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

Biodiversity Information Standards

TDWG

The Proceedings of TDWG

Provisional Abstracts of the 2010 Annual Conference of the Taxonomic Databases Working Group

26 September – 1 October 2010
Woods Hole, Massachusetts, USA
(Hosted by Marine Biological Laboratory)

Edited by Anna L. Weitzman

Published by Biodiversity Information Standards (TDWG)

Woods Hole, Massachusetts, USA, 2010

TDWG 2010 was sponsored by:

To be cited as:

Weitzman, A.L. (ed.). Proceedings of TDWG (2010), Woods Hole Massachusetts, USA.

This book contains abstracts of the papers, posters and computer demonstrations presented at the Annual Conference of the Taxonomic Databases Working Group held 26 September – 1 October 2010 at the Marine Biological Laboratory in Woods Hole Massachusetts, USA.

The meeting attracted more than 212 participants from 27 countries and over 118 prestigious scientific research institutions, museums and companies.

The editor gratefully acknowledges with thanks the vitally important contributions of William Ulate Rodriguez and Lee Belbin towards the preparation of this publication and Chris Freeland for other assistance and support. I apologize to those whose institutional affiliations are not recorded. Only the information included with the abstract submission was included for most contributors.

Published and distributed as an Adobe® Portable Document Format (PDF) document for free download from the Conference web site at www.tdwg.org/conference2010/

Proceedings of TDWG

1 – Plenary Session. Introduction

1.1 Briefing from iEvoBio 2010

Cynthia Parr[1], Hilmar Lapp, Rod Page, Rob Guralnick, Cecile Ane

[1] EOL, National Museum of Natural History, Smithsonian Institution, Washington DC, USA, parrc[at]si.edu

This summer, the inaugural conference for Informatics for Phylogenetics, Evolution, and Biodiversity (iEvoBio 2010, www.ievobio.org) was held in association with the joint annual meeting of the Society for the Study of Evolution (SSE), American Society of Naturalists (ASN), and the Society of Systematic Biologists (SSB), in Portland Oregon USA, 29-30 June 2010. This talk reports on the nature of the conference and its relevance to the TDWG community.

Inspired by the Bioinformatics Open Source conference (www.open-bio.org/wiki/BOSC_2010), iEvoBio included both formal and informal presentations and discussion opportunities. Topics ranged from analysis methods to cyberinfrastructure to visualization tools – all aiming to advance research in phylogenetics, evolution and biodiversity. In addition to invited keynote talks, a limited number of full talks were selected through peer-review; many mentioned TDWG standards. Five-minute lightning talks were also competitive and were particularly well-received. A software bazaar session allowed hands-on demonstrations, while birds-of-a-feather gatherings enabled everyone to share thoughts in semi-spontaneous break-out groups. Finally, a challenge focusing on visualization invited developers to showcase their most recent software solutions and conference participants voted on them during the conference. All presented software had to be licensed with a recognized Open Source License (www.opensource.org/licenses), and available in source code form prior to the conference, which allowed the participants to learn from and build on the presented work. Sponsors included the US National Evolutionary Synthesis Center (NESCent; www.nescent.org) and SSB, with additional assistance from Encyclopedia of Life (www.eol.org).

The conference was a success by many measures. Over 300 individuals registered for iEvoBio 2010; most were co-registered for Evolution 2010. Many who could not attend followed an active Twitter feed (#ievobio). TDWGians served on the organizing committee, and many were featured as speakers. Andrew Hill (a frequent TDWG attendee) and his collaborators R Guralnick and S Pick, won the challenge with PhyloBox (phylobox.appspot.com), a browser-based phylogeny visualization that uses the emerging PhyloWS standard (https://www.nescent.org/wg/evoinfo/index.php?title=PhyloWS). Perhaps the greatest accomplishment was the increase in awareness (particularly among biologists who might not otherwise have been exposed to it) that biodiversity and evolutionary informatics is a rapidly developing, exciting research field.

Organizers plan to publish a report from the conference as an open-access journal article. Participants indicated strong interest in continuing as a regular international meeting. As a research-and-applications-focused venue, iEvoBio complements conferences such as TDWG (data standards) and e-Biosphere (visioning for the field of biodiversity informatics). In particular, it affords an excellent opportunity to interact with stakeholders in the evolutionary biology community.

2 – Plenary Session. Names and Concepts

Session Chair: Rich Pyle, Bishop Museum, Hawaii
The impact of Codes on technical standards, how systems/standards represent names & concepts.

2.1 e-Publication and Plant Names – using standards

Arthur D. Chapman
Australian Biodiversity Information Services, Toowoomba, Australia, biodiv_2[at]achapman.org

At the International Botanical Congress in Vienna in 2005, a Committee was re-established to examine the issue of the electronic publication of plant names and to submit proposals to alter the International Code of Botanical Nomenclature at the next Congress to be held in Melbourne, Australia in 2011. The Committee is now finalising its deliberations and is proposing a number of changes to allow for the effective publication of plant names in electronic journals and monographs from January 1, 2013. The proposed changes are based around the use of international standards such as PDF (ISO/IEC 32000-1:2008), PDF/A (ISO 19005-1:2005), ISSN (International Standard Serial Number) and ISBN (International Standard Book Number).

The Committee identified a number of key issues that needed to be addressed if electronic publication was to be allowed without it developing into a chaotic situation whereby anyone could simply publish names in one-off, privately, casually, or even unintentionally published Web pages. These issues included discovery, access, immutability, and long-term archiving.

The Committee believes that mandating the use of PDF – an international standard for exchange of electronic publications – in conjunction with ISSN and ISBNs solves many of these issues – including discovery, access and immutability. In addition, by recommending PDF/A – a standard established for long-term archiving of electronic documents – in conjunction with digital repositories such as JSTOR (www.jstor.org), The Biodiversity Heritage Library (BHL – www.biodiversitylibrary.org), PubMedCentral (www.ncbi.nlm.nih.gov/pmc), The Scientific Electronic Library Online (SciELO – www.scielo.br), Portico (www.portico.org), LOCKSS (Lots of Copies Keep Stuff Safe – lockss.stanford.edu/lockss/Home), and others, including National digital archiving repositories – solves the issues of archiving and immutability.

The proposal still has a long way to go, and has to get at least 25% support of postal votes from members of the International Association of Plant Taxonomists before being able to be submitted to the Congress next year. It then needs to pass on the floor of the Congress before being finally accepted.

2.2 PESI: Experiences Implementing Data Standards and Persistent Identifiers for Taxa

Roger Hyam[1] and Yde de Jong[2]
[1] The Natural History Museum, London, UK, roger[at]hyam.net; [2] University of Amsterdam, Netherlands, yjong[at]uva.nl

PESI stands for the Pan-European Species dictionaries Infrastructure (www.eu-nomen.eu/pesi). It is a mechanism to deliver an annotated checklist of the species occurring in Europe. At the core of PESI are three databases PlantBase (www.emplantbase.org), Fauna Europaea (www.faunaeur.org) and ERMS (European Register of Marine Species; www.marbef.org/data/erms.php). These three databases feed into a single data warehouse that runs the PESI portal (www.eu-nomen.eu/portal). Surrounding this core PESI includes interactions with geographic focal points and a network of experts and global species databases.

PESI involves merging data from multiple sources and then publishing it to a wider audience. This requires a mapping between the different schemas used by the different datasources and/or an implementation of standards within those datasources.

A pragmatic approach was taken for the three main databases. A bespoke procedure was developed at BGBM (Botanic Garden and Botanical Museum Berlin-Dahlem, www.bgbm.org) using the Common Data Model to merger PlantBase and Fauna Europaea. This data is merged with ERMS and other annotation data (such as images) in the data warehouse at VLIZ (Flanders Marine Institute; www.vliz.be). This procedure is repeated periodically to keep the data warehouse synchronised with the source databases.

In order for the datasets to be merged in this way they need to share common vocabularies for some fields – these included: taxon status, nomenclatural status and occurrence status. For each of these statuses a wide range of terms could theoretically be used but it was found that the vast majority of cases were covered with very simple vocabularies of only a few terms. This casts doubt on the benefit of building large vocabularies of terms in the absence of clearly defined applications.

PESI also provides species lists based on geographic regions. The TDWG World Geographical Scheme for Recording Plant Distributions (www.tdwg.org/standards/109) was used as a basis for land regions and the (non-standardised) polygons available for these regions exploited. This standard does not cover the seas and so VLIZ extended the standard to include the seas along with their maritime political boundaries. This has led to several open questions. Should the maritime boundaries be taken through the TDWG standards process? Where should the polygon information be stored for posterity? Should TDWG be recommending that people score data to geographic regions when GPS (global positioning system) data is available? Should we have a standard data store for arbitrary geospatial polygons?

PESI wants to release its data as Linked Data (linkeddata.org) and make it as widely available and citable as possible. To this end we discussed creating persistent identifiers for taxa in the list. Technically implementing these identifiers is straight forward but what is not clear is the contract between PESI and the consumers of the identifiers. The principle stumbling block to an agreed system is an accepted way to handle changes in the data. What behaviour is expected by a user of the system? Are the taxonomists able to support such behaviour or would it require a change in working practices?

2.3 Implementing field based data validation

John Deck
University of California at Berkeley, CA, USA, jdeck[at]berkeley.edu

The Moorea Biocode project has the ambitious goal of DNA barcoding an entire ecosystem: the island of Moorea, located in French Polynesia. This ecosystem has over 5,000 identified species and the project itself has collected and sequenced over 30,000 specimens. As part of the collecting effort in the Moorea Biocode project, we have developed informatics tools to track data from the collecting event, specimen identification, photograph, laboratory, and ultimately to host institution and sequence repositories. This talk highlights both a validation protocol and a tool that has evolved from this effort that enables field based data collection efforts from a diverse group of mostly non-technical biologists.

Data validation must precede the assignment of unique identifiers while lack of a reliable internet connection on the island of Moorea requires rules to be interpreted and enforced offline. At the same time, the validation rules need to be updated constantly. The validation rules are written in XML and define data types, numeric ranges, values in lists, and geographic extent. These elements are served in a central location and cached by the client application. The validation rules

file has been kept as simple as possible to enable further expansion into a broader global standard to be implemented for other biodiversity inventory projects.

The validation rules file is interpreted by bioValidator (biovalidator.sourceforge.net), a java-based application which can be run on Mac, Windows, or Linux systems. bioValidator interprets validation rules and runs the rules against researcher's excel spreadsheets. Once the data has been validated, it is then possible to build reliable unique keys against the data and perform other functions on the field-based data such as photo-matching, higher taxonomy lookups, and upload to a web-accessible database. These tools enable the expansion and utilization of data against known standards while working in the field. The advantages to this approach are instant feedback on data quality, enabling immediate assignment of GUIDs and simultaneous incorporation into parallel databases, and allowing for integration of multimedia and supporting data at the point of collection.

Support acknowledged from the Gordon and Betty Moore Foundation

2.4 Cross-mapping between taxonomies in the i4Life project: techniques for identifying relationships and the role of GUIDs

Andrew C Jones and Richard J White

School of Computer Science & Informatics, Cardiff University, UK, Andrew.C.Jones[at]cs.cardiff.ac.uk, R.J.White[at]cs.cardiff.ac.uk

The EC Framework 7 i4Life project, led by the University of Reading, aims to establish a Virtual Research Community involving major global programmes exploring the full extent of life on earth. The Species 2000 Catalogue of Life will complement the catalogues and indexes used by these programmes. A key issue is the ability to relate the scientific names used in each of these programmes to each other. For this a cross-map will be constructed, supporting interoperation at the taxonomic level between the i4Life project participants. Cross-map construction would be a time-consuming and error-prone process if performed by hand, and one of Cardiff's roles in the project to develop new software that will help to automate the process, reducing the number of decisions that need to be left to a human cross-map editor to make.

This software will draw on our previous experience in creating the LITCHI (Logic-based Integration of Taxonomic Conflicts in Heterogeneous Information systems) software. In LITCHI 1, constraints representing taxonomic practice were implemented in the Prolog programming language, and conflicts between checklists could be detected; in LITCHI 2, this was replaced by a deterministic Java program which built a cross-map between supplied checklists. In the i4life software we plan to re-introduce a more declarative approach to the definition of the rules to be used for generating a checklist, and in the presentation we shall discuss options for doing this. In previous work we have concentrated on the examination of the scientific names contained in checklists and their relationships to each other. With the introduction of Globally Unique Identifiers (GUIDs), we will now be making use of available GUIDs to inform the creation and maintenance of the cross-map.

Further information about the topics covered in this session can be found at biodiversity.cs.cf.ac.uk/i4life/.

2.5 A Darwin-Core Archive solution to publishing and indexing taxonomic data within the Global Biodiversity Information Facility (GBIF) network

David Remsen and Markus Döring
GBIF Secretariat, Copenhagen, Denmark, dremsen[at]gbif.org, mdoering[at]gbif.org

The Darwin Core is a body of standards, ratified by TDWG in 2009, that includes a set of terms relating to taxa and their occurrence in nature, and a set of practices regarding the use of these terms in the publication of biodiversity data and information. GBIF has adopted the Darwin Core standard as the basis for publishing and integrating taxonomic data, specifically annotated species checklists, taxonomic catalogues and nomenclatural lists.

GBIF utilises a text-based solution for employing the Darwin Core terms that provides a relatively simple and extensible data format referred to as a Darwin Core Archive (DWCA). The simplicity of the format provides a relatively non-technical option for publishing biodiversity data that does not require complicated installations of data publication software. Darwin Core Archives can be published via a simple web address or URL.

In this session we will provide a brief introduction to the structure and scope of the format, some sample data sets, and GBIF's plans for deploying it.

2.6 Methods and Tools for Name Discovery/Identification and Mapping

Lakshmi Manohar Akella[1], Holly Miller, Catherine N. Norton
[1] MBLWHOI Library, Marine Biological Laboratory, Woods Hole, MA USA, lakella[at]mbl.edu

Identification of scientific names and mapping them to identifiers in a database are considered essential for many text mining tasks like recognizing gene names [1,2], finding/extracting organism specific information lifespan/life history, geographic distributions, etc. from the literature. A scientific name finding system would also enable the extraction of all contexts/usages in which the name appears in biomedical and biodiversity literature. Organism name can serve as an important metadata element for linking and organizing information from various biological sources [1,3,4,5], so a species identification system would enable such kind of integration.

Most of the previous approaches to address the problem of name finding were primarily dictionary based, where a dictionary of names was used to identify names in text. Many biodiversity literature and legacy text sources like BHL (Biodiversity Heritage Library) contain many names with Optical Character Recognition (OCR) errors, alternative names and misclassified names. One species is discovered a day in the Galapagos islands itself with some belonging to an entirely new genus, thousands of new species are discovered every year and many are reclassified. Some names are spelled the same as geo-locations or people names and therefore disambiguation of names is required. We developed approaches and built tools that address all the above. The tool, NetiNeti, will be a one-stop solution for name identification and discovery. This tool enables finding of names with OCR errors and variations. The system is based on probabilistic machine learning methods where a given string has a certain probability for being a scientific name or for not being a scientific name depending on the name string itself and the context in which it appears. We would also present a very fast currently under development to map the extracted names to database identifiers or a master list of names.

1. Gerner M, Nenadic G, Bergman CM: LINNAEUS: a species name identification system for biomedical literature: BMC Bioinformatics. 2010 Feb 11;11:85
2. Kappeler T, Kaljurand K, Rinaldi F: TX Task: Automatic detection of focus organisms in biomedical publications. Proceedings of the BioNLP 2009 Workshop: June 4-5 2009; Boulder, Colorado: Association for Computational Linguistics 2009, 80-88.
3. Leary PR, Remsen DP, Norton CN, Patterson DJ, Sarkar IN: uBioRSS: tracking taxonomic literature using RSS. Bioinformatics 2007, 23(11):1434-1436.

4. Page RD: TBMap: a taxonomic perspective on the phylogenetic database TreeBASE. BMC Bioinformatics 2007, 8:158.
5. Sarkar IN: Biodiversity informatics: organizing and linking information across the spectrum of life. Briefings in Bioinformatics 2007, 8(5):347-357.

3 – Plenary Session. Citizen Science

Session Chair: Cynthia Parr, EOL, NMNH, Smithsonian Institution
Citizen science holds much promise for crowd-sourced data gathering and enhanced interaction among scientists and enthusiasts. Speakers in this session will share their experiences and thoughts on improving citizen science. In particular:

- How can standards and tool design foster improved data quality and sharing?
- What can we learn from systems designed to support education and professional communities?
- Where does the field of citizen science need to go in the future?

It is notable that citizen science is an underlying theme of the entire TDWG meeting. This session will raise issues to be considered as we mount a biodiversity survey of Woods Hole on Wednesday. Many TDWG participants will be testing tools and exploring data all week.

3.1 The Encyclopedia of Life and Citizen Science

Jeff Holmes[1] and Marie Studer
[1] Harvard University, Museum of Comparative Zoology, MA, USA; jholmes[at]eol.org

The Encyclopedia of Life (EOL, www.eol.org) aims to provide detailed information for all 1.9 million known species in an easy-to-navigate online format. The project brings together citizens and scientists from around the world to create a dynamic, interactive place to learn and share information. This talk highlights results from the EOL Learning and Education Group's recent collaborations as well as findings from recent surveys and interviews targeting citizen science interests and needs (education.eol.org).

With the U.S. National Parks Service and National Geographic, EOL brings together students, educators, interested citizens, and scientists for annual 24-hour BioBlitz events to document local species. EOL constructs an on-line field guide as a resource and receives data from participants. Better species identification tools are needed to encourage exploration and deeper learning. In addition, we need mechanisms for sorting, reviewing and displaying BioBlitz content simply and effectively.

In another collaborative project with the East Bay Regional Park District and KQED Public Media, students explore mudflats and eelgrass beds and make observations during low tide. Information is both consumed (such as EOL species pages) and generated as participants collect and summarize their findings. Students create video summaries, so appropriate species-oriented venues are needed for this content. Participating students learn more and are more interested in continuing their learning about nature through summer camps or after school programs.

We recently conducted a series of workshops with informal education experts, administered an online survey, and interviewed participants with extensive citizen science experience. Results fell into four categories:

Content: Citizen Scientists want reliable species information that is vetted by scientists, especially if it is placed in context. For example, information about keystone organisms, predator-prey relationships, or niches can motivate citizen scientists to participate. We will need data structures and standardized terminologies to support management and display of inter-species relationships.

Community involvement: Participants noted that communities are vital to the communication and education necessary to mobilize greater portions of society. Communities also support sharing stories, questions, concerns, and methodologies with other citizens and scientists. International communities go beyond science, increasing understanding and awareness across political and cultural boundaries. To support citizen science efforts, we must develop methods for engaging communities and provide clear entry points for citizen science data and involvement.

Tools and services: Participants asked for list-generating capabilities. These range from field guide tools to specialized lists that highlight inter-species relationships or categorize organisms by use. Having the ability to customize pages or to re-organize information for specific purposes was also seen as important. Equally valuable are species identification tools. Services that push news items such as local observations were also requested. To fulfill these needs, we need standards and methodologies to enable us to build tools that can sort, transform, explore, and re-package biodiversity information flexibly.

User interfaces: Participants desire a balance of easy access to verified information and moderated community input to ensure quality. If we are to serve different audiences, each will require a targeted interface. Thus systems should support content filters for intended audiences. Lastly, interfaces to biodiversity information via mobile applications were commonly requested to promote increased participation.

3.2 Taxonomic and data-quality challenges of large-scale citizen science: Examples from eBird

Marshall J. Iliff
eBird Project Leader, Cornell Lab of Ornithology, NY, USA; miliff[at]aol.com

Any project that gathers observations of organisms face considerable challenges with respect to taxonomy; while species names form the basic units for reporting of observations, ensuring consistency and data quality is paramount. For example, eBird (www.ebird.org) an internet-based bird checklist project which engages tens of thousands of volunteers is unique in its global scope. All eBird data is contributed to the Avian Knowledge Network (www.avianknowledge.net) which contributes all public data to GBIF; eBird is GBIF's largest single data contributor. To ensure consistency eBird has adopted a unified global avian taxonomy based upon the popular Clements Checklist of the Birds of the World. For common and scientific nomenclature, our taxonomy defers to the decisions of regional authorities, thus ensuring that eBird uses bird names that are familiar to contributors. eBird also promotes the use of taxonomic concepts based on field-identifiable forms which include taxa above and below the species level, including species-pairs, genera, and family-level identifications, as well as species, subspecies, subspecies groups, and even yet-to-be-described taxa. eBird also takes advantage of a strong body of knowledge about bird status and distribution by using experts to define regional filters, to vet outlying records, and to request voucher information (e.g., photographs or sound recordings) from users. Despite our promotion of the taxonomic concept, we occasionally must address species-level splits; assumptions about the intended taxon must occasionally be made and the taxonomic concept revised to a more specific taxon. These revisions are made rarely and with great care, but are often possible given the body of knowledge about bird distribution and the ability to maintain species-pairs as a taxonomic concept within eBird. The data quality practices and attention to taxonomic standards have resulted is a massive but well-vetted database of bird observations that is now producing extremely accurate spatio-temporal models of bird occurrence throughout the year.

3.3 Mobile Observation System for Handheld Field Data Collection

Lori Scott and Dave Hauver
NatureServe, Arlington, VA, USA, Lori_Scott[at]natureserve.org

Relatively inexpensive Global Positioning Systems (GPS) units have vastly improved the ability of field biologists to accurately geo-reference their observations. The promise of these technologies for facilitating the capture, management, and sharing of field data has yet to be fully realized, and there is a powerful demand for field-capable handheld devices specifically designed for this purpose. At the same time, there are an increasing number of smart phone-based citizen science applications coming to market with the potential to generate a flood of new observation data. How can observation data standards and the experience of practitioners working with handheld field data capture devices apply to these emerging citizen science projects?

NatureServe will introduce a new handheld field data capture device (www.natureserve.org/projects/handheld/index.jsp) and associated software applications designed to improve the ability of field biologists to record data conforming to modern geospatial concepts. The goal of this Mobile Observations System is to boost the efficiency and accuracy of the field scientists who collect observation data while providing biologists, ecologists, researchers, and conservation practitioners with quicker and easier access to observation data by reducing the time from data capture to data sharing.

One of the system's three main components is an observation template library – an online collaboration tool to define and share data structures that will be used to collect observations in the field and manage observation data online. Representing a wide variety of observation protocols, the library will enable search and reuse of the available templates to promote interoperability and data standards – at the survey protocol level or at the individual attribute definition level.

A desktop application provides a local editing environment that serves as an intermediary between the observation template library, a handheld device, and one or more long term data repositories. The desktop application aids in creating field forms for one or more specified observation templates, cleaning up and validating data that has been collected in the field, and transferring the data to another system for long term storage and analysis.

The handheld device, a GPS-enabled mobile data logger, is loaded with field data collection forms and any necessary reference data or maps required for the field survey. The system automates the creation of the field forms and temporary staging database based on the selected observation template(s).

The prototype system is currently being field tested by collaborating observation-oriented institutions. These alpha testers are attempting to identify software defects, test compatibility with existing systems and workflow processes, evaluate satisfaction of collaborators and data sharing partners, and help identify essential elements for a training program. Early feedback suggests that the system's core components could be extended to support a wider range of mobile devices, including smart phones, thereby creating an opportunity to connect trained citizen naturalists with organized scientific data collection efforts through the mechanism of sharing standardized data collection forms via the system's Template Library.

This material is based upon work supported by the National Science Foundation under Grant No. DBI–0547630—Improving Geospatial Data Capture of Biological Features: Development of a Handheld Tool for Field Inventory and Mapping. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The project team also gratefully acknowledges Trimble for the donation of four Juno ™ handheld units for development, prototyping, and testing of the system.

3.4 Amphibian Conservation Education Project

Emily Brown[1] and Ian Cottingham[2]

[1] Omaha's Henry Doorly Zoo, educate[at]omahazoo.com; [2] Computing Innovation Group, University of Nebraska-Lincoln, Nebraska, USA

Globally we are facing an amphibian decline which is paralleling the extinction of dinosaurs. Close to 6,000 known species of amphibians live in our world, yet almost 2,000 species are threatened with extinction. Research indicates amphibians are disappearing due to habitat loss, pollution and spread of the diseases, such as Chytrid fungus.

Omaha's Henry Doorly Zoo (www.omahazoo.com) has become one of the leaders in a global amphibian educational awareness campaign. In addition, Omaha's Zoo actively participates in research to determine how Chytrid fungus is spread. This effort is designed to create an awareness of the critical state of amphibians and to empower individuals to make changes in how they are utilizing natural resources.

Through the Omaha Zoo's global effort in amphibian conservation, it was discovered that little was known about native amphibian populations and the effect of Chytrid fungus in Nebraska. This is due to the fact that extensive field research is required to analyze these trends. In response, Omaha's Henry Doorly Zoo launched the Amphibian Conservation Education Project, a citizen science campaign. Through the joint efforts of science educators and students we are able to collect critical data to determine the health and vitality of Nebraska amphibian populations. Participants monitor Chytrid fungus, water quality and conduct population surveys.

Omaha's Zoo has partnered with the University of Nebraska-Lincoln Department of Computer Science and Engineering who have provided a web-based data repository for citizen scientists to record their findings, called The Biofinity Project. This database provides a pipeline of information throughout the state allowing herpetologists and participants to identify trends in amphibian populations, water quality and the spread of Chytrid fungus (biofinity.unl.edu/HDZ/amphibian/create).

The strengths of The Biofinity Project as a tool to support citizen science, such as the Amphibian Conservation Education Project, lies in the framework architecture and the My Labs feature. The system allows small tools to be easily created that are a subset of The Biofinity Project, but provide access to complete implementation. Two examples of this are implementation for the TDWG BioBlitz and the collaboration with the Omaha's Zoo. In the case of the BioBlitz, a small application for citizen scientists was quickly created to use for collection of occurrence observations. The application itself delivers only a small subset of functionality (the ability to capture location and images) but integrated with My Labs, it provides powerful tools for archival and data creation. With these tools, Biofinity allows easy linkage between bench scientists and citizen scientists. In the case of Omaha's Zoo, a My Labs setup was quickly repurposed to allow non-lab members (in this case students and teachers) to have a normalized way to enter data into a robust database, making the data accessible to a broad community through our publication mechanisms. The Biofinity Project in support of citizen science is bridging communities and streamlining tools for specific purposes.

3.5 Biodiversity Snapshots: Engaging students in citizen science using mobile tools

Elycia Wallis

Museum Victoria, Melbourne, Australia, EWallis[at]museum.vic.gov.au

Biodiversity Snapshots (www.biodiversitysnapshots.net.au) combines environmental education and mobile device tools to provide school students with an experience of being citizen

scientists. The project has been developed by Museum Victoria in Australia, with funding from the local Department of Education and Early Childhood Development, and the Atlas of Living Australia. The program consists of a web-based tool, available in either a desktop presentation or on a mobile device, for logging sightings of a variety of types of animals in the field. In this case, 'the field' may be a school yard and the exercise may be an audit of biodiversity in the school, or may be a local park, wetland, or beach. A field guide and simple "Help me ID" tool provide students, who may never have observed wildlife before, with a mechanism to identify and record sightings of birds, mammals, reptiles, frogs, insects, spiders and other garden invertebrates. Lesson planning and curriculum materials help both teachers and students to better understand why scientists gather this type of data and what research it supports. This real world project may appear simple but the innovation is in the user experience rather than in the technology. Students may never have observed wildlife before, nor may they have used mobile devices as tools to record observations. This project delivers a holistic learning experience for the user and provides many lessons for developing and delivering citizen science projects in the real world.

3.6 An overview of Citizen Science Needs for Biodiversity Standards and Tools

R.D. Stevenson
University of Massachusetts, Boston, USA, robert.stevenson[at]umb.edu

In the last 20 years Citizen Science (CitSci), also called Volunteer Monitoring, has gained momentum as a viable approach to gathering useful scientific data in domains including astronomy, bird studies, and water quality. Projects with a biological focus investigate a wide range of organisms from microbes and butterflies to invasive plants and reef fish. Most of the efforts are driven by major environmental concerns such as biodiversity loss and climate change. The data collected vary from species occurrence records to more structured information about population densities, life stages, and migration timing. Many programs are organized at the local level, within a park or watershed, but a few such as eBird (www.ebird.org) and REEF (Reef Environmental Education Foundation, www.reef.org) are now global in scope and are proving to be of great scientific value because no other research efforts cover such spatial expanses. Furthermore, these programs have the potential to contribute large amounts of data to GBIF as eBird is demonstrating. Here I present a framework for identifying standards and tools that will benefit citizen science projects by comparing the training levels and scale of the projects to conventional scientific enterprises. A key characteristic of citizen science projects is the relative inexperience of the participants in gathering data. Correct taxonomic identification is of major concern. Organizations running CitSci projects have adopted several approaches to ensuring data quality including 1) training and rating of observers, 2) development of quality assurance plans and 3) review of data by electronic filters and human experts. A number of scientific papers document the scientific validity and limitations of specific CitSci programs but there are currently no standards for data quality procedures across programs. A promising approach for improving identification accuracy is the use of electronic vouchers. The application of the Multimedia Resources Task Group standards in combination with the Darwin Core is designed to do just that but better tools are needed to make it easy for CitSci groups to use electronic vouchers. A second key feature of CitSci projects is keeping the language understandable, a prime example being their use of common rather than scientific names. The development of global names architecture service that supports common names and standard common names across languages and regions will aid CitSci projects. A third key characteristic of many CitSci projects is the ecological rather than simply taxonomic orientation of their studies. The Darwin Core now supports many of the concepts needed for these protocols but mapping of CitSci protocol variables to Darwin Core variables presents a barrier to CitSci groups. Tools that use controlled vocabularies that can be understood by citizens will expedite the adoption of Darwin Core standards. A fourth key development is the use of mobile technologies. It is likely

that the quality and amount of data will increase significantly when one can easily provide applications to CitSci participates that allow them to record and upload data with their cell phones. The growth of Internet and user computer skills has provided a way to connect people and projects over wide spatial scales at relatively low cost. While hurdles to publishing CitSci data remain, improvements in biodiversity standards and tools will lower the technological barriers for project managers and participants to share their observations.

4 – Plenary Session. Keynote Presentation

Informatics: The Quest for Information Interoperability

Michael J. Ackerman

Dr. Ackerman is the National Library of Medicine's Assistant Director for High Performance Computing & Communications, and is a pioneer within the field of medical informatics through his active research in medical imaging, high performance computing, and telemedicine; www.lhncbc.nlm.nih.gov/ohpcc/home/mjabio.html, mackerman[at]mail.nih.gov

Medical Informatics has been defined as the study of the flow of information through the healthcare system. The study rapidly leads to the conclusion that "med-speak" is more than one language, dictionaries are lacking, and the flow is often blocked. The same can often also be said about the compatibility of medical devices. I will explore the healthcare domain and show how existing standards are being put together to create both hardware and software interoperability thus fostering the rapid flow of understandable information.

5 – Plenary Session. Agrobiodiversity

Session Chair: Elizabeth Arnaud, Bioversity International
Covering issues related to agriculture and crop biodiversity (carrying over a major theme from TDWG 2009).

5.1 Biodiversity Information in developing countries: opportunities and challenges for promoting TDWG standards in Africa

Charles Kahindo[1], Theeten Franck[2], Cael Garin[2], Noé Nicolas[3], Manuana Jean-Pierre[4], Kasajima Motonobu[5], Tchibozo Sévérin[6], Mergen Patricia[2]

[1] UOB, Université Officielle de Bukavu, DR Congo, ckahindo[at]yahoo.com; [2] Royal Museum For Central Africa (RMCA), Tervuren, Belgium; [3] Belgian Biodiversity Platform, ULB, Brussels, Belgium; [4] CEDESURK, Centre de Documentation de l'Enseignement Supérieur Universitaire et de la Recherche de Kinshasa, Kinshasa, DR Congo; [5] UniversiTIC, DR Congo; [6] CERGET, Centre de Recherche pour la Gestion de la Biodiversité et du Terroir, Cotonou, Benin

Africa is endowed with a great biological diversity and many countries in eastern, central, and southern parts have embarked on developing databases for sustainable management of specimens and collections. Recent and ongoing expeditions on the continent are likely to produce valuable data from poorly known areas.

With awareness raised, there is increased willingness to share data within African countries at national, regional and international levels. It has been realized that there is a great potential for biodiversity information techniques to provide a platform for developing countries to apply state of the art bioinformatics methods to large datasets, in a practical way in order to promote networking and to address pressing issues of biodiversity conservation and management. Many governmental and non-governmental bodies have demonstrated interest in this field of activities.

GBIF and the Belgian Development Cooperation financed project CABIN (Central African Biodiversity Information Network) have played a pioneering role in supporting local initiatives in several African countries in terms of capacity building for data digitizing and TDWG standards dissemination. However a number of challenges still have to be addressed to allow African countries to fully benefit from advanced information technologies and to contribute extensively to worldwide initiatives:

- Need for infrastructure to enable large scale digitalization and related specimens preservation
- Overall limited funding compared to the challenges
- Need for more involvement from the local related policies
- Insufficient means for applying good governance guidelines

At TDWG 2009, it was identified that there is a need for an Interest Group on Biodiversity Information in Developing Countries. It is the aim of the authors to be the African voice of this new Interest Group.

This talk will give an overview of the current state of the art in Africa, highlight the challenges and the efforts to overcome them, by reporting on recent Biodiversity Information related projects and capacity building initiatives. To conclude, a work plan will be suggested to set up the new Interest group on Developing Countries concretely.

gbif.africamuseum.be/CABINPortal/ (CABIN Portal CEDESURK, Kinshasa
gbif.africamuseum.be/CABINPortal/ (Mirror at RMCA, Tervuren Belgium
Powered by BioCASE (Biological Collection Access Services): www.biocase.org

The authors gratefully acknowledge funding from DGDC (Belgian Development Cooperation), CUD (Commission universitaire pour le Développement), VLIR-UOS (Vlaamse Interuniversitaire Raad – University Development Cooperation), AUF (Agence Universitaire de la Francophonie), FFI (Fonds Francophone des Inforoutes), UNESCO (United Nations Educational, Scientific and Cultural Organization), GBIF (Global Biodiversity Information Facility), European Commission. RMCA (Royal Museum for Central Africa)

5.2 Accessing the original observation data captured during plant exploration missions for collecting crop diversity

Hannes Gaisberger, Federico Mattei, Massimo Buonaiuto, Andrea De Pirro, Valentina Barbiero, Simone Mori, Elizabeth Arnaud
Bioversity International, Rome, Italy, H.Gaisberger[at]cgiar.org

Crop diversity is collected through germplasm samples from the wild or on-farm and sent to genebanks for safety duplication, conservation and potential distribution. To ensure complete data is available in genebanks, collectors record key sample information in field books. These hand-written documents contain observations about a sample's environment, related cultural practices, traditional uses, disease symptoms and the presence of pests. Curators allocate a unique accession number to each sample when entered into the genebank collection. Collected information is included in the sample's Passport data, a standard dataset developed jointly by FAO (Food and Agriculture Organization of the United Nations) and Bioversity to facilitate germplasm information exchange. Passport data includes the accession's taxonomy, location of collection, georeferences, collector's name, data of collect, sample code allocated by the collector, etc.

Scanning the original mission documents and making them available online is key to tracking the original sample, which serves as the basis for an accession pedigree and supports the identification of potential duplicates. Original information is also crucial for the quality-checking of Passport data recorded in databases and for supporting the identification of lost samples and gaps in collections.

An online central repository was developed in 2010 to provide easy access to 27,000 scanned files (pdf format) representing collecting missions for African yam, Bambara Groundnut,

Barley, Beans, Cassava, Chickpea, Cowpea, Finger millet, Forages, Groundnut, Maize, Musa, Pear millet, Pigeonpea, Potato, Rice, Sorghum, Tree genetic resources, Wheat, Wild Vigna and Yam. Documents were scanned provided by the Agricultural Research Centre (ARC) of Lao People's Democratic Republic through joint Lao-IRRI collecting missions, International Rice Research Institute (IRRI), AfricaRice, International Institute on Tropical Agriculture (IITA) and Bioversity International (Bioversity).

Each scanned file is associated with metadata compatible with the germplasm extension of the international standard, DarwinCore, and relevant documentation such as mission reports and field books. While public users can access to the central repository through a simple search mask, content providers can also manage imported documents through a reserved area. Work regarding metadata allocation is ongoing.

Since 1974, Bioversity International (including IBPGR – International Board for Plant Genetic Resources and IPGRI – International Plant Genetic Resources Institute) has supported more than 550 germplasm collecting missions, organized with national and international partners, yielding 225,875 samples and covering 4,300 species from 137 countries. Some of these were threatened by genetic erosion in situ. Samples were shipped to genebanks, accompanied by basic Passport data, for safe conservation and continued use. One sub-sample was kept by the originating country, if adequate conservation facilities were available, while other sub-samples were stored in appropriate genebanks maintaining international base collections. This process aimed at safely store collected germplasm that would remain available to users worldwide. Many of the field books maintained by Bioversity are unique copies. This wealth of information is regarded as being a global public good. As such, Bioversity aims to complete the scanning of all documentation by the end of 2010, for public use. To date, the quality of 95,750 Passport data records has been improved through data extracted from scanned documentation. The next steps is to publish the collection mission database online, link the Passport data to the full text of the corresponding mission reports, check the georeferences, possibly with Biogeomancer, before mapping the areas where the diversity was collected.

www.central-repository.cgiar.org/crop_collecting_missions.html
code.google.com/p/darwincore-germplasm/

5.3 Beyond DarwinCore: Challenges in mobilizing richer content

Samy Gaiji[1], Dag Endresen[2], Jonas Nordling[2], Sonia Dias[3], Elizabeth Arnaud[3]
[1] The Global Biodiversity Information Facility (GBIF) Secretariat, Copenhagen, Denmark, sgaiji[at]gbif.org; [2] Nordic Genetic Resource Centre (NordGen), Alnarp, Sweden; [3] Bioversity International, Rome, Italy

One of the challenges of the scientific community is access to richer biodiversity content than the DarwinCore (DwC) concepts in order to span to phenotypic as well as genomics and ecosystems domains. The further engagement of the scientific community requires the biodiversity informatics community to provide the infrastructure response to such fundamental need. In 2009, the genebank community developed its first enriched extension to the new DwC version covering the description of phenotypic traits. In 2010, within the strategy of GBIF to expand its global infrastructure a feasibility study was initiated with NordGen and Bioversity International to assess the scalability of the GBIF infrastructure in meeting the needs of the European genebank community. Various installations of the GBIF Integrated Publishing Toolkit (IPT) were deployed within the European Plant Genetic Resources Catalogue (EURISCO) network of publishers using the DwC genebank extension.

Access to such richer biodiversity information through the European Plant Genetic Resources Catalogue (EURISCO) is critical to the scientific and policy-making users (e.g., identification of the most valuable germplasm with economically valuable traits). This presentation will highlight

the lessons learnt of this feasibility study and provide recommendations on ways forward for the expansion of the GBIF infrastructure in support of scientific communities.

6 – Plenary Session. Mass Digitization and Electronic Publications for Scientists, Librarians, Publishers and Informaticians

Session Chair: Chris Freeland, Missouri Botanical Garden
The online availability and accessibility of legacy literature and contemporary publications continues to expand through mass scanning projects and "born-digital" electronic works. This panel presentation will review issues surrounding these resources and the impacts they have on five traditional & emerging roles in natural history museums: a taxonomist, a librarian, a publisher, an informatician with a focus on schema building & data exchange, and an informatician with a focus on tool building & data extraction.

6.1 Publication, paper, and data in systematics: More or Further?

N. Dean Pentcheff
Natural History Museum of Los Angeles County, CA, USA, pentcheff[at]gmail.com

Biological taxonomy and systematics, the naming of names and the organization of taxa, have a relationship to their publication history that is unique among the sciences. Conventional scientific publication serves as a report of work and discoveries. Taxonomic publication additionally serves as the definer of biological taxa. The published description of a taxon (along with type specimens, for modern work) is the enduring definition of the taxon. Future workers can certainly publish additional information and revisions, but the publications themselves remain the defining works.

That defining role of the published literature in taxonomy and systematics has some interesting implications. One is that partial access to taxonomic literature is insufficient for taxonomic work: taxonomists must be able to see all publications relevant to a taxon to define new taxa or revise existing taxa. Another implication is that bibliographic information is important. Unlike other fields, where bibliographic information solely facilitates access to earlier work, a taxonomic reference indicates the defining text of a taxon. Differing revisions may have differing interpretations of a taxon, so the bibliographic information associated with the use of a taxonomic name defines the interpretation of that name.

For taxonomy and systematics, therefore, the publications and the bibliographic information of the field really are part of the data of the discipline, and are not just reports of the work. Researchers are understandably sensitive about proposals to change the perceived stable system of publication: peer-reviewed publication on archival paper. As we move inevitably forward into electronic publication, there are going to be issues that will need to be solved in a way that is both forward-looking and satisfies the concerns of currently practicing systematists.

Some key issues that face us as we travel forward into a more modern biology include the following:

- Permanence and immutability. Paper publication has been seen as the gold standard of permanence and resistance to modification. Electronic publications for taxonomy will need to meet and exceed those qualities of paper publications to be acceptable to practicing taxonomists.

- Access to the historical literature. When taxonomy was done in classical institutions, access to literature was assured by massive historical libraries. Now that taxonomic work is moving out of the classical European and North American institutions, that access is going to have to become electronic. Achieving that requires both technical and legal advances.
- Literature as parsable data. Historically, the only way to get data from a taxonomic publication was for a human to read it. We are now in a position to move towards taxonomic publications that combine human readability with machine-parsable components. That will allow for much better automated data collection and analysis.
- Increased speed of taxonomic work. While classical publication-based taxonomy and systematics has worked well, it has worked slowly: we have described only a fraction of the world's biodiversity. To increase the pace of taxonomic work, we are going to have to be clever and creative in using the electronic and communications infrastructure that is now available. Simply modelling paper publication in electronic form will fail to yield the efficiency we need to address the key biodiversity questions that need good science-based answers.

6.2 Libraries and the Code: The changing role of botanical libraries in the age of electronic publication.

Douglas Holland
Missouri Botanical Garden, St. Louis, Missouri, USA, doug.holland[at]mobot.org

Natural history and botanical libraries have traditionally been the repositories of books and journals containing organismal names, descriptions and classifications that attempt to organize the world's biodiversity. For centuries these libraries have diligently, acquired, preserved, cataloged and provided access to this vast and Byzantine corpus of literature in the service of taxonomy. But the shifting sands of electronic publication and scholarly communication have left libraries scrambling to find their place in this new paradigm. Their role as storage warehouse is changing, and may be shaken to its foundations as taxonomists move away from the fundamental concept that all new organism names must be available on printed paper and deposited in libraries.

The *International Code of Botanical Nomenclature* (ICBN) is the set of rules and recommendations governing with the formal botanical names that are given to plants. The ICBN is revised every six years at the International Botanical Congress (IBC). The next IBC will be convened in July of 2011 in Melbourne, Australia.

Currently the ICBN stipulates that descriptions of new plant names or changes to plant names must be published and distributed on paper in print form. They cannot be published in electronic only format. At the IBC in 2011 a special committee will propose changes to the ICBN which will allow electronic only publication of new plant names and descriptions.

These changes whether accepted or not portend significant changes in the way libraries, publishers and plant taxonomists have done business for the past 250 years. If the ICBN does not change to allow electronic only publication of plant names it risks becoming superfluous. Taxonomists will certainly begin to work around the rules through loopholes in the ICBN, or subvert it entirely. A provision in the ICBN allows limited deposit of paper copies of otherwise electronic only publications in libraries. If authors are forced to exploit this provision it will create reams of disassociated and fragmentary publications that librarian's will struggle to organize for decades to come. If the changes are accepted we will face a different challenge; monitoring, assembling, and preserving born digital information, while at the same time maintaining all those activities and services for print publications.

During the course of the panel discussion I will address some of the proposed changes to the ICBN and some of the challenging debates within the committee proposing those changes, including "immutability" and "archiving" of born digital publications as well is issues such as open access and

peer review. I will also attempt to forecast how these changes will affect libraries and the role they play in biodiversity information and informatics into the future.

6.3 Semantic tagging, semantic enhancements and XML-based editorial workflow from the viewpoint of a biodiversity publisher

Lyubomir Penev[1,3], Donat Agosti[2], Teodor Georgiev[3], Terry Catapano[2], Vladimir Blagoderov[4], David Roberts[4], Vincent S. Smith[4], Norman F. Johnson[5], Guido Sautter[2,6], Robert A. Morris[7], Vishwas Chavan[8], Tim Robertson[8], Pavel Stoev[9], Jeremy Miller[10], Sandra Knapp[4], Cynthia Parr[11], W. John Kress[12], Terry Erwin[12]

[1] Bulgarian Academy of Sciences, Sofia, Bulgaria, lyubo.penev[at]gmail.com; [2] Plazi, Bern, Switzerland; [3] Pensoft Publishers, Sofia, Bulgaria; [4] The Natural History Museum, London, UK; [5] The Ohio State University, Columbus, OH, USA; [6] IPD Bohm, Karlsruhe Institute of Technology, Germany; [2,7] University of Massachusetts, Boston, USA; [8] Global Biodiversity Information Facility, Copenhagen, Denmark; [9] National Museum of Natural History, Sofia, Bulgaria; [10] Netherlands Centre for Biodiversity Naturalis, Leiden, The Netherlands; [11] Encyclopedia of Life, Washington, DC, USA; [12] Smithsonian Institution, Washington, DC, USA

In 2010, Pensoft invested considerable effort in developing innovative ways to publish and disseminate biodiversity information, prototyping and implementing through its flagship journal ZooKeys. The following workflow is now established: (1) Submission phase – acceptance of automatically generated XML-tagged manuscripts from Scratchpads, authors' databases, and GBIF Integrated Publishing Toolkit (IPT); (2) Editorial phase – a specially designed Pensoft Mark Up Tool (PMT) provides in-house, fine granularity XML tagging of manuscripts based on the NLM TaxPub XML scheman (sourceforge.net/projects/taxpub); (3) Publication phase – PMT generates a semantically enhanced HTML version of the paper, providing: (i) visualization of main tag elements within the text (e.g., taxon names, taxon treatments, localities); (ii) internal cross-linking between paper sections, references, tables, figures, and keys; (iii) mapping of localities listed in the entire paper or for separate taxon treatments; (iv) gene sequences auto-tagged and linked to Genbank (www.ncbi.nlm.nih.gov/genbank) and Barcode of Life (www.barcodinglife.com); (v) collections acronyms linked to the Biological Collections Index (www.biodiversitycollectionsindex.org); (4) Dissemination phase – (i) final XML output of the paper validated for archiving in PubMedCentral (www.ncbi.nlm.nih.gov/pmc); (ii) PDFs with bibliographic metadata embedded, tested for uploading to the Biodiversity Heritage Library (BHL, www.biodiversitylibrary.org); (iii) taxon treatments supplied through XML to Encyclopedia of Life (EOL) and Plazi; (iv) occurrence datasets indexed by GBIF IPT (Integrated Publishing Toolkit); (v) published papers linked to their "dynamic" versions on Scratchpads (scratchpads.eu/about) and GBIF Metadata Catalogue. Another specially designed tool, the Pensoft Taxon Profile (PTP), dynamically links any taxonomic name of any rank listed within the HTML version of the paper to a wide array of leading biodiversity resources, including GBIF, EOL, NCBI – National Center for Biotechnology Information, Barcode of Life, BHL, Plazi, Tree of Life, Catalogue of Life, IT IS – Integrated Taxonomic Information System, ZooBank, the International Plant Name Index (IPNI), Index Fungorum, the Lichens Database (LIAS), Tropicos, PLANTS database, The Gymnosperm Database, Hymenoptera Name Server, Chilobase, Diptera.org, Wikipedia, Wikimedia, Wikispecies, etc. The PTP also displays links of taxon names to literature references in PubMed, Google Scholar, and BHL. Images of the taxon, if available, are provided from Morphbank (www.morphbank.net) , Wikimedia or Yahoo. With publishing of its special issue Taxonomy shifts up a gear: New publishing tools to accelerate biodiversity research (pensoftonline.net/zookeys/index.php/journal/issue/view/52), ZooKeys became the first taxonomic journal to provide a complete XML-based editorial, publication and dissemination workflow implemented as a routine and cost-efficient practice. This workflow will also soon be implemented in botany and mycology through PhytoKeys. What benefits would these developments bring to biodiversity scientists and society? XML technologies provide a new way for automated discovery and indexing of data that greatly facilitate their use and reuse; texts can be harvested, indexed and

browsed at taxon treatment level; e-publications are properly archived using PDF, XML and separate images files (PubMedCentral's practice); text and data are online and freely available, thus authors gain much broader recognition for their work (higher citation rates, new possibilities for collaboration, etc.); dynamic taxon profiles save a great amount of time to readers (e.g., taxonomists, ecologists, conservationists, citizen scientists, policy makers, etc.), because they can get information accumulated in leading biodiversity sources about any taxon, in any point of the world, just in seconds and for free.

We thank all our authors, editors, reviewers and partners for their valuable support. The partnering organizations are listed at: pensoftonline.net/Parnering%20organisations-special%20issue-ZooKeys-50.pdf.

6.4 The future of Informatics in digital literature – or literature and it's (digital) future

Donat Agosti and Terry Catapano
Plazi, Bern, Switzerland, agosti[at]amnh.org

Every year up to 20,000 new species are described, many others redescribed, and millions of pages of legacy descriptions are digitized. Beyond the domain of taxonomy is an even larger body of ecological literature. However, an almost negligible amount of original descriptions find their way to web pages and not into increasingly common communication tools such as iPads, smart phones, and other devices that potentially could provide immediate access to this content.

In this presentation we will briefly focus on what we imagine to be the future of taxonomic publishing and then describe what we see one of many possible strategies for dealing with the widening gap between the potential and the current information environment, still very much biased towards traditional print publication.

We will first try to describe our guiding view of the user demands and the technical possibilities that we consider the most relevant for the future of publishing. We then make a point that the handling of legacy and prospective publishing should be separated and dealt with separately, followed by a brief overview of how we at Plazi (plazi.org) deal with both of them. Following is a brief description of tools to deal with legacy publications under two extreme assumptions: a focus on a digitizing as large volume of pages while also encoding the content of interest at a suitable degree of precision. For prospective publishing, we will discuss our extension of the National Library of Medicine / National Center for Biodiversity Informatics Journal Archiving Document Type Definition (taxpub NLM DTD: sourceforge.net/projects/taxpub) for markup of taxonomic treatments and that is used already in the production of Zookeys (pensoftonline.net/zookeys/index.php/journal/issue/view/52). The presentation will conclude with ways of how to disseminate this content as widely as possible. If possible, we comment the various steps, alternatives and pitfalls we encountered or see.

The examples are drawn from an ongoing collaboration between Plazi, Pensoft, the Biodiversity Heritage Library, Global Biodiversity Heritage Library, Encyclopedia of Life and the National Library of Medicine.

7 – Plenary Session. Hardware and Infrastructure

Session Chair: Javier de la Torre, vizzuality, Madrid, Spain
Focusing on innovative solutions for managing biodiversity data, implications for smaller institutions.

7.1 The Google Earth Engine: A computational platform for global-scale analysis of Earth observation data

Dave Thau
Google Earth Engine, CA, USA, thau[at]google.com

Performing analyses on Earth observation data at a planetary, continental, or similarly large scale can be a time consuming and lonely task. Performing a satellite-based continental analysis of deforestation, for example, requires downloading all the necessary satellite data, preprocessing it to suit a given algorithm or workflow, and then running the analysis on a desktop computer, or a cluster of CPUs. Depending on the scale of the data, the complexity of the workflow, and the capacity of the available hardware, each of these steps could take days or weeks.

Google.org, the philanthropic arm of Google, is creating a platform called the Google Earth Engine that will greatly simplify and expedite the process of discovering and analyzing Earth observation data. The platform provides APIs for developing algorithms that can be run across the Google cloud, and for visualizing and interacting with data products via user-designed Web interfaces. Many algorithms analyzing Earth observation data can be applied on a "per-pixel" basis, and the Earth Engine algorithm development environment is explicitly designed to parallelize these algorithms so that analyses can be distributed over wide swaths of Google CPUs.

The Earth Engine platform has been designed to encourage the sharing and reuse of algorithms. Some operations, such as removing clouds from satellite imagery, are performed by most people who analyze satellite data. The Earth Engine provides a suite of expert-provided algorithms as well as the means for algorithm developers to share their algorithms with others. In addition to supporting communities of scientists, this functionality can promote scientific openness, allowing others to review and replicate analyses, as well as attempt the same analyses with alternative algorithms. A storage component is coupled to the processing and visualization aspects of the platform, providing access to large datasets, such as historic and contemporary data from NASA's Landsat 5 and 7, and MODIS satellites, and allowing users to store and make available the products of their analyses.

The Earth Engine platform, though not yet publicly available, is currently operational and open to trusted testers. This talk will introduce the Earth Engine platform in the context of biodiversity and conservation studies, review its current status, and detail some of its more complex features.

blog.google.org/2009/12/earth-engine-powered-by-google.html
googleblog.blogspot.com/2009/12/seeing-forest-through-cloud.html

7.2 Challenges of operating a global biodiversity index

Tim Robertson
GBIF Secretariat, Copenhagen, Denmark, trobertson[at]gbif.org

The Global Biodiversity Information Facility (GBIF) maintains a biodiversity Data Portal (data.gbif.org) which provides discovery and access to the content published through the GBIF network. To date, more than 200 million taxon occurrence records are indexed through the portal, originating from over 3000 datasets.

The present GBIF Data Portal has certain limitations related to the indexing latency, data throughput and functionality, which in part can be attributed to the reliance on a single database. Additionally, the Portal has limited capabilities for checklist and taxonomic content, and for metadata, and the dataset information required to determine the fitness for use is lacking.

Overcoming these limitations is a challenge for an index such as the Portal as it requires the integration of highly relational content with a large volume of spreadsheet style data, calling for

differing technologies. This presentation will present the challenges and introduce the planned evolutions to meet those challenges including:

- The inclusion of dataset level information through a metadata catalogue.
- Improved attribution and citation of resources through an enhanced GBIF Registry.
- Workflows for publishing and indexing checklist content in the GBIF ChecklistBank and the integration of those with other data types.
- Revised API for taxonomic, occurrence, and dataset metadata.
- Annotation brokerage.
- Improved processing options and output formats (including geospatial and statistical).
- The maintenance of multiple indexes for specific user groups within the GBIF Network.
- Provision of open access to the index on public platforms such as the Amazon EC2 cloud.

7.3 Updates on the Global BHL Cluster

Phil Cryer[1] and Anthony Goddard[2]

[1] Biodiversity Heritage Library, Missouri Botanical Garden, USA, phil.cryer[at]mobot.org;
[2] Biodiversity Heritage Library, Marine Biological Laboratory, MA, USA

Since the beginning of 2010, the Biodiversity Heritage Library (BHL) has loaded over 65 terabytes of biodiversity literature from the Internet Archive onto their first distributed storage cluster, which has a capacity of 100 terabytes, in Woods Hole, MA. This storage cluster was designed and assembled by BHL, using commodity hardware and open source software. By focusing on open source solutions and commodity hardware, they have been able to create a reproducible and redundant platform that allows BHL to determine the best data hosting and serving options for the project. The next phase of the project is to fully distribute the BHL corpus to other global BHL partners. 2010 saw the global growth of the BHL into China, Australia, Brazil and Europe. Initial distribution of data to Europe began in the summer of 2010 and is presently in progress. During this next phase, systems for distributing and syncing large volumes of data are being developed and tested and range from the initial, physical delivery of content to the syncing of metadata and content between partners and projects. The final phase of the project will be in the development and application of distributed processing for data stored on the cluster. During this final phase, Map/Reduce (labs.google.com/papers/mapreduce.html) and similar technologies will be leveraged to allow for large scale data mining and processing of the stored content in collaboration with informatics researchers and projects.

Challenges faced in the development and delivery of such a system include geographical load balancing and failover, data consistency and validation, bandwidth and related costs as well as physical interoperability with global systems. Technologies utilized include the GlusterFS (www.gluster.com/community/documentation/index.php/GlusterFS_User_Guide) distributed filesystem, cluster monitoring and alerting software, as well as methods and software implemented for syncing content over wide area networks. Tools such as OpenSSH (www.openssh.com), rsync (www.samba.org/rsync), lsyncd (code.google.com/p/lsyncd) and inotify (www.linuxjournal.com/article/8478) are used to automatically detect content changes and distribute them from the primary cluster to secondary server clusters. These tools and the lessons learned during the development of this project will be applicable to other large scale informatics projects that are facing the same challenges as BHL. As such, BHL global will present a roadmap showing how others can use this same software to build their own cluster using simple graphical user interface (GUI) based tools.

7.4 Use of Google APIs for Biodiversity Informatics.

Kathryn Brisbin, Rebecca Shapley
Google, CA, USA; kbrisbin[at]google.com, rshapley[at]google.com

Google has more than 80 APIs now available for developers. Many Biodiversity Informatics projects are currently making or planning to make use of some of them, e.g.:

- Mapping: Protected Planet (www.protectedplanet.net), Global Mountain Biodiversity Portal (www.mountainbiodiversity.org), BioGeomancer (www.biogeomancer.org), GBIF
- AppEngine (code.google.com/appengine): VertNet (vertnet.org), Map of Life (www.mapoflife.org)

There are several new Google APIs that can be useful for Biodiversity Informatics. Examples include:

- Prediction (code.google.com/apis/predict)
- BigQuery (code.google.com/apis/bigquery)
- Google Storage (code.google.com/apis/storage)
- AppEngine (code.google.com/appengine) with multi-tenancy, high performance image serving, and MapReduce (code.google.com/edu/parallel/mapreduce-tutorial.html)
- Re-Captcha (www.google.com/recaptcha)
- Picasa API (picasa.google.com)
- Fusion Tables (tables.googlelabs.com)

This session will offer an overview of why these APIs may be valuable for researchers and the TDWG community when planning technical projects.

7.5 SilverLining: Biodiversity data meets cloud computing

John Wieczorek[1], Aaron Steele, Dave Vieglais
[1] University of California, Berkeley, CA, USA; tuco[at]berkeley.edu

Alarm over global climate change and associated loss of biodiversity has resulted in international demand for quick, reliable access to high quality data on the spatio-temporal occurrence of species and their relation to environment. Responses to this demand have led to the development of four NSF-funded distributed database networks (FishNet2, MaNIS, HerpNET, ORNIS), which currently include 171 collections from 12 countries and 52 additional collections (20 countries) committed to participation. Collectively, these networks have successfully demonstrated community data sharing and cooperative data management.

Participation in each of these networks has far exceeded expectations, resulting in growing problems of scalability, performance, sustainability, and ability to incorporate new members. SilverLining is an EArly-concept Grant for Exploratory Research (EAGER) funded by the US National Science Foundation with the goal of evaluating the viability of public cloud computing for addressing these problems. A public cloud refers to computing resources (e.g., data storage and processing) available over the Internet that are owned, operated, and maintained by organizations such as Google and Amazon.

Initial assessments by SilverLining were made of Amazon Web Services, AppScale, and Google App Engine (GAE) for performance characteristics, development environment, and overall service costs. Based on these categories, GAE was chosen as the cloud environment in which to conduct detailed experiments to better understand the nature and cost of uploading, searching, and batch processing biodiversity data. The results of experiments provide strong evidence that cloud computing systems in general and GAE in particular offer viable solutions. An overview of the development experience, services implemented, and cost measurements obtained thus far will be presented.

7.6 Vision and Ambition for LifeWatch ICT Infrastructure

A. Poigné[1], V. Hernández-Ernst, A. Hardisty, H. Voss, and W. Berendsohn
[1] Fraunhofer Institute Intelligent Analysis & Information Systems, IAIS, Department Knowledge Discovery,
axel.poigne[at]iais.fraunhofer.de

The European Strategy Forum on Research Infrastructures (ESFRI) identified the opportunity to strengthen biodiversity research in Europe by selecting LifeWatch (www.lifewatch.eu) for its first Roadmap. The LifeWatch project has brought together eight European networks in biodiversity science, 19 national governments and, through its committees, further platform providers and users. The European Commission is funding the preparatory phase (2008 – 2011) through the Framework Program (FP7) "Infrastructures", were a roadmap as well as the legal, financial, and technical arrangements for an operational phase of 30 years is been prepared.

The LifeWatch Information and Communication Technology (ICT) infrastructure is envisioned as a network of services providing secure access across multiple organisations to biodiversity and related data and to relevant analytical and modelling tools by individual and collaborative groups of researchers, architected using ideas of Open Distributed Processing, Spatial Data Infrastructures, and Grid Computing enabling scientists to create 'e-Laboratories'. While the emphasis will be on sharing data and workflows (and associated provenance information) with others, researchers may control access rights. Resources will be available through a Service Oriented Architecture (SOA) that integrates across and based on standards as provided by standardisation bodies such Open Geospatial Consortium (OGC), TDWG, European Grid Initiative (EGI), Web 2.0 (W3C), and the Semantic Web.

Key technical result of the preparatory phase is a reference model that builds on principles such as reusability, modularity, portability, interoperability, discoverability, and compliance with standards. The *LifeWatch Reference Model* specifies the architectural approach of the LifeWatch ICT infrastructure and constitutes a solid conceptual basis for supporting these principles. The architecture is based on mature approaches (OGC Reference Model, ISO/IEC 10746) as foundation to equip the infrastructure with:

An iterative design process, distinguishing between abstract and concrete (platform specific) specifications,

- A service oriented architecture based on standards,
- Rules to define information models and services,
- Purpose-oriented information models for meta-data,
- Generic interfaces and service types,
- A workflow framework for orchestrating service chains, and
- Validation rules to determine the conformance level of concrete specifications and implementations.

Separating the abstract and the concrete design allows LifeWatch to adapt to technological changes and to provide some amount of freedom for implementations. Pre-defined information models, service interfaces, service types, and specification and validation rules will be supplied for the definition of services as a means to enhance interoperability.

The provision of provenance information and scientific workflows will be key distinguishing features of LifeWatch, complemented by mechanisms for semantic mediation and unique identification, offering a new level of support for biodiversity science.

The presentation provides an overview of LifeWatch's architecture, services to be provided, aspects of the reference model, technological options for implementation, and the timeline for construction.

7.7 The Biofinity Project – Application architecture for open data integration

Ian Cottingham[1], Stephen Scott, Leen-Kiat Soh
[1] Dept. of Computer Science & Engineering, University of Nebraska, USA, sscott[at]cse.unl.edu

The Biofinity Project (biofinity.unl.edu) is a holistic, cloud computing, approach to federated biodiversity data management. Our suite of tools focuses on the support of biodiversity and genomics research by relating data of participating organizations with those of larger, public databases like GBIF, NCBI (National Center for Biotechnology Information), and EOL (Encyclopedia of Life) as well as robust analysis tools like BLAST and ClustalW2. The Biofinity Project provides users transparent access to data search, browsing, and analysis tools that are fully integrated with an array of web-based tools and data. As a result, we have a single-source, web-enabled interface to all aspects of the framework that allows users to interact with data from diverse sources as if they were from a single provider. We will discuss the framework architecture and data specification that supports our application model and data integration components. We will specifically address the challenges of bi-directional communication with other sources for information sharing and biodiversity data and how these challenges can be overcome through the use of distributing computing architecture and storage concepts commonly referred to as cloud computing. Throughout the presentation we will focus on how these concepts can be implemented to create an open environment for inter-project collaboration, focusing on the strengths of the approach towards facilitating protocol transparency, preventing the creation of a "walled garden" and providing greater support for users of our applications and consumers of our data. We will also discuss the techniques we have employed to leverage our architecture to support citizen science efforts, providing as a specific example our use of Twitter and My Labs feature (http://biofinity.unl.edu/biofinity/lab/info) as a basis for data sharing with the TDWG bioblitz.

8. Parallel and Working Sessions

8.1 Future directions and recommendations for enhancing fitness-for-use across the GBIF network

Andrew W. Hill, Javier Otegui, Arturo H. Ariño[1] and Robert P. Guralnick
[1] Dept. Zoology & Ecology, University of Navarra, Spain, artarip[at]unav.es

In the midst of a biodiversity crisis of yet unknown magnitude, the community is working hard to coordinate the sharing and using of biological datasets from the diversity of natural sciences. In those efforts, geospatial data are a key component that can help us join biodiversity information with data from other sources to study where species exist and how they are responding to a changing environment. Assuring that biodiversity data from taxon- or biome-specific networks is as accurate as reported is essential given the myriad uses of such data in biological research, conservation assessment and education. Fortunately, the community has actively developed standardized approaches and methods for sharing biodiversity records. However, despite the best efforts of all involved, undocumented problems with geospatial data still persist. Each user therefore must vet records carefully to determine their fitness-for-use: often, a time consuming task. Although user vetting will always happen, the key discussion point here is what can be done prior to user access of data to enhance and better report the data's fitness-for-use.

Fitness-for-use refers to a scale of data quality that changes with the varying data accuracy, precision and intended use. For some applications and in the context of geospatial data, data quality can be relatively low and still fit for use. For example, coarse scale geospatial data may only be usable for continental or global analyses but certainly not for local analyses. We can split fitness-for-

use into two broad categories: (1) Are the geospatial data correct? And, (2) Are the geospatial data usable at the geographic scale of the question?

The community knows errors without annotation exist in the Global Biodiversity Information Facility (GBIF) network and this erodes the community confidence in all the data. While multiple methods of documenting fitness-for-use have been employed by both the primary institutions that curate the data as well as the organizations that assist in sharing that data, much more can and should be done. Three areas in particular require attention: improvement of revision and republication methods for data publishers, new and improved methods for documenting different areas of geospatial fitness-for-use, and adoption of new technology to increase the speed at which fitness-for-use enhancement can be performed on the entire available biodiversity information dataset. While much of the groundwork for discussing these concepts was developed in Chapman's "Principles and Methods of Data Cleaning", we attempt to build on that work to highlight several future directions for enhancing geospatial fitness-for-use in biodiversity data.

We focus on the annotation schema as one mechanism to handle the two-way exchange of information at data publishing nodes in the biodiversity network. To determine if geospatial data are correct, we explore technologies such as the GBIF filter and existing georeferencing tools such as BioGeomancer and GEOLocate among others, as a primary means to avoid geospatial errors and to generate georeferences according to best practices.

Six recommendations are proposed for enhancing fitness-for-use, ranging from support of georeferencing initiatives to enabling data access using cloud-based infrastructures through improvements in the existing data filter.

8.2 Biodiversity Information and informatics for Cooperation and Development

Draft list of Authors/discussion panel: Charles Kahindo[1], Alex Asase[2], Jean Ganglo[3], Frank Oguya[4], Innocent Akampurira[5], Patricia Mergen[6], Eric Chenin[7], Elizabeth Arnaud[8]

[1] UOB, Université Officielle de Bukavu, DR Congo, ckahindo[at]yahoo.com; [2] GBIF node manager in Ghana; [3] GBIF node manager in Benin; [4] GBIF node manager in Kenya; [5] GBIF node manager in Uganda; [6] Royal Museum For Central Africa, Tervuren, Belgium; [7] Centre IRD d'Orléans, Sud Experts Plantes, France; [8] Bioversity International, Rome, Italy

Many developing or emergent countries maintain the most biodiversity-rich environments and face extensive challenges in balancing socio-economic development with conservation practices.

Today, governments are becoming more aware of the importance of biodiversity and the need to protect natural habitats. Biodiversity information plays an important role in supporting conservation-related decision-making, particularly with regard to land planning and use and the designation of protected areas.

Access to useful biodiversity information is influenced by many factors, ranging from the proper collection of specimens and observation in the field to ensuring collected information is available online using appropriate standards and formats for various target users.

- Challenges to biodiversity information access include:
- Infrastructure to enable large-scale digitalization and related specimens preservation
- Limited funding with regard to the scope and scale of project objectives
- Involvement of relevant local policy-makers
- Available means for applying good governance guidelines
- Active involvement of local populations
- Capacity of national institutions.

Overall enhancement of local capacity and a 'train the trainers' network are needed to improve access to biodiversity information. GBIF CEPDEC (Capacity Enhancement Programme for Developing Countries) projects, as well as multiple National Cooperation and Development

Ministries and International Agencies have invested in improving the use and knowledge base of biodiversity information.

Based on working group discussions at the TDWG 2009 meeting, participants suggested that a new focus group on 'Biodiversity Information for Cooperation and Development' be established and that TDWG play an active role in its development.

The objective of this working session will, therefore, be to introduce this idea and discuss possible alternative solutions, within and across regions.

Presentations of exemplar projects in the domain have been selected and speakers identified, but the session remains open to additional contributors and participants.

Provisional session schedule:

- Introduction and objectives (Elizabeth Arnaud)
- Exemplar projects:
  - CEPDEC- Sud-Expert Plant (Eric Chenin)
  - Biodiversity information in developing countries: opportunities and challenges for promoting TDWG standards in Africa And the CABIN project (Charles Kahindo)
  - GBIF tools deployment and trainings in Benin, Ghana and Kenya, GBIF-Tanzania, training in Mauritania
  - Projects and tools promoted by Bioversity International and partners to support capacity building in information management on conservation and uses
  - Other participants interested to intervene (i.e., PI Lake Victoria Basin Biodiversity Informatics (LAVIBI) project.
- Discussion on the scope and objectives of the focus group.

Session outcome: A network of members available to establish the new TDWG focus group, to produce the texts for a Charter and coordinate future discussions and activities in order to find synergies and exchange advice and lessons learned. This will support the ultimate goal of the better understanding and application of biodiversity information for cooperation and development.

8.3 Data Citation Mechanism

Vishwas Chavan[1], Rod Page and Roger Hyam

[1] GBIF Secretariat, Copenhagen, Denmark, vchavan[at]gbif.org

The growth of online data resources in the area of biodiversity presents complex challenges with regards to their citations. However, appropriate, and deep data citation mechanism is essential to provide due incentives, recognition and rewards to all players involved in data resource conceptualization, collection, collation, to data publishing. This is not alone a technical and/or infrastructural issue, but also has deep rooted socio-political angle, and cultural change in how electronic resources are cited. This session is aimed at brainstorming the need of biodiversity data publishers to adequately cite the data resources. This session is further aimed at enlisting technical, infrastructural, social, political and cultural challenges and solutions to overcome this impediment.

8.4 Multimedia Resources Metadata Schema

Vishwas Chavan[1], Robert Morris[2]

[1] Global Biodiversity Information Facility Secretariat, Copenhagen, Denmark, vchavan[at]gbif.org; [2] University of Massachusetts, Boston, USA

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 (www.keytonature.eu), the NBII (National Biological Information Infrastructure) Digital Image Library, Morphbank (www.morphbank.net) , 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). GBIF commissioned the 'Multimedia Resources Task Group (MRTG)' in March 2008 and it was approved in December 2009 by 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). MRTG has been submitted to the TDWG standards track for review. During this session we intend to discuss the schema, its usefulness and improvisation, and potential implementation of schema through global infrastructures such as GBIF, EOL, Atlas of Living Australia (ALA) and national, regional, thematic and global multimedia repositories.

8.5 Investigating Advanced User Support Desired by Computational Biologists

Adam Eck and Leen-Kiat Soh

Department of Computer Science and Engineering, University of Nebraska-Lincoln, USA, aeck[at]cse.unl.edu, lksoh[at]cse.unl.edu

Recent years have seen an explosive growth in the use of technology to support biology research, enabling and enhancing projects ranging from gene sequencing in large labs of professional scientists collaborating from around the world to local in-field biodiversity data collection by citizen scientists. This growth is due to advancements in three primary areas: 1) infrastructure support, including the curation and publication of large-scale databases of biological data, ontologies for mapping relationships between diverse scientific terms and interdisciplinary fields of research, and cloud computing concepts for managing simultaneous, distributed access to information and tools; 2) hardware support, including the use of high performance computing, mobile devices, wireless networking, and interactive displays to increase the ability of scientists to collect and analyze data and share results both within and outside the lab; and 3) software support, including GIS tools for merging geographical context with scientific data, collaborative tools such as wikis and content management systems for shared authoring and dissemination of results, and specialized tools for performing specific, repetitive tasks.

While much prior attention has been paid to constructing systems composed of these various technologies in order to support ongoing computational biology research, less attention has been given to improving the **usability** of such systems. In this working group session, we discuss approaches for providing advanced support of user activities above the level of *simply making the system functional for research* to *making the system actively work for users.* Specifically, we are interested in the application of **intelligent support** to enhance the user experience and productivity of computational biologists to ultimately facilitate faster, better, and more interdisciplinary scientific discoveries. These approaches include providing automation for common tasks, user and student training in both system interaction and scientific discovery, and intelligent data mining and recommendations, to name a few. Drawing from advancements within the intelligent user interface (IUI) community and the TDWG community's experiences with existing computational biology tools, we aim to address five objectives during this working group session: 1) openly discuss what works and what doesn't in existing systems with ideas from the TDWG community on how to improve the usability of computational biology tools, 2) identify technologies from the IUI community and other fields within computer science to support the usability needs of computational biology researchers, 3) address potential pitfalls of adding advanced user support to computational biology tools, such as user frustration and trust issues, as well as data security concerns, 4) inquire about leveraging social networking or crowd sourcing in scientific discovery and how to intelligently support these activities

within biology research, and 5) investigate avenues for productive research between the computer science and TDWG communities to both improve computer system usability and enable more efficient and effective biology research. Discussions during this working group session will produce guidelines, use cases, and a roadmap for guiding the development of the next generation of user interfaces for computational biology systems, including within the Biofinity (biofinity.unl.edu) project from the University of Nebraska-Lincoln.

8.6 Data Hosting Infrastructure: Challenges and Potentials

Anthony Goddard, Phil Cryer, Nathan Wilson and Vishwas Chavan[1]
[1] GBIF Secretariat, Copenhagen, Denmark, vchavan[at]gbif.org

Whilst an unprecedented volume of primary biodiversity data is currently being generated worldwide, it is perceived that significant amounts of data get lost or will be lost after project closure. However, mechanisms to rescue, archive, and publish such data are currently lacking. This calls for the urgent implementation of federated 'data hosting infrastructure' at a local, national, and global scale to prevent the loss of valuable primary biodiversity data. The main objective of this session is to initiate dialog with key stakeholder communities and understand the nature of biodiversity data; the social barriers and constraints surrounding this data; as well as gaps in existing tools, standards and processes that act as impediments to the hassle free archiving and hosting of orphaned of to be orphaned biodiversity data resources. The session is further aimed at brainstorming the criteria for establishing and endorsing 'data hosting centers', and best practices to prevent loss of data.

8.7 TDWG Observations Task Group: Designing an Exchange Serialization Syntax for Scientific Observations Data and Models

Matthew B. Jones[1], Hilmar Lapp[2], Mark Schildhauer[1], Shawn Bowers[3]
[1] NCEAS, jones[at]nceas.ucsb.edu; [2] NESCent, [3] Gonzaga

A model of scientific observations that is shared and semantically expressive is one of the most promising means to discover, access, aggregate, and analyse on a broad scale scientific data about the earth and the organisms that inhabit it. To this end, the TDWG Observations Task Group aims to produce a specification for exchanging observational data from scientific disciplines that are relevant to TDWG, including species occurrences, species characteristics, ecological data, and environmental measurements. At this 2010 workshop, the Task Group will examine select observational data models, with the specific purpose to converge on an exchange specification for observational data that can serialize data from those models without losing semantic or quantitative information. The models to initially focus on were chosen for their semantic expressivity, potentially broad scope of data domains, and widespread current application for biodiversity science-relevant data. They include OBOE, the Extensible Observation Ontology, a semantic model originally developed to expose the semantic content of heterogeneous ecological and environmental observations; the OGC Observations and Measurements model, developed to exchange environmental observations as part of the Sensor Web Enablement suite; the Entity-Quality (EQ) model, developed for making descriptive biological observations computable; and Darwin Core, a vocabulary and also an XML format in widespread use to exchange species observations. The activities planned for the workshop include 1) beginning to define a syntax for the exchange standard using a few alternative approaches (such as XML Schema and RDF), and 2) for each serialization approach, identifying issues in scalability, expressivity, tool compatibility, etc. early on. Participants will draw upon existing serialization work in observations data modeling, including the use of the OBOE (www.americancoders.com/OpenBusinessObjects), O+M, and EQ serialization

approaches. The development of the exchange specification, and thus this workshop, is coordinated by a partnership of the TDWG Task Group with the NSF-funded Scientific Observations Network (SONet) and the Joint Working Group on Observational Data Models and Semantics. These initiatives have highly synergistic goals regarding the interoperability and semantic richness of observational data, but have a much broader remit.

https://sonet.ecoinformatics.org/workshops/tdwg-2010-meeting/observational-data-models-tdwg-2010

8.8 e-Infrastructure for the *4Life projects

Andrew C Jones[1], Richard J White[1] and Frank A Bisby[2]

[1] School of Computer Science & Informatics, Cardiff University, UK, Andrew.C.Jones[at]cs.cardiff.ac.uk, R.J.White[at]cs.cardiff.ac.uk; [2]Species 2000 Secretariat, University of Reading, Reading, UK, f.a.bisby[at]reading.ac.uk

In this parallel session we will discuss issues relating to the computing provision for the EC Framework 7 4D4Life and i4Life projects. The 4d4Life project is seeking to establish the Catalogue of Life as a sustainable, extensible system supported by a suitable architecture. The i4Life project aims to establish a Virtual Research Community involving major global programmes (Encyclopedia of Life, LifeWatch (www.lifewatch.eu), EBI/ELIXIR (European Bioinformatics Institute, European Life Sciences Infrastructure For Biological Information; www.elixir-europe.org), GBIF, IUCN (International Union for Conservation of Nature, www.iucn.org), CBOL (www.barcodinglife.org)) exploring the full extent of life on earth. Both of these projects are led by the University of Reading.

A key issue in 4d4Lifeis to adopt an architecture which is open and based on widely-accepted standards. A Service Component Architecture (SCA) approach is being adopted, due to its implementation language-independent means of defining services, the ease with which new services can be introduced, etc. This architecture will support management and maintenance of the Catalogue of Life, interoperation with other clients that interact with the Catalogue of Life without human intervention, etc., and allow the integration of "business rules" to control these processes.

In the i4Life project, it will be necessary to interrelate the catalogues and indexes used by the various global programmes. The Species 2000 Catalogue of Life will complement the catalogues and indexes used by these programmes. To facilitate this, a "cross-map" supporting interoperation at the taxonomic level between the project participants will be needed, due to differences in classification. Since the project participants work in quite diverse ways, with differing interfaces to their systems, etc., the transmission and use of relevant data between systems will also present challenges. It is expected that techniques to be discussed will include techniques for populating the cross-map by semi-automated means, for using the cross-map, and for transmitting and using relevant data (including augmenting partners' catalogues and indexes using data from other partners). Presenters will discuss experience gained in relevant past projects where possible.

This session will include presentations from project partners which:

- Give an overview of the 4d4life and i4life projects, and the importance of the infrastructures being developed;
- Explain the key elements of the new 4D4Life "e-2 architecture";
- Discuss the "pipelines" between the global biodiversity projects in i4Life, and the role of the Catalogue of Life; and
- Discuss cross-mapping techniques in i4Life.

Following these presentations there will be the opportunity for members of the audience to give feedback on the project plans, and on how they hope to use the services that will be developed.

Further information about the 4D4Life and i4Life projects can be found at these URLs 4d4life.eu/, biodiversity.cs.cf.ac.uk/4lifeprojects/.

8.9 Bringing it all together – How to build rich biodiversity data portals using the EDIT Platform for Cybertaxonomy

Andreas Kohlbecker[1], Anton Güntsch[1], Agnes Kirchhoff[1], Florian Causse[2], James Davy[3], Andreas Müller[1] & Walter G. Berendsohn[1]

[1] Botanic Garden & Botanical Museum Berlin-Dahlem, Germany, a.kohlbecker[at]BGBM.org; [2] Universite Pierre et Marie Curie Paris, France; [3] Royal Museum for Central Africa Tervuren, Belgium

The European Distributed Institute of Taxonomy (EDIT) is an EU-funded project aimed at integrating taxonomic research and research infrastructures in Europe. One of the project's main outputs is the "EDIT Platform for Cybertaxonomy". The Platform brings the taxonomic workflow to the Internet, providing an open architecture to connect and integrate existing applications and for developing new tools where gaps in the workflow exist.

The Platform provides a set of tools to facilitate fieldwork, analyze data, assemble treatments, and publish them efficiently. Reliability and reusability of data are key requirements for each of these tools and thus for the Platform as a whole.

The workshop will give a practical introduction to the Platform concepts and architecture centred on the EDIT Common Data Model (CDM) and the CDM software library. We will use the example of constructing a rich biodiversity data portal which displays alternative classifications, full featured synonymies, distribution maps, descriptions, media and more. We will demonstrate how a variety of different EDIT Platform software modules and services can be combined into an integrated view on all kinds of data related to a specific taxonomic group.

A focus will be on the EDIT Taxonomic Editor ("EDITor"), which plays a key role in the process of data integration. We will also explain how data can be integrated using one of the Platform data import modules, e.g., for TCS (Taxon Concept Schema), ABCD (Access to Biological Collections Data) and SDD (Structure of Descriptive Data).

Finally, the workshop will demonstrate the CDM Dataportal. This platform module is build using the Drupal content management system and accesses the web services accessing the CDM Community Store. The Dataportal is a fully configurable user-friendly and feature-rich taxonomic data portal that can be tailored for a specific User community.

8.10 TDWG Phylogenetics Standards Programming Workshop

Hilmar Lapp[1] and Nico Cellinese[2]

[1] US National Evolutionary Synthesis Center, Durham, North Carolina, USA, hlapp[at]nescent.org; [2]University of Florida, Gainesville, Florida, USA, ncellinese[at]flmnh.ufl.edu

Since 2009 several of the emerging standards for accessing and exchanging phylogenetic data have been used or implemented in applications ranging from research-level interoperability studies to production community resources. For example, the programmatic data access interface of the newly released version of TreeBASE (www.treebase.org) now supports the PhyloWS (evoinfo.nescent.org/PhyloWS) web-service standard, returns records in the NeXML (nexml.org) exchange standard (among others), and uses the Comparative Data Analysis Ontology (CDAO, www.evolutionaryontology.org) (as well as the Dublin Core and Darwin Core vocabularies) to express the semantics of metadata elements. Other efforts include converting the contents of TreeBASE into a format used by generic machine reasoners and building visual data browsers on top of it. Initiatives such as these are pivotal in establishing proof-of-principle for interoperability through standards application, and in identifying gaps requiring further standards development. This one-day, hands-on working meeting aims to support these on-going initiatives by providing an opportunity for developers to work together on issues ranging from gaps in phylogenetics standard compliance, to improving the existing standards, to increasing the depth and coverage of their

documentation, all with a specific perspective on biodiversity informatics needs. Several potential targets have already been identified by members of the Phylogenetics Standards Interest Group (www.tdwg.org/activities/phylogenetics), including increasing the support for the emerging data standards (e.g., NeXML) among tree visualization tools; exposing the taxonomic data in online biodiversity data resources (such as EOL, www.eol.org and Scratchpads, scratchpads.eu) through a standard data access interface (e.g., PhyloWS); and defining for commonly used exchange formats how the leaf nodes of trees can be linked to those metadata that best facilitate integration of phylogenetic data (such as species names in a taxonomy, and geo-coordinates of the corresponding specimens in a museum collection). Another target is further developing the Phyloreferencing standard (evoio.org/wiki/Phyloreferencing_subgroup) that was initiated in 2009, and which forms the basis of the topological query support in PhyloWS. The targets will be further narrowed down to a feasible scope prior to the conference through online discussion on the tdwg-phylo mailing list (tdwg-phylo[at]tdwg.org), and everyone interested in participating should either join the discussion there, or contact the Interest Group conveners directly. We anticipate that the tasks eventually selected will need metadata and documentation experts as much as developers.

8.11 Annotation of biodiversity data: Toward standards related to their structure, capture, transport, and management.

James Macklin
Harvard University Herbaria, Cambridge, Massachusetts, USA, jmacklin[at]oeb.harvard.edu

There is an urgent need to discuss, evaluate, propose, and standardize mechanisms for the collection, distribution, and storage of annotations of biodiversity data. The Filtered Push and Morphbank (www.morphbank.net) projects have recently begun examining requirements for the representations of annotations and their exchange. These and other projects have submitted a TDWG Annotations Interest Group proposal now nearing the end of the acceptance process. It is clear that we all need to insure that our resources can communicate with each other and that we have similar understandings of the nature and concerns of annotations, and how annotations should interact with transport mechanisms and with existing domain standards. The workshop will be broken into four working sessions: 1) Documentation of use cases and requirements, 2) Annotation semantics and ontologies, 3) Domain specific interactions, particularly the roles of existing TDWG standards, 4) Transport and Interchange of annotations.

8.12 GBIF Knowledge Organization

Robert A. Morris[1], Terry Catapano, Donald Hobern, Hilmar Lapp, Norman Morrison, Natasha Noy, Mark Schildhauer
[1] University of Massachusetts, Boston, USA, morris.bob[at]gmail.com

The team listed below is charged by the Global Biological Information Facility (GBIF) with developing a position paper for it about its future needs for Knowledge Organization Systems (KOS), a rubric which includes thesauri and other controlled vocabularies, ontologies, gazetteers, and other such resources. Of course, the paper will cover tools, impediments to deployment, training requirements, community mechanism requirements, etc. A draft for public comment will be available at about the time of TDWG 2010. This session will discuss the deliberations to date, solicit further input, and hope to generate further community input to the process. Several members of the team will be available throughout TDWG 2010 for informal discussion.

8.13 Nomina VII: Progress and priorities with a global names architecture – discussion session

Coordinated by D. Patterson
Marine Biological Laboratory, Woods Hole, Massachusetts, USA, dpatterson[at]eol.org

The names of organisms label almost every piece of information about organisms. Names have the potential to be used to index and organize biological information. Names are also the foundation of our tally of living and extinct species. The vision of a names-based cyberinfrastructure is of a virtual layer that interconnects expert sources of names information to serve the needs of nomenclaturists, taxonomists and managers of biodiversity information. The goal has been pursued through a series of Nomina workshops that created a road map roadmap for the 'Global Names Architecture', and a number of initial components. These include CiteBank and a Global Names Index as repositories of raw name strings and citations (respectively). Through reconciliation, these are mapped into clean forms that can be stored by the central data store – a global names usage bank (GNUB). GNUB is being developed to also serve the functions of animal and fungal nomenclators. The initial components share information via Darwin Core Archive (DWCA), and include an interface to allow experts to view and improve the underlying infrastructure. This workshop (Nomina VII) seeks input on the overall architecture, functional, and technical requirements. The outcome will be a list of priorities. Issues to be covered may include the Global Names Index, CiteBank, GNUB, ZooBank (and other nomenclators), reconciliation challenges, names services (such as services that will normalize names from different sources) name-linking services, semantic data-linking, data exchange, and governance.

8.14 Citizen Science

David Remsen[1], Joel Sachs
[1] GBIF Secretariat, Copenhagen, Denmark, dremsen[at]gbif.org

This will be an informal session, broken into two parts. The first will be a post-mortem on the bioblitz, with discussion on what worked, what didn't work, etc., with a focus on being relevant to "real world" citizen science. The second part will be more forward looking, and focused on the roles TDWG could/should play in citizen science efforts. We will begin the session by crafting our agenda – discussion topics can be suggested in advance, or during the session.

Topics for Part 1 will likely include:

i) Which applications rose to the top and why?
ii) What were the core unmet needs during and immediately after the blitz?
iii) Were the existing standards up to the task?
iv) Pulling together a lessons learned manuscript for publication.

Topics for Part 2 will likely include:

i) the role of standards in citizen science;
ii) strategies for data quality; and
iii) the relationship between citizen science and the new field of "community remote sensing".

Other possible topics include grass-roots activities; integrating activities; tool development; use of social and semantic computing; and visions for the future.

8.15 Persistent Identifiers Guide workshop

Kevin Richards
Landcare Research, New Zealand, RichardsK[at]landcareresearch.co.nz

The Beginners Guide to Persistent Identifiers booklet is being drafted in response to a request by GBIF to provide documentation about persistent identifiers (or GUIDs) to the many and varied users whom have little experience the with concept and associated technologies. The brief of the document is to educate novice users about issuing, management, tools and uses of Persistent Identifiers for biodiversity resources. The booklet is currently at the community review stage and feedback on the structure and content of the document has been requested.

The intention of this workshop is to run through the current version of the document and discuss any issues, concerns and comments about each section.

The attendees of the workshop should read through the document prior to the workshop and be equipped with questions and comments to address during the session. The document is available at community.gbif.org/pg/file/eotuama/read/8160/

9. Posters

9.1 Implementation of GBIF's Architecture in GBIF France: from ideas to reality

Michael Akbaraly, Delphine Gasc, Anne-Sophie Archambeau
GBIF France, akbaraly[at]gbif.fr

The Global Biodiversity Information Facility (GBIF; www.gbif.org) is an international organisation that focuses on making scientific data on biodiversity available via the Internet using web services. GBIF's information architecture makes these data accessible and searchable through a single portal. Currently the data portal provided more than 203 million occurrences from many institutions from around the world. Data available through the GBIF portal are primarily distribution data on plants, animals, fungi, and microbes for the world, and scientific names data.

Priorities, with an emphasis on promoting participation and working through partners, include mobilising biodiversity data, developing protocols and standards to ensure scientific integrity and interoperability, building an informatics architecture to allow the interlinking of diverse data types from disparate sources, promoting capacity building and catalysing development of analytical tools for improved decision-making.

To fulfill this mission, GBIF has established a network of national nodes whose mission is to promote the existing information in its territory. The national node uses tools developed by GBIF to connect and publish data. Thus, GBIF France (www.gbif.fr) gathers data hosted by France and the node's team provides support to data providers.

Besides these activities, GBIF France participates in GBIF's enhancement by acting as a tester and pioneer in the use of the newly developed tools. Thus, when GBIF has decided to redefine their architecture, GBIF France has been one of the first to adopt and promote this new strategy.

This strategy is to facilitate and enhance the data retrieval by creating a 'highway to data' by using three components:

- IPT (Indexing Publishing Toolkit): a software platform providing an efficient publishing of biodiversity data on the Internet, through the GBIF network.

- HIT (Harvesting and Indexing Toolkit): an open-source, Java-based web application that simplifies the otherwise complicated process of harvesting biodiversity data from a distributed network of data publishers.
- And the implementation of a node's portal

By describing our one-year experience in this poster, we aim to highlight the challenges that necessarily appear when implementing this "highway to data".

9.2 The Biofinity Project – Managing Data in the Cloud

Shawn Baden, Ian Cottingham, Stephen Scott and Leen-Kiat Soh
Department of Computer Science and Engineering, University of Nebraska---Lincoln; sbaden[at]cse.unl.edu, icottingham2[at]unl.edu, sscott[at]cse.unl.edu, lksoh[at]cse.unl.edu

The Biofinity Project (biofinity.unl.edu) is a holistic, cloud computing, approach to federated biodiversity data management. Our suite of tools focuses on the development of a software framework that supports biodiversity and genomics research by relating data of participating labs with those of larger, public databases. It also gives users access to data search, browsing, and analysis tools that are fully integrated with the interface. Our efforts have resulted in the creation of a single-source, web-enabled interface to all aspects of the framework that allows users to interact with data from diverse sources as if they were from a single provider. The unification of many disparate, existing data sources provides community support for independent research activities while encouraging data sharing and collaboration. The Biofinity Project includes data management workflows that allow users to easily manipulate data in a standard format, allowing it to be readily published and shared as part of the data sets that are federated by the system. The workflows not only provide data management functionality, but also give users access to analytical tools through the same common web-enabled data set. Users can run standard bioinformatics tools on both private and public data sets, benefiting from the data integration components of the system. This allows users access to data sets significantly larger than what could be generally maintained in a local lab environment.

Our poster presents an overview of The Biofinity Project and will discuss how researchers can utilize features of the system to import, analyze, edit, publish, share, integrate, and otherwise manage complex biodiversity and genomic data sets. The poster includes an overview of integrated technologies relative to the data management approach taken by the system including:

- Google Maps Integration;
- Mobile Device (iOS and Google Android) Services;
- Integration with External Data Repositories including: GBIF, NCBI (National Center for Biotechnology Information), Encyclopedia of Life, uBio (www.ubio.org), and others; and
- Social Networking for Biology using Twitter for Citizen Science and Flickr for image management.

Our poster discusses how these technologies, and our integrated data management system, can be used by researchers to streamline complex data management tasks from a single web-based gateway to a rich set of Cloud Services, data APIs, and integrated data and tools.

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

Melita Birthälmer[1], Boris Jacob[2]

[1] Museum für NaturkundeLeibniz Institute for Research on Evolution and Biodiversity at the Humboldt University Berlin, Melita.Birthaelmer[at]mfn-berlin.de; [2] Botanic Garden and Botanical Museum Berlin-Dahlem, Germany, jacob.boris[at]googlemail.com

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 information can be found in scientific books and journals of the past centuries. Natural history museums and botanical gardens collectively hold the majority of the world's published knowledge on the discovery and subsequent description of biological diversity up to the present, and the only way to access this knowledge is to visit a number of different, geographically dispersed, specialist libraries. This hinders fundamental research in the domain of biodiversity as it depends – more than any other natural science – upon historic literature. 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; www.bhl-europe.eu) 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 focuses on interoperability of existing European digital libraries and repositories with the goal of providing open access to the wider public, scientists, and decision makers. It aims to implement technological solutions for search and retrieval and long-term sustainability of digitized content. BHL-Europe is designed as a best practices network, not focused on the task of digitization, which is left to each content provider, but assisting in future scanning activities.

BHL-Europe will make biodiversity literature freely available on three online platforms: a multilingual BHL-Europe portal for search and retrieval, the Global Reference Index to Biodiversity (GRIB), and Europeana.

The technical architecture of BHL-Europe is based on the Open Archival Information System (OAIS) reference model and enables content ingestion, archival storage and delivery of content to users of already digitised content. A prototype of the multilingual BHL-Europe portal with sophisticated search tools will be available end of fall 2010.

The GRIB is built in cooperation with the European Distributed Institute of Taxonomy (EDIT) and will serve as a central index in the BHL-Europe infrastructure. The GRIB will be a bibliographic database including the library catalogues of BHL-Europe partners and will provide deduplication and scanning management functionalities. A prototype of the GRIB is working already and the final system is expected to be finished in spring 2011.

In addition to the BHL-Europe portal, biodiversity literature will be as well accessible through Europeana, the portal of the European Digital Library. Since June 2010, more than 82,000 books are currently accessible in Europeana, and this number will increase continuously while BHL-Europe is harvesting digital content from its content providers.

Making digital content of numerous biodiversity libraries from all over the world interoperable through the BHL-Europe portal will bring the vision of a global digital open access library for biodiversity literature to life.

9.4 Digitizing Engelmann's Legacy

Mike Blomberg, Chris Freeland, Doug Holland, Stephanie Keil
Missouri Botanical Garden, St Louis, USA, Mike.Blomberg[at]mobot.org

The approximately 8,000 plant specimens contained within the Missouri Botanical Garden (MGB) Herbarium's George Engelmann Collection represent plants gathered during pioneering expeditions into the Native American West following those of Lewis and Clark. These specimens represent the first scientific record of the plants growing in the vast wilderness west of the Mississippi River. As such, they form the earliest verifiable documentation of species occurrences before the rapid migration west permanently altered that pristine landscape through human alterations and the introduction of invasive species. These specimens have been digitized and a representative set have been geocoded in Tropicos (www.tropicos.org/project/engelmann), MBG's botanical information system, with visualization and analysis facilitated through integration of ArcGIS Server. These specimens and the resulting interactive maps help inform scientists, students, and the public about the historic distributions of species throughout the Native American landscape.

9.5 Toward a plant taxonomic name resolution service

Brad Boyle[1], Sheldon McKay[2], Brian Enquist[1], Jerry Lu[2], Nicole Hopkins[2], William Piel[3], Kathleen Kennedy[2], Matthew Helmke[2], Evan Deabl[2]
[1] University of Arizona, Tucson, USA, bboyle[at]email.arizona.edu; [2] iPlant Collaborative, Tucson, Arizona, USA; [3] Yale University, Connecticut, USA

A cross-cutting data integration challenge for virtually every field of organismal biology is the resolution of erroneous, synonymous, or other conflicting taxonomic names. In particular, large databases of specimen observations (for example, GBIF, www.gbif.org; SpeciesLink, splink.cria.org.br), ecological inventories (VegBank, www.vegbank.org; SALVIAS, www.salvias.net), species traits (TraitNet, www.columbia.edu/cu/traitnet), gene sequences (GenBank, www.ncbi.nlm.nih.gov/Genbank) and phylogenies (TreeBASE, www.treebase.org) are plagued by high rates of taxonomic error and uncertainty. The consequences of such errors for integrative analyses based on these data are severe. Unfortunately, correcting and harmonizing taxonomy remains a time-consuming responsibility of individual investigators. The need for a sustainable, high-throughput solution to the problem of taxonomic resolution has never been more urgent.

A pilot project being developed via a collaboration between the iPlant Tree of Life project (iPToL, www.iplantcollaborative.org/grand-challenges/about-grand-challenges/current-challenges/iptol), the Botanical Information and Ecology Network (BIEN, www.nceas.ucsb.edu/featured/enquist), the Missouri Botanical Garden (MBG, www.mobot.org) and others will develop a suite of tools for automated and user-assisted standardization of taxonomic names of vascular and non-vascular plants. The iPlant Taxonomic Name Resolution Service (TNRS) will build upon and extend existing open source solutions and use established exchange standards to the fullest extent possible. The TNRS will operate both as a web interface and as a RESTful web service, capable of interactive name discovery and unsupervised batch-processing. Key features will include parsing and correction of misspelled names and authorities, conversion of synonyms to accepted names, flagging of inherently ambiguous names (e.g., *pro parte* synonyms) and the application of alternative higher classifications.

Initial data streams of reference names and synonymy will be provided by MBG's Tropicos database (www.tropicos.org), supplemented with additional digitized regional and monographic checklists such as the New York Botanical Garden's Lecythidaceae names checklist (sweetgum.nybg.org/lp/index.html). By tiling together these resources using algorithms described

below, we expect to provide nearly complete coverage for plants of the New World, with increasingly broad coverage for the rest of the world as additional sources are brought on board.

The TNRS is being developed as iPlant's "incubator project", a new, accelerated collaborative development model that brings together scientific advisers from working groups, the iPToL engagement team, members of the iPlant core services, and core software development groups.

www.iplantcollaborative.org/grand-challenges/about-grand-challenges/current-challenges/iptol-portfolio/taxonomic-name-res

9.6 The cybertaxonomy unit of the Royal Museum for Central Africa – the TDWG connection

G. Cael, F. Theeten, K. Jacobsen, S. Cooleman, L. Smirnova, J. Davy, & P. Mergen
Royal Museum for Central Africa, Tervuren, Belgium, garin.cael[at]africamuseum.be

As one of the very first European institutions, the Royal Museum for Central Africa has a distinct unit of 'Biodiversity Information and Cybertaxonomy'. It manages the institution's involvement in the ever-changing, but ongoing projects (EU, Belgian Cooperation, Belgian Science Policy, in-house, etc.). Some of which will be briefly presented here:

- SYNTHESYS and SYNTHESYS 2: SYNTHEsis of SYStematic resources 1 and 2. In total a 9 year project aims to create an integrated European infrastructure for researchers in the natural sciences.
- EDIT: European Distributed Institute of Taxonomy. A network of excellence comprising of 29 leading European, North American and Russian institutions. The overall objective is to build a world leading capacity by creating a European virtual centre of excellence, which will increase both the scientific basis and capacity for biodiversity conservation.
- CABIN: Central African Biodiversity Information Network. The aim of this project is to implement a network of databases on biodiversity information, in collaboration with several research institutions based in Central Africa and ease the integration of local researchers in networks such as GBIF (both as data consumers and data providers).
- STERNA: Semantic web-based Thematic European Reference Network Application. A 'best practice' network project supporting the objectives of the European Digital Library by pioneering the integration of semantically enriched digital resources on birds in the field of natural science, biodiversity and conservation.
- BHL and BHL-Europe: Biodiversity Heritage Library – Europe is a consortium of major natural history museum libraries, botanical libraries, and research institutions organized to digitize, serve, and preserve the legacy literature of biodiversity.
- Daubenton: A project to enable the exchange of collection managers and technicians within several leading European institutes of natural history.
- GBIF: Global Biodiversity Information Facility. An international government-initiated and funded initiative that enables free and open access to biodiversity data online.
- CETAF: Consortium of European TAxonomic Facilities is a networked consortium of scientific institutions in Europe formed to promote training, research and understanding of systematic biology and palaeobiology, Together, CETAF institutions hold very substantial biological (zoological and botanical), palaeobiological, and geological collections and provide the resource for the work of thousands of researchers in a variety of scientific disciplines.

The most important common theme in these projects is that the Biodiversity Information Standards presented each year at TDWG have always been the foundation for the technology and/or data involved in each project. This perfectly illustrates the usefulness and dissemination of TDWG standards.

9.7 Using Google Maps to Display Large Amount of Biodiversity Data

Elie Chen, Kun-Chi Lai, Elisha Hsu, Burke Chih-Jen Ko, Kwang-Tsao Shao
Taiwan Biodiversity Information Facility, Biodiversity Research Center, Academia Sinica, eliechen[at]gate.sinica.edu.tw

TaiBIF (Taiwan Biodiversity Information Facility; taibif.org.tw) is the Taiwan node of GBIF. Its main purpose is to implement at the national level the global strategy of GBIF to advance the flow of biodiversity raw data, and to provide information technology services to increase usage of information and let data publishers receive appropriate credit. TaiBIF joined the "Taiwan e-Learning and Digital Archives Program (TELDAP)" in 2007 in order to expand the integration of biodiversity data from various institutions and help with international exchanges of these data.

The majority of the species occurrence data in the TaiBIF database come from TELDAP (specimen collections of the Biosphere & Nature thematic group). Another source is the non-TELDAP institutions with their digitized specimen or observational data. The database has accumulated more than 550,000 records so far.

To integrate distributed databases, TaiBIF utilizes TapirLink as the main data-sharing tool and Darwin Core as the metadata standard for species occurrence data. TaiBIF also established a schema file to collect information recorded in Traditional Chinese characters. All the data providers who set up TapirLink not only can share data with GBIF through this schema file in the XML format, but also can meet the TaiBIF purpose of integrating Traditional Chinese data. TaiBIF has successfully promoted this kind of framework to five institutions in Taiwan.

The key back-end integration technology in the data portal of TaiBIF is TAPIR and Darwin Core. The core architecture of the front-end (user side) system is jqGrid (a jQuery plugin) and Google Maps. When species occurrence data of a species are queried, the system applies a search function to look up synonyms from the Catalogue of Life in Taiwan so that all the specimens which were given different scientific names at different times can be accessed and presented.

TaiBIF performs several functions using Google Maps:

1. Display of a large amount of data. TaiBIF converts the coordinates of species occurrence data into the grid squares of different spatial scales (scale control). The grid sizes can be 40x40 km^2, 10x10 km^2, or 2x2 km^2. Different colors are assigned to denote the ranges of data numbers in each grid.
2. Quadrilateral search. It carries out spatial search of data within a scalable quadrilateral.
3. Region search. Using the point-in-polygon algorithm, it carries out spatial search of data within a selected region.
4. Display of data across time. Using AJAX time slider, it displays data distributions of one single species or institution through a period of time.
5. Drawing of elevation profile of an area.

taibif.org.tw/?locale=en&tid=485
taibif.org.tw/taibif_search/serviceTimeSlider.php?Family=Araliaceae&institution=HAST&locale=en
taibnet.sinica.edu.tw/
140.109.29.92/tapirlink/darwinxml/darwin_c.xml
140.109.29.92/tapirlink/model/output_record.xml

9.8 The Oregon Flora Project

Thea J. Cook and Linda K. Hardison
Dept. Botany & Plant Pathology, Oregon State University, Corvallis, OR 97331-2902,
cookthe[at]science.oregonstate.edu, hardisol[at]science.oregonstate.edu

The Oregon Flora Project (OFP) is developing a flora and related data about the vascular plants of Oregon. In addition to a printed single-volume format, the OFP is presenting the flora

digitally in conjunction with extensive data from other facets. Multiple entry keys as well as dichotomous keys will allow identification of the ca. 4,500 taxa found within Oregon. The Oregon Plant Atlas provides interactive mapping of over 540,000 records of plant observations and specimens from 37 herbaria representing government agencies, small colleges, and major collections. The Photo Gallery presents field photos and herbarium specimen images. Incorporating citizen scientist input into the Atlas and Photo Gallery is a key principle for the OFP. An interactive synonymized Checklist will be available online in early 2011.

The data underlying the photo gallery, atlas and multiple entry keys hinge upon the taxon name associated with each record. Data associated with each name are stored in the checklist database, which organizes these scientific names into taxonomic concepts via concept numbers, the basis for the OFP synonymized checklist. Queries to the atlas or photo gallery utilize scientific names, but the underlying database actually retrieves all records matching the taxonomic concept of that scientific name and shows the user a list of the taxa queried. Because synonymies are in flux, we do not update scientific names throughout our datasets to currently recognized names. Instead, as the checklist database is updated, the names representing each concept are re-defined so that the OFP web interface always reflects a complete list of scientific names associated with a taxonomic concept. The online Checklist tool will also provide links to related images and a distribution map for any taxon. Please learn more about the project by visiting our website at www.oregonflora.org!

9.9 A Pan-European Species-directories Infrastructure (PESI)

Yde de Jong[1], Juliana Kouwenberg, Phillip Boegh, Mark Costello, Charles Hussey, Roger Hyam, Thierry Bourgin, Anton Güntsch, Walter Berendsohn, Ward Appeltans

[1] Zoological Museum Amsterdam, Faculty of Science – University of Amsterdam, Amsterdam, The Netherlands; yjong[at]uva.nl

The correct use of names and their relationships is essential for biodiversity management; therefore the availability of taxonomically validated, standardised nomenclatures is fundamental for biological e-infrastructures. PESI (www.eu-nomen.eu/pesi) is the next step in integrating and securing taxonomically authoritative species name registers, serving to underpin the management of biodiversity in Europe.

PESI is a joint initiative of two Networks of Excellence: EDIT (European Distributed Institute of Taxonomy, www.e-taxonomy.eu) and MarBEF (Marine Biodiversity and Ecosystem Functioning, www.marbef.org), funded by the European Commission under the Seventh Framework Capacities Work Programme – Research Infrastructures – and is led by the University of Amsterdam. It was started in May 2008 and will last three years, involving 40 partner organisations from 26 countries and several non-contracted associated partners.

PESI defines and coordinates strategies to integrate the infrastructural components of four major community networks on taxonomic indexing and their respective knowledge (social and technical) infrastructures; those of marine life, terrestrial plants, fungi and animals, into a joint work programme. These include the three main all-taxon registers in Europe, namely the European Register of Marine Species (www.marbef.org/data/erms.php), Fauna Europaea (www.faunaeur.org), and Euro+Med PlantBase (www.emplantbase.org/home.html) in coordination with EU-based nomenclators, i.e., Index Fungorum, International Plant Names Index, and AlgaeBase, plus the network of EU-based Global Species Databases (GSDs).

The integration of social expertise networks will result in functional knowledge systems of taxonomic experts and regional focal points, which will collaborate on the establishment of standardised and authoritative taxonomic data and the development of approaches for their long-term sustainability. The sustainability of these taxonomic expert networks is considered the most threatening issue to PESI's success, since Europe is experiencing a decline in the number of professional taxonomists. PESI is addressing this concern by advancing the abilities of the Society for

the Management of Electronic Biodiversity Data (SMEBD, www.smebd.eu), by collaborating with the EDIT project and the Consortium of European Taxonomic Facilities (CETAF, www.cetaf.org), as well as reaching out to non-professional taxonomists and taxonomic societies to revive this vital science.

The technical integration of these checklists into a joint 'European Taxonomic Backbone' relies on the Common Data Model (CDM), ensuring the conceptual mapping of taxonomic databases. This is hosted in the CDM store as a denormalised relational database management system (the so-called 'PESI data warehouse'). The CDM represents a component of EDIT's Cybertaxonomy Platform.

PESI is also involved in supporting international efforts on the development of the 'Global Names Architecture' by building a common intelligent name-matching device in consultation with principal initiatives like GBIF