available via Web Services. In accordance with the timeline specified in the directive, metadata for selected areas will be coming online in 2010, with dates for metadata and data on various concepts being made available being spread out over the next decade and completion planned for 2019.

In order to maintain complicity with standards, the relevant ISO and OGC standards have been selected for the implementation of these services. On the Metadata level, the ISO 19115 Suite of standards are to be used. On the data level, current plans call for the use of GML for encoding of the data, and the use of Web Map Service (WMS) and Web Feature Service (WFS) for the provision of data services. In addition, further OGC standards, such as the Observations & Measurements Schema (O&M) and the Sensor Web Enablement Suite (SWE) are currently being tested for suitability. While this does not directly affect TDWG, it does have the following ramifications:

• A vast amount of environmentally relevant data will be made available in Europe over the next 10 years, which will be valuable to augment the biodiversity data provided by theTDWG community. This includes concepts such as (subset):

o Hydrography
o Protected sites
o Land cover
o Bio-geographical regions
o Habitats and biotopes
o Species distribution
o Soil
o Population distribution — demography

• A great deal of effort is being placed into the extension of these existing standards, as well as the technology for discovery, access and harvesting. While the OGC palette of standards does not currently sufficiently cover the biodiversity domain, the framework is quite suitable for extension into this domain. As many of the concepts required for TDWG standards (locations, responsible parties & roles, observations) are already covered and maintained by OGC, this would allow TDWG to focus on their area of expertise, while gaining access to the resources provided by a much wider community.

To my view, whatever route TDWG chooses to take in the future, this development should be taken into account, be it in the joining of forces with OGC for the definition of sound biodiversity components for the existing standards, or be in only in the integration of the data holdings being made available by this initiative for further analysis purposes.

INSPIRE: http://inspire.jrc.ec.europa.eu/
OGC: http:// www.opengeospatial.org/

16. Working Sessions

16.1. The jKey wiki key player and builder ∞

Stephan Opitz, Gregor Hagedorn
Federal Biological Research Center (JKI)

jKey is a combination of standard mediawiki templates with javascript code to provide a simple solution for a simple task: publish plain traditional dichotomous or polytomous keys as documents on the web, edit and illustrate them collaboratively on the web, and finally either view them in printable overview mode or play them interactively (step-by-step, with revisionable/confirmable history and uncertainty flagging). jKey uses a mediawiki installation as its basis. Keys can be edited in the normal wiki-editor as well as in a custom, form-based editor. All javascript code is GPL licensed. It is available and editable directly on the Wiki.

See http://www.keytonature.eu/wiki/JKey_Player for further information.

16.2. Crop Ontology: a Reference Controlled Vocabulary on Crop Trait Information ∞

Elizabeth Arnaud[1], Rosemary Shrestha[2], Martin Senger[3], Mauleon Ramil[3], Milko Skofic[1]
[1] Bioversity International, [2] CIMMYT, [3] International Rice Research institute (IRRI)

Within the Consultative Group on International Agricultural Research (CGIAR; www.cgiar.org) consortium, the volume of agriculture-related information is increasing enormously. The CGIAR has accumulated historic crop data that are related to phenotype, breeding, germplasm, pedigree, traits, etc., for the past six or seven decades. The Generation Challenge Programme (GCP; http://www.generationcp.org) central data repository system (http://gcpcr.grinfo.net/) also contains data related to both genotype and phenotype. To facilitate smooth data exchange across databases and data annotation, controlled vocabulary system is urgently needed. Moreover, most terms associated with phenotypes are not well covered by current existing ontologies. Therefore, GCP deployed crop ontology (CO) which characterizes the computational architecture of a knowledge-based system. CO intends to globally and uniformly identify ontology terms

of nine ontology domains such as the GCP Domain Model, the General Germplasm and Passport, the Taxonomic, the Plant Anatomy and Development, the Phenotype and Trait, the Structural and Functional Genomics, the Location and Environment, the General Science and other sub-domain or site specific ontology domain.
At present, CO focuses on developing crop-specific trait ontology for chickpea, maize, Musa, potato, rice, sorghum and wheat crops. The GCP Crop Ontology browser is available at http://koios.generationcp.org/ontology-lookup/ for searching ontology terms or specific ontology hierarchy present in CO. To facilitate development of the GCP ontology, a site for curators and collaborators the Pantheon project web site (look on http://pantheon.generationcp.org under GCP Semantics ... GCP Ontology menu item) is available. Complementing the Pantheon website is a project created for GCP ontology on the CropForge software project management site (http://cropforge.org/projects/gcpontology/). The Pantheon site is mainly targeting software and ontology developers. To cater to scientific end users of the GCP ontology, an additional site is being established at http://mcclintock.generationcp.org/. This site uses the GCP ontology inventory page to index summary pages for ontology with links to available ontology files and link to other sites that are related to external ontology. All information described on the website is freely available and all ontology flat files in OBO-format are also available to download. In the near future, the ontology project team will commission a query interface which will enable researchers to query a comprehensive CO database using the keywords that are related to traits, plant structure, growth stages and molecular functions. Other queries will direct users to associated GCP phenotype, genotype and other related crop datasets. Crop Ontology will use text mining tools such as MSWord-2007 Ontology add-in and Terminizer (http://terminizer.org/) to capture ontology terms in documents and publications. Collaboration with Gramene Trait Ontology (TO) and Plant Ontology (PO) is continued by submitting CO terms as new terms to these databases. Further collaboration is being processed with AGROVOC, Thai Rice Ontology, Plant ontology consortium, SGN genomic network (Potato trait ontology), and Maize GDB.

Support is acknowledged from: Generation Challenge Programme, Bioversity International,. CIMMYT, IRRI, ICRISAT, CIP,

16.3. Fine-Grained Semantic Markup of Descriptive Data for Knowledge Applications in Biodiversity Domains

Hong Cui[1], James Macklin[2], Chunshui Yu[1], Partha Pratim Sanyal[1]
[1] University of Arizona, [2] Harvard University

This project will develop a set of unsupervised machine learning algorithms & create a suite of domain-independent, high-throughput software that marks textual descriptive data of taxonomic treatments to support various knowledge applications, including producing character matrices & identification keys for different taxon groups. We demonstrate a recent version of the unsupervised semantic parser, at TDWG 2009, which takes a complete volume of descriptions in MS word format. The parser marked two volumes of Flora of North America (FNA) (V.5 & 19) & two volumes of Flora of China (FOC) (V.5 & 22) with an estimated accuracy of 97% at clause level. Character level markup is to be evaluated. We are enhancing the parser to do character level markup for descriptions in other formats such as journal articles & OCRed pages under National Science Foundation grant # EF-0849982. More information about the project is at http://sirls.arizona.edu/cui/project.
System Description
1. General: The user needs to enter the configuration, source & target directories to save all intermediate & final markup results. The source folder should contain the original volume in MSWord format. It also asks the user to enter a dataset prefix that will be assigned to the volume that has been selected for markup. Saving the information would allow resuming a stopped process, later, by loading the information & going directly to the stop point.
2. Segmentation: Segments a volume to individual taxon files which are listed & saved in the in the target directory.
3. Verification: Checks individual taxon against a list of taxa in the volume if such a list is provided. It reports discrepancies between the two, which must be resolved before proceeding to the next. The verification result shows files, styles seen in files, taxon numbers are all in agreement between the taxon list file & segmentation results. Once errors are corrected, user can re-click on Start to re-verify till everything is in agreement.
4. Transformation: Transforms style tags to semantic tags. While nomenclature, habitat, distribution, discussion, & key sections are marked, the morphological description paragraphs are extracted & saved in a separate folder for more in-depth markup.
5. Structure Name Correction: The unsupervised markup algorithm is called to perform the markup on the morphological descriptions. It marks up description paragraphs clause by clause by assigning each clause a structure/organ name. The organ names learned by the algorithm will be listed for the user to review. The user can remove wrong organ names from the list, which can be retagged as unknown.
6. Unknown removal: Clause tagged as unknown is displayed for the user to assign tags. When a clause is checked, a context panel displays the context where the clause appears in the text. A drop-down selection allows the user to select/end a modifier for a tag. When a clause is saved, the unknown list is refreshed. When the user returns, he will see the previous unknown clauses.

7. Finalizer: Marks up character/states in descriptions & outputs complete XML files for individual taxon. The algorithm that performs character/state markup needs improvements.
8. Glossary: When a glossary accompanying the volume is provided, the software compares glossary with organ names & character states learned by the algorithms. It reports terms (structure names/character states) not covered by Glossary & how the algorithm understands these new terms.

Support is acknowledged from: We acknowledge the Flora of North America Project & the National Science Foundation for their continued patronage & support for our ongoing research project.

16.4. Harnessing the long tail: Small biodiversity data publishers

Vishwas Chavan, Eamonn O'Tuama, David Remsen
Global Biodiversity Information Facility

Access to, and sharing of biodiversity data of sufficient quality and volume are essential for biodiversity conservation and sustainable utilization of biotic resources. However, large volumes of biodiversity data are held by small individual and institutional data publishers. These data holders constitute the long tail of science. Most of the investment in biodiversity informatics to date has focused on the requirements of the big players. The development of tools, processes and infrastructure for small publishers that has been neglected in the past is now being addressed.

Standards, protocols, processes and tools need to be developed in such a manner that (a) they are an integral part of recording devices and instrumentation for seamlessly documenting observations; (b) authenticity, reliability, and data quality can be evaluated when data are generated or are at an early stage within the data management chain; (c) the creation of metadata is automated as much as possible and recorded as data are generated; and (d) they encourage contributions by non-English speaking data providers.

This session is aimed at debating the barriers to development of tools and infrastructure for small publishers. The session will also discuss some of the existing tools and devote time to prepare a list of requirements of small publishers who wish to be part of mainstream biodiversity informatics.

16.5. Taxonomic Name Processing ∞

David Remsen, Markus Döring
GBIF

Taxonomic names are a key component of biodiversity information. Names provide labels for taxa and nearly all information about species is tied to a scientific name. Effectively linking information about a species to an authority file or simply linking two separate data items tied to the same species name requires the ability to recognize equivalence. Computerised methods that rely on matching strings of text often have difficulty determining if two forms of a scientific name are equivalent. Scientific names have many complex and optional components that affect their composition. Many elements may be abbreviated or misspelled. Failure to reconcile this sort of variation in a name may have serious ramifications in integrating and accessing species information.

Recently there have been efforts to identify and consolidate work on name processing in order to evaluate and share approaches in addressing common problems, avoid duplication of effort and common mistakes, and identify new uses for methods and services that have been developed. The results of these efforts have enabled a range of useful tools and services as well as identified areas for further refinement and research.

In this interactive session we will provide an overview of recent developments in taxonomic name processing that include:

- Identification and distinction of key name-processing focal areas
- Development of shared vocabularies, tests, and grammars
- Review of common respositories, documentation and source code
- Reference implementations of name processing tools and services
- Identification and discussion of current challenges and key areas for continued work
- Extension of the developed methodologies into other areas of biodiversity information.

16.6. The role of persistent identifiers in tracking taxon changes

Andrew C Jones, Richard J White, Ewen R Orme

Cardiff University

The importance of persistent resolvable identifiers such as Life Science Identifiers (LSIDs) and HTTP URIs to provide metadata about objects and concepts of interest in biodiversity informatics is becoming increasingly recognised. The GBIF LSID-GUID Task Group (LGTG) has recently investigated and reported general guidelines and recommendations for the use of persistent resolvable identifiers for biodiversity data objects and concepts. However some concepts differ from data objects in that they may change through time. For example, taxon concepts may change as knowledge improves, while taxon names may or may not stay the same.

Taxon Concept Schema (TCS) metadata obtained by resolving taxon identifiers can describe taxon concepts and the relationships among taxa and names in a single taxonomic view such as a consistent monograph. It can also describe relationships between taxa as they change through time, for example during revisions, or the relationships between alternative taxonomic views and biological classifications.

This capability is not in itself enough to answer the needs of biologists who need to relate data from multiple sources subject to changing taxonomy and alternatives views. There is a need to track and document the changes in taxon concepts through time in an attempt to improve the reliability and precision of taxon referencing.

In this presentation we address the need for capturing changing views of the taxonomy of organisms, and the ability of TCS to document such changes in a machine-readable way to support services which help users interpret and use biological data. The information about taxon changes and the services which provide it will be richer and more informative if the publishers and providers of taxon data document their taxon concepts and track the changes they make and the reasons for them. If this is not done explicitly, it is possible to infer some of this missing information by comparing the information in available checklists.

Using the example of the Catalogue of Life, supported by GBIF, TDWG and the EU-funded 4D4Life project, we show how the lineage of taxon changes can be made explicit in the metadata, to support clients and services which need to understand the nature of the changes which have occurred in order to interpret and assemble data from multiple sources correctly or explain difficulties and ambiguities to users.

Support is acknowledged from: GBIF, TDWG

16.7. An Open Source Annotation Service for the Atlas of Living Australia

Ron Chernich[1], Stephen Crawley[1], Donald Hobern[2], Jane Hunter[1]
[1] University of Queensland, [2] Atlas of Living Australia

In the last quarter of 2008, the eResearch Group [1] at the University of Queensland began development on an annotation service for the Atlas of Living Australia (ALA). Twelve months on, the project has reached a state of stable production readiness, having met all of the initial objectives of the functional requirements for use in the ALA Portal. The development team has built two main open source components. The first, code-named Danno, is a Java based annotation server that implements the W3C Annotea public Draft Protocols [2], and the Open Archives Initiative Protocols for Metadata Harvesting (OAI-PMH) [3]. Users may interact with the server using the browser-independent client-side user interface components developed in parallel with Danno which we have named Dannotate, or through existing, browser-specific Annotea clients such as Annozilla [4]. Together, Danno and Dannotate form a comprehensive HTTP based annotation service for Internet or intranet content. The granularity of annotations and associated reply chains may be a full HTML page, selected areas of text on a page, and/or selected regions of images on the pages, or images served stand-alone. We describe the major technical design challenges and decisions made, highlighting the implications of some of these choices as they apply to users and administrators of the ALA, or any other systems that elect to utilize the Danno and/or Dannotate components to provide annotation services.
The Danno server implements two well known protocols. The first is Annotea, a protocol for creating and publishing sharable annotations of web documents [5] issued by the W3C as a public Draft for discussion. The second is the popular Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) which provides a convenient way for the annotations and replies to be selectively harvested by date range, or set identifier.
References
[1] http://www.itee.uq.edu.au/~eresearch
[2] http://www.w3.org/2001/Annotea/User/Protocol.html
[3] http://www.openarchives.org/OAI/openarchivesprotocol.html
[4] http://annozilla.mozdev.org/
[5] http://www.w3.org/2002/12/AnnoteaProtocol-20021219

Support is acknowledged from: Atlas of Living Australia, National Collaborative Research Infrastructure Strategy

16.8. A Species Conservation Information System for the State of Louisiana

Nelson Rios, Henry Bart

Natural History collections contain a wealth of data on biodiversity. Computer information systems based primarily on these data are increasingly being used for biodiversity conservation. A variety of software tools have been developed in recent years for databasing and georeferencing natural history data (http://specifysoftware.org/, http://www.museum.tulane.edu/geolocate/), networking and retrieving this data from distributed databases and data portals (http://digir.sourceforge.net/, http://www.gbif.org/), and using the data in biodiversity research. However, these technologies have yet to coalesce into an information system for identifying a species' critical habitat, analyzing its population trends, and using this information to plan for the species conservation. To address this issue we developed a prototype information system focused of fishes for in the state of Louisiana. The information system known as the Louisiana Fish and Wildlife Conservation Portal allows researchers to dynamically assemble data from various agency, institutional, and data aggregator sources; unify taxonomic names accross data sources, plot & display the data on digital maps, identify gaps in knowledge spatially and temporally, view static models of population trends and niches of particular fish species to identify critical habitat and assess threats to species survival.

Support is acknowledged from: Louisiana Department of Wildlife and Fisheries

16.9. WebBiodiverse - a tool for visualising biodiversity distribution and the taxonomic, phylogenetic and spatial relationships between taxa ∞

Ajay Ranipeta[1], Paul Flemons[1], Shawn Laffan[2], Dan Rosauer[2]
[1] Australian Museum, [2] University of NSW

Biodiverse is a tool for the spatial analysis of biodiversity, providing an interface for visualisation of patterns and tools for Moving Window, Clustering and Randomisation analyses. It provides a platform for the calculation of over 100 diversity metrics including Endemism, Phylogenetic Diversity and beta diversity. The WebBiodiverse system was built to allow a simpler usage and visualisation tool to achieve some of the Biodiverse features through a simpler interface. The web version also means that it can reach a wider audience without users installing the application on their systems.

While simple, feature-rich and widely available, there were a few constraints faced while working on the web version, primarily due to the lack of standards on some of the visualisation concepts across browsers. This meant that the WebBiodiverse application, currently, only supports a certain subset of features. Constraints include displaying a large phylogenetic tree, a requirement to display data as an equirectangular grid (meaning no background imagery) and browser memory limitations. However, there are existing tools and libraries available that provide a scaffolding to allow for most of the visualisation effects; they look after a common set of features across all major browsers.

One aim of the workshop/presentation would be to discuss any new avenues to provide good quality standards, tools and libraries that can provide a framework for a feature-rich, yet simple usability of a biodiversity analyses application.

Biodiverse can be accessed through purl.oclc.org/biodiverse. webBiodiverse is accessible through http://www.biomaps.net.au/WebBiodiverse/index.htm

Support is acknowledged from: University of Florida, Australian Museum, University of NSW

16.10. Wiki publishing workshop

Gregor Hagedorn, Stephan Opitz
Federal Biological Research Center (JKI)

Most web publishing and data management frameworks are centered on the presentation and data management needs of big organizations. The security, trust, and communication management required by these organizations does not reflect the traditional peer-reviewed publishing and recognition system of science. On the other hand, the majority of this publishing system is dedicated to paper/PDF publishing workflows, lacking support for data markup, creative commons licenses, and ongoing collaborations. Furthermore, whereas the conventional publishing industry recognizes the value of mini-reviews in medicine or molecular biology, the value of "taxon mini-reviews" is little recognized and thus largely relegated to informal web publications (i.e., species or taxon pages).

The development of the internet towards collaborative communities and hosted services, commonly termed "Web 2.0," demonstrates the potential of peer-based, self-organizing systems with minimal hierarchy. With respect to quality, it is possible to combine public mass-review with expert review (see, for example, recent developments in Wikipedia using

Flagged Revisions). A general prejudice against this approach is that it has little value as structured and reusable data. While largely true for the word-processor-to-PDF publishing workflow, this is not essential for all forms of document-based publishing. An object-oriented Wiki-approach can offer free-form text as well as structured and reusable data (e.g. in DBpedia.org and Open Data Linking projects). The software, "Semantic MediaWiki," (semantic-mediawiki.org) could bring the benefits of the semantic web and its technologies (like Web Ontology Language, OWL, and Resource Description Framework, RDF) within the reach of many biologists. Semantic MediaWiki allows the presentation of ontologies as glossary entries to biological experts, while ontology experts can add the necessary information to achieve the most suitable representation for machine reasoning. All information in Semantic MediaWiki is automatically exposed as OWL/RDF.

The workshop will show examples of integrating free-form text with wiki-based structured data for identification keys, to show the potential of the wiki technology.

16.11. Recording and sharing annotations during two stages of museum specimen digitization: Apiary and Atrium

Amanda K Neill[1], Jason H Best[1], John P Janovec[1], Mathias A Tobler[1], William E Moen[2], Tiana F. Franklin[1], M. Brooke Byerley[1]
[1] Botanical Research Institute of Texas, [2] University of North Texas

Annotations of specimens held in natural history collections are essential modifiers of primary records, and include expert taxonomic determination of species identity, as well as reference to studies based on the voucher. While the primary label data for a collection will be little modified after its digitization, annotations may continue to be added to the physical museum object or to the virtual representation of that object's metadata. It is imperative that these additional data, which make specimens more valuable, useful, and relevant, be recorded and shared as easily as primary specimen label data.

We are currently developing two informatics projects that seek to accommodate annotation data in all its forms and maximize the sharing of annotations, between annotated specimens and those who study them, between duplicate specimens held at multiple institutions, and between the digital record and the physical object. To effectively preserve and disseminate the annotation records generated by projects such as these, the biodiversity informatics community must develop standards and mechanisms by which these data can be shared.

The Apiary Project (www.apiaryproject.org) is addressing these challenges by exploring and developing transformation processes in a digital workflow that yields high-quality, machine-processable label data from images of specimens in a cost- and time-efficient manner. The University of North Texas's Texas Center for Digital Knowledge (TxCDK) and the Botanical Research Institute of Texas (BRIT), with funding from an Institute of Museum and Library Services National Leadership Grant, are conducting fundamental research with the goal of identifying how human intelligence can be combined with machine processes for effective and efficient transformation of herbarium specimen label information. Recording legacy annotations attached to specimen sheets is an important component of this project.
The Atrium Biodiversity Information System (www.atrium-biodiversity.org) is a web-based application developed at BRIT to facilitate the management and dissemination of biodiversity data. Atrium allows the sharing of specimen data and images with experts who can add new annotations through online determinations. Annotations can be searched and sorted, shared through workspaces, and annotation labels can be produced and printed. The recently-released Atrium version 1.7 enhances the application's annotation features, introducing confirmations with level of confidence, and the ability to record the basis of the determination.

Support is acknowledged from: Moore Foundation, Beneficia Foundation, Institute of Museum and Library Services, BRIT

16.12. Key construction with polymorphic characters

Zhimin Wang, Robert A. Morris
University of Massachusetts Boston

To produce identification keys from character matrices, information gain (also known as entropy reduction)[1], the Gini index[3], and their variants have been used to evaluate descriptive characters for ordering them in a key. However, these methods can only deal rigorously with characters that divide taxa into non-overlapped groups. In the case of polymorphic characters, which are very common in taxonomy, some heuristical adjustments to such measurements mentioned have been employed [1] [2]. To systematically address this problem, we develop a a new measurement ("Disconnectivity") based on a natural generalization of entropy in information theory, which directly handles polymorphic characters. In this talk we will present its intuition and give some examples of its application.

References:
[1] Dallwitz, M.J. 1974. A flexible computer program for generating identification keys. Syst. Zool. 23, 50–57.
[2] Hagedorn, Gregor 2007. Structuring descriptive data of organisms - Requirement analysis and information models. Ph. D. Thesis, University of Bayreuth. 270-274
[3] M. Delgado Calvo-Flores, W. Fajardo Contreras, E.L. Gibaja Galindo, R. Perez-Perez, XKey: A tool for the generation of identification keys, Expert Systems with Applications, Volume 30, Issue 2, February 2006, Pages 337-351, ISSN 0957-4174, DOI: 10.1016/j.eswa.2005.07.034.

16.13. TDWG Collaboration and Fund Raising in the Year of Biodiversity

Lee Belbin
Blatant Fabrications Pty Ltd

One of my responsibilities as part of the TDWG Infrastructure Project (TIP: http://www.tdwg.org/activities/tip/) was to consider a funding model for TDWG. When TIP started, TDWG had 23 individual members and 17 institutional members. It is amazing that an organization in TDWG's position has so few members. TDWG develops standards for sharing biodiversity information. Most of us know just how central that is for work such as GBIF, CoL, EoL, ALA, EDIT, OBIS, ITIS....and museums, herbaria etc – a HUGE international community! 'Biodiversity' also sustains our lives.
TDWG started as a club of IT inclined biotypes playing with databases. The explosion of IT has radically changed that. TDWG is now at a point where the biotypes can't easily understand the few (unbelievably valuable and overworked) people with IT skills that TDWG has fortunately attracted.
Some have not joined TDWG because they see it as 'bureaucratic': To quote Rod Page and Nike "...Just do it...". I think TDWG has addressed that. Movement is now more dependent on the individual than the committee.
It was obvious to me that we needed to build TDWG membership before we look to more handouts. With a substantial membership, we make a stronger case to potential supporters, and we can better spread the TDWG load. In 2008, we had 32 Individual members and 47 institutional members - better but this is a miniscule percentage of the domain.
As an Australian media personality likes to say (in typical Oz fashion) - "It is better to have people inside the tent pissing out than outside pissing in."
What can each of us do to help build a more effective TDWG in the Year of Biodiversity? How can we collaborate more effectively? How can we build a membership sufficient to maintain an effective infrastructure and spread the workload? How do we best attract additional funding for projects that won't get done without it?
Please come along to the session. We need some wild ideas for 2010.

Support is acknowledged from: TDWG, The Atlas of Living Australia

16.14. A graphical tool for computer-assisted plant identification

Pierre Grard[1], Pierre Bonnet[2], Juliana Prosperi[1], Thomas Le Bourgeois[1], Claude Edelin[3], Frédéric Théveny[1], Alain Carrara[1]
[1] CIRAD, [2] INRA (French Nat. Inst. for Agricultural Research), [3] CNRS

Species identification is a major constraint for biodiversity conservation.
Conventional identification tools are usually difficult to use for non specialists, mainly because they require important botanical knowledge during the identification process. For this reason, we developed a graphical identification approach that resulted in the IDAO (IDentification Assistée par Ordinateur) software. Through simple clicks on vector drawings, the user selects morphological (shape, size, position, color and texture of organs) or ecological characters corresponding to the plant he/she wants to identify, thus building a sort of "identikit picture" for the species. The software compares this set to all those available in its database with a simple matching coefficient, and provides a probable identification. At any time during the process, the user may consult species description files. Missing information is tolerated, users can thus access to a result of their identification, without the use of all characters of identification. Numerous illustrations are present in each species description in order to facilitate identification.
This graphic multi-entry identification system has been adapted to various floras (weeds, trees, orchids) around the world (West Africa, India, Cambodia, etc.), for weed control or biodiversity conservation. It is accessible on-line on Internet (http://umramap.cirad.fr/amap2/logiciels_amap/index.php?page=idao), or available on Cd-rom. Current developments on the new version of this identification tool include (i) a free open version, which will allow adaptation of the graphic interface by users according to their own flora, (ii) generalisation of the use of open drawing format (SVG: Scalable Vector Graphics), (iii) the extension of this approach to new characters (such as anatomical characters of the wood), and floras (such as paddy fields weeds). There is no constraint for the use of this tool for the identification of animal species; it wasn't realized however until today.

Support is acknowledged from: EU

16.15. Pl@ntNote: a flexible software for the management and share of data on plants

Philippe Birnbaum[1], Jean-François Molino[2], Jérôme Perez[2], Frédéric Théveny[1]
[1] CIRAD, [2] IRD

A key issue for biodiversity, agronomy and ecological studies is the availability of large and reliable databases on plants. Today, internet tools provide several services that allow the aggregation of various kinds of data into large databases. However, the preliminary and fundamental work of collecting and synthesizing data still relies on individual scientists or amateurs—each often with distinct objectives, tools and methods—who manage their data by themselves, rather than through network databases.
The resulting multiplicity of database and data file formats impedes data standardization and exchange. Furthermore, large collective databases are usually made by aggregating duplicates of individual datasets, yet often lack reliability because any control of data quality is made in the original dataset, but not in the duplicate.
Pl@ntNote is a free, easy-access and powerful solution to these problems, allowing the individual management of potentially any kind of botanical data. Its core includes generic methods independent from data type, such as editing, viewing, exporting, sharing, mapping, and searching. The user freely designs a database structure according to his/her own requirements. In a second step, structures and/or data can be exchanged within a community of users.
The distinction between structure and methods makes Pl@ntNote particularly well suited to meet the needs of scientists from various disciplines. Indeed, this tool is now in current use for the management of herbarium specimens, pictures, plant descriptions, growth surveys, weeds surveys, field inventories, living collections, genetic resources as well as palaeobotanic records. It is available in a stable beta version.

16.16. Interoperability of databases on useful plants

Helmut Knuepffer[1], Michel Chauvet[2]
[1] Leibniz Institute of Plant Genetics & Crop Plant Research, [2] Agropolis-International

Many attempts have been made in the past to compile online and printed inventories of useful plants, including cultivated plants. They differ in thematic and geographical scope, and in database structure. In an era when sustainable development is high on the political agenda, a comprehensive system of information about plant resources of the world is badly needed. It should encompass agronomical as well as botanical data and cover all the kinds of uses. Such a challenge can be reached only by a collaborative effort, involving sharing tasks, implementing standards, promoting interoperability and exchanging data sets.

An overview is given of the various existing inventories and information sources dealing with useful and/or cultivated plant species. The focus will be on species- (or taxon-) related inventories and information system, rather than on accession- (or unit-level-) based systems such as the European Search Portal for Plant Genetic Resources (EURISCO), numerous European “Central Crop Databases” and the planned “Global Information on Germplasm Accessions (GIGA). Nonetheless, cross-references between taxon-level and accession-level information systems are of vital importance.

Among the systems discussed will be taxonomic databases for cultivated plant species (Mansfeld’s World Database on Agricultural and Horticultural Crops, the Plant Taxonomy of GRIN – Genetic Resources Information Network of the U.S.), but also more general systems containing information on cultivated and useful plant species besides other plants, such as the Encyclopedia of Life or Flora Europaea. In addition, a number of standard reference books for useful and cultivated plant species will be mentioned. Both global information sources and regional ones will be taken into account. The sources considered may have a broad scope (i.e. all relevant plant species) or focus on particular groups of plants.

We offer this overview as a base and starting point to discuss the ways and means to improve the situation and better meet the needs of users. The discussion should focus on issues of harmonisation of data sources, TDWG and other standards applicable for the interoperability of such data sources, completing the list of relevant information systems and books, and possible funding opportunities for the development of a portal for centralised access to the various existing information sources. User requirements will also be addressed.

16.17. EDIT Advanced Scratchpads tutorial

Simon Rycroft, Kehan Harman, Dave Roberts, Vince Smith
Natural History Museum, London

As part of the European Distributed Institute of Taxonomy (EDIT) the Scratchpads (http://scratchpads.eu) have been developed as a data-publishing framework with which groups of people have created their own virtual research communities supporting natural history science. The system has been in use since March 2007 and currently supports more than 100 communities and 1,000 users. This workshop will provide a developer-level overview of the Scratchpad project including information on hosting a Scratchpad server, extending the Scratchpad functionality through the addition of new modules and the integration of data from other projects. The session will take the form of an overview presentation, followed by a question and answer session on the following topics:

- Scratchpad server installation
- Code repositories and module installation/updates (http://drupal.org, Subversion control system, http://svn.scratchpads.eu)
- Site installation profiles
- Taxonomy management
- Mirroring
- Future directions (ViBRANT - Virtual Biodiversity Research and Access Network for Taxonomy)

At the end of the session, attendees will know how to install their own Scratchpad server and engage with the development of the Scratchpad system, including the development of modules supporting additional functionality.

Support is acknowledged from: EU FP6

16.18. Multimedia Resources Metadata Schema

Vishwas Chavan[1], Robert A Morris[2]
[1] Global Biodiversity Information Facility, [2] University of Massachusetts Boston

The Multimedia Resources Metadata schema ("MRTG schema") is a set of representation-neutral metadata vocabularies for describing biodiversity-related multimedia resources and collections. The MRTG standard is the culmination of work on multimedia resource descriptions carried out by Key To Nature, the NBII (National Biological Information Infrastructure) Digital Image Library, MorphBank, and others, together with input from a number of other stakeholder communities including Encyclopedia of Life (EOL), the Biodiversity Heritage Library (BHL) and UMASS-Boston (University of Massachusetts Boston). The Global Biodiversity Information Facility (GBIF) commissioned the 'Multimedia Resources Task Group (MRTG)' in March 2008 and it was approved in December 2009 by Biodiversity Information Standards (TDWG) as the 'Joint GBIF-TDWG Task Group on Multimedia Resources in Biodiversity'. The standard was developed by the Joint GBIF- TDWG Multimedia Resources Task Group to fit with the suite of data standards being developed on behalf of the Global Biodiversity Information Facility (GBIF) by Biodiversity Information Standards (TDWG). During this session we intend to discuss the schema, its usefulness and improvisation, before it is ready for formal ratification by TDWG.

16.19. Bringing advanced statistical data analysis to the web: the integration of R with the Atrium Biodiversity Information System

Mathias W Tobler, John P Janovec, Jason H Best, Amanda K Neill, Anton Webber
Botanical Research Institute of Texas

Online biodiversity databases generally provide users with interfaces for searching and viewing primary biodiversity data with very limited options for other use or analysis. Advanced data processing and analysis usually takes place on the user's desktop computer, taking advantage of a wider range of available GIS and statistical software. When we developed a new module for managing vegetation survey data in the Atrium Biodiversity Information System (www.atrium-biodiversity.org), we wanted to provide users with the ability to perform statistical analysis such as hierarchical clustering, ordinations, species accumulation curves and more through an online interface without the need for downloading the data. This led us to develop a framework that integrates the free open source statistical software R with our online system. R has a large user base in the scientific community and a growing number of packages for a wide range of analysis. R is built on a powerful programming language providing many functions for data formatting and processing prior to analysis.

We created a simple graphical user interface (GUI) within Atrium that lets users select the vegetation plots they want to include in the analysis, choose a specific analysis, define input parameters (e.g. clustering method, distance measurement) and then display the resulting graphs and text output. Results can be exported as a PDF text file.

On the server side Atrium formats the R script needed to perform the analysis and customizes the SQL query and the analysis parameters based on the user input. The script is then passed to R for processing. R connects directly to the

MySQL database to obtain the data needed for the analysis and writes text and graphical output to a temporary directory from whence these are integrated into the results page. The framework is easily extendible allowing administrators to add new analyses in the form of R scripts and to edit existing scripts through an online interface. The framework allows us to perform virtually any analysis that can be performed in R.

Support is acknowledged from: Gordon & Betty Moore Foundation, Beneficia Foundation, Conservation International, National Science Foundation

16.20. Linked Literature and the Biodiversity Heritage Library

Chris Freeland
Missouri Botanical Garden

The Biodiversity Heritage Library (BHL) is an international consortium of the world's leading natural history libraries, online at http://www.biodiversitylibrary.org. BHL has been digitizing its partner libraries' collections since 2007 and now contains 40,000 volumes and 16 million pages of core biodiversity literature. New interfaces and services are required to expose this resource via programmatic means, enabling the automated linking of complementary databases such as nomenclators and biodiversity catalogues. This working session will review OpenURL and other technologies that facilitate these linkages, with a goal of highlighting new services and techniques that should be incorporated into BHL.

Support is acknowledged from: Encyclopedia of Life, IMLS, MacArthur Foundation, Moore Foundation

16.21. Experiences Building a Species Profile Model Web Service

Terry H. Catapano[1], Robert A. Morris[2]
[1] Columbia University, [2] UMASS-Boston and Harvard University

With funding from the Global Biodiversity Information Facility (GBIF) on behalf of the Encyclopedia of Life (EOL), Plazi (http://plazi.org) has implemented a Species Profile Model (SPM) service for the provision of taxonomic descriptions and other data extracted from published legacy literature and current electronic biosystematics publications. These now number over 11,000 from nearly 6,000 taxonomic treatments extracted from approximately 500 publications. EOL harvests them for inclusion in species pages. Generally standard tools were successful in supporting this service, whose access interfaces are defined at http://wiki.tdwg.org/twiki/bin/view/SPM/PlaziEOLProject and we found that a useful and robust service is possible with SPM based on adequately marked up documents.

SPM provides SPM InfoItem (SPMI) classes to provide information about taxa . In this presentation we sketch a number of issues we encountered surrounding SPM and SPMI that we recommend be addressed to provide robust use of SPM for data integration. These include:
1. SPM is based on the TDWG Ontologies, and specified in the Web Ontology Language (OWL). OWL is amenable to machine reasoning, but no clear reasoning goals have been articulated by TDWG, which leaves ambiguous several technical choices implementors might make in using or extending SPM.

2. The SPM concept associatedTaxon is underspecified. It does not provide a robust mechanism for specifying the nature of the association. It is possibly that this can be remedied with an appeal to the TDWG Taxon Concept hasRelationship, although that presently has overly narrow range.
3. Some vocabulary items in SPMI lack definition or guidance for their use. For example, One type of InfoItem is the Description. But this term is rather broadly used in biology. In systematics literature it is ambiguous whether the concept should apply to the entire section designated as the taxonomic treatment of a taxon in the article, or should refer only to the morphological description section.
4. Insufficient SPMI concepts have been specified. For example, we found no simple general way to signal the important "Materials Examined" section of typical systematics papers. This might make it difficult to mine an SPM service for occurrence data.
5. There are three different concepts in SPMI about description. These are the InfoItem subclasses Description, GeneralDescription, and DiagnosticDescription. Lacking definitions it is impossible to determine what relations these have to one another.
6. Lack of Metadata about the served SPM: We found no clear way to document within the SPM file how the SPM itself was produced. We resorted to XML comments, but it is unclear whether some standard RDF annotation mechanism might be better. Of special importance might be provenance of the SPM, including original source, changes, versions, etc.

Some classes of issues are not about SPM, but are shared with other consumers and producers of biodiversity data and

we recommend the SPM community participate in addressing them. The most recurring one surrounds various aspects of Universal Resource Identifiers (URIs) and many such issues are identified in Cryer et al. (2009). Time permitting, we will discuss some of those.

Reference:
Phil Cryer, et al. "Adoption of Persistent Identifiers for Biodiversity Informatics Draft recommendations of the GBIF LGTG", 18 August 2009, final report impending at GBIF.

Support is acknowledged from: Encyclopedia of Life; Global Biodiversity Information Facility, Plazi; University of Massachusetts at Boston Department of Computer Science; U.S. National Science Foundation.

16.22. EDIT tutorial for programmers - CDM Library

Andreas Kohlbecker[1], Ben Clark[2], Pepe Ciardelli[1], Niels Hoffmann[1]
[1] Botanic Garden & Botanical Museum Berlin-Dahlem, [2] Royal Botanic Gardens, Kew

The European Distributed Institute of Taxonomy (EDIT) is an EU-funded project that helps integrate the traditionally disparate field of scientific taxonomy as practiced in Europe.
The EDIT Platform for Cybertaxonomy provides taxonomists and those working with biodiversity data with a set of loosely coupled tools to facilitate fieldwork, analyze data, assemble treatments, and publish efficiently. Reliability and reusability of data are key requirements for each of these tools and thus for the Platform as a whole.

In order to guarantee reusability of data and to facilitate full interoperability between the various Platform components, a new data model, the “Common Data Model” (CDM) was developed. The CDM is strongly influenced by both the TDWG Ontology (http://wiki.tdwg.org/twiki/bin/view/TAG/TDWGOntology) and the Berlin Model (http://www.bgbm.org/BioDivInf/Docs/bgbm-model/); other models and standards have influenced the modelling as well. The model describes all the commonly used data that is dealt with in the platform, and therefore covers taxonomic names and concepts; literature references; authors; (type) specimen; structured descriptive data; molecular data; related (binary) files such as images or compiled keys; controlled vocabularies and terms; and species related content of any kind like economic use or conservation status.

On top of the CDM, the CDM Library - a generic open source library - has been implemented in Java, offering a local API, web service based RESTful remote API and transformation services for all major taxonomic standards from and to the CDM, making the CDM the “glue” between Platform components. This trinity of APIs not only facilitates the development of core CDM Applications such as the CDM Community Server, the CDM DataPortals, and the Taxonomic Editor. By providing basic functionality that can be extended for a particular purpose they also are ideal basis for almost any software development in the area of biodiversity. Using the CDM Library new applications can be rolled out quicker than it would take to develop a system de-novo.

The purpose of this workshop is to introduce interested software developers to the CDM Library and APIs. During the workshop the participants will learn by example how to implement a simple “Taxon of the Day” service. This REST service will deliver every day exactly one interesting taxon.

16.23. A matrix based character editor for Scratchpads

Kehan Tristam Harman, Simon David Rycroft, Ben Scott, David McL. Roberts, Vincent Simon Smith
Natural History Museum, London

A number of tools are available for editing matrix style character data, but few facilitate collaborative, web-based data editing. The Scratchpad project (http://scratchpads.eu) is a data publishing platform providing tools for taxonomists to share and manage their data on the web, in a structured way that is loosely integrated with other data resources on the web. We have developed a web-based character editor that is integrated within the Scratchpad framework. This allows users to define characters of different classes (controlled/multi-state, text and numeric) that can be edited by multiple users in a spreadsheet-style interface. Import and export will be implemented for a number of different exchange formats including Nexus for phylogenetic analysis and, eventually, SDD (Structured Descriptive Data) / DELTA (Description Language for Taxonomy) for descriptive standard compatibility. We are planning to integrate analytical features to support phylogenetic analysis, multi-access identification keys and natural language descriptions.

Support is acknowledged from: The Scratchpad project is financially supported by EDIT - a European Union framework 6 funded Network of Excellence project. The NHM, London and the Global Biodiversity Information Facility (GBIF) have provided additional grants supporting developers.

16.24. IBIS-ID, an Adobe FLEX based identification tool for SDD-encoded multi-access keys

Mircea Giurgiu[1], Andrei Homodi[1], Gregor Hagedorn[2]
[1] Technical University of Cluj-Napoca, [2] Julius Kuhn-Institute

IBIS-ID (Interactive Biodiversity Identification Software) is a software tool created to help the users in the process of identification of species or other taxa, by using the multi-access keys described in a SDD (Structure of Descriptive Data) file. The tool has been developed in the framework of "KeyToNature" (www.keytonature.eu) Project funded in the frame of the EC eContentPlus Programme. It is based on the Adobe Flex technology, a well suited candidate because of its effectiveness for data driven interactive applications and native support for dealing with data organized in XML (eXtensible Markup Language) structured files through the support of the ECMA e4x (ECMAScript for XML) standard.

The SDD format has been developed by the Taxonomic Databases Working Group and is an XML file that holds one or more identification keys (datasets). A dataset is structured into multiple blocks including TaxonNames, Specimens, Characters, CharacterTrees. The important blocks for developing the application are Characters, CharacterTrees and CodedDescriptions. From CharacterTrees the application creates a tree view of the Characters, each character defining its states (values). The CodedDescriptions block holds data that links states to specimens. This block is used every time an identification step is taken. The SDD files can be located on the same server as the application, embedded in a mediawiki page or at a different location on the Internet. Identification is carried out through the interface in which various filter functions eliminate taxa until only one is left.

The application layout is similar to Lucid and consists of four main panels: available features, states, features selected so far and remaining taxa. Identification of a specimen is done by selecting specific states of a feature that the specimen possesses. The lower left panel will keep track of states a user has already selected for a feature. These are the identification steps and by clicking on any step, the user can change his selection. The lower right panel is the most important as it holds possible identification candidates. After each successful identification step, a number of taxa are removed, keeping only those with the features described in the identification steps. The eliminated taxa are less important, but they can still be accessed via a tab in the upper right panel. The tab also displays the number of eliminated taxa from the total number.

The application has been tested with several SDD files and already integrated in the ILIAS (German for Integriertes Lern-, Informations- und Arbeitskooperations System, Open Source Learning Management System) e-Learning environment in order to log the identification steps performed by the users. This approach has been proved to be a successful one, as it allows the extraction of identification key usage statistics with important pedagogical value and with relevance for the creators of SDD-based identification keys. Further developments of tool are planned with respect to access to media resources, exploration of additional data inside the SDD data source, and improved multi lingual support.

Support is acknowledged from: EC in the Programme eContentPlus

16.25. The Environment Ontology – Linking Environmental Data

Norman Morrison
NERC Environmental Bioinformatics Centre & The University of Manchester

Every biological specimen that is collected or sampled — whether for a museum collection, for an epidemiological or population study, for research into ecology, evolution, biodiversity or sustainability, indeed any biological research — comes from a particular habitat where particular physical conditions prevail. Though there are a variety of controlled vocabularies for bioregions or biomes, there is no accepted semantic (e.g. machine-reasoning-friendly) standard for describing the environment from which these biological samples are collected. This is a serious problem for anyone wishing to retrieve and compare environmental data.

Under the umbrella of The Environment Ontology Consortium (www.environmentontology.org) work has begun to provide an integrated approach to the problem of linking environmental data. The aims of this effort are to support the semantically consistent description of, and computational reasoning over, environmental information associated with biological data of any organism or biological sample.

The task of describing the environments of organisms and biological samples has been divided into two orthogonal yet complementary sub-projects. EnvO is an ontology that describes environment types (coral reef, tundra, savanna). Gaz represents a first step towards an open source gazetteer, constructed on ontological principles, that describes places (Paris, Texas, Mount Kilimanjaro) and the relations between them. Combined, an environment ontology and associated

gazetteer, in which place names are annotated with environmental information and with GPS coordinates, will provide a format that can be read and used by software agents, thus permitting them to find, share and integrate information that ordinarily would have required human intervention.

In this presentation I will discuss how the development and application of EnvO and Gaz can serve as a testbed for establishing robust guidelines for biologists and others recording information about environmental context and other geo-spatially indexed locations. Potential domain applications include epidemiological studies; studies of immigration and emigration patterns; studies of aspects of human environments related to health and disease; studies of the effects on agriculture and livestock of climate change; all of which require controlled vocabularies for describing environmental information in association with spatial descriptions of biological phenomena of different sorts. These examples highlight just some of the areas in which an Environment Ontology would be of benefit to the scientific community. Indeed, it is our hope that the Environment Ontology will form an essential component of the nascent semantic web and that in the future it will be used for the annotation of any record that has an environmental component.

Support is acknowledged from: Meetings have been supported by NERC & NSF

16.26. TERMINIZER - ASSISTING MARK-UP OF TEXT USING ONTOLOGICAL TERMS

David Hancock
University Of Manchester

Ontologies offer the potential to assist in both searching for documents and data sets, by enabling smarter matching and automatic generation of search terms, and the interpretation thereof, where the unambiguous nature of ontological annotation leads to improved comprehension.

We present an easy to use tool that promotes the inclusion of ontologies in scientific data by assisting in the detection of ontological terms found in free text. The resulting terms are displayed either overlaid on the original text or in a list organised by the ontology and frequency. The user can interactively accept or reject each match, or try to find a more appropriate match by exploring the network of ontology concepts themselves.

The initial Terminizer service offers the full set of ontologies from the OBO Foundry, a collection of over 40 general purpose biological ontologies. In collaboration with colleagues at the International Rice Research Institute we have built a crop-specific version using the full set of ontologies from the Generation Challenge Programme.

The system is implemented as a Web service. Both the term detection service and the interactive presentation layer can be easily incorporated within other Web sites or programs.

The Terminizer system has been built using the omixed framework, an architecture supporting the rapid deployment of collaborative, Web-facing databases. More information about omixed and terminizer, including a live demonstration of the service, is available on our website: http://terminizer.org/

Support is acknowledged from: NERC : Natural Environment Research Council (UK)

16.27. GBIF LSID-GUID Task Group report and discussion on persistent identifiers

Greg Riccardi[1], Richard J White[2], Eamonn O Tuama[3]
[1] Florida State University, [2] Cardiff University, [3] GBIF

Effective identification of data objects is essential for linking the world's biodiversity data. The Global Biodiversity Information Facility (GBIF) has identified the provision of identifiers for biodiversity objects as one of the central challenges to developing a global bioinformatics infrastructure. GBIF's plans envisage using TDWG standards to "allow all data objects to be identified using standard actionable globally unique identifiers".

GBIF convened a task group, the "LSID-GUID Task Group" (LGTG) to explore the issues and offer recommendations on the use of persistent identifiers, including LSIDs (Life Science Identifiers) and other GUIDs (Globally Unique Identifiers), to facilitate the sharing of biodiversity information, with particular reference to the GBIF network. This will enable GBIF to provide architecture leadership and best practices for implementation, and become a stable, long-term provider of identifier resolution services.

The recommendations of the LGTG can be summarised as follows: GBIF's data portal is a focal point in the flow of biodiversity data. GBIF should place the use and re-use of identifiers as a high priority in assessing the quality of data, and move to a position where it mandates the use of identifiers and well known vocabularies for all data accepted by the

portal. GBIF should provide literature and training courses to ensure that all users appreciate the importance of issuing identifiers for their data and re-using identifiers from other people's data. However, many suppliers of good quality data are unable to provide reliable resolution of the identifiers they issue, and GBIF should provide services to support resolution of these identifiers and the hosting and maintenance of essential vocabularies.

This discussion session will introduce the report and its recommendations to the TDWG community, provide a forum for their discussion, and start to explore some questions which could not be answered definitively by the LGTG because they require wider discussion before a consensus can be reached. The report is seen as a step towards that goal.

The session will consist of three or four brief presentations to introduce the topics for discussion, with intervening periods for discussion. The topics will include

(i) The aims of the LGTG, issues in the use of persistent actionable identifiers, and resolution as the action which is most important in sharing and linking biodiversity data. Metadata (RDF and vocabularies), how to deliver metadata in response to resolution requests (LSIDs, HTTP URIs and Linked Data), and the need to adopt good practices for issuing persistent identifiers.

(ii) The use of persistent identifiers: the perspectives of the biodiversity user, the data provider, and the computer scientist. Issues not addressed by the LGTG, for example what kinds of objects and concepts should be given persistent identifiers, and their use to provide means for tracking changing concepts, such as taxon concepts.

(iii) The LGTG report's recommendations: the services which are needed to support persistent identifiers, their resolution and use; the role of GBIF in promoting persistent identifiers; and what the TDWG community could and should do to support GBIF and the vision of a global pool of shared biodiversity data and services.

Support is acknowledged from: GBIF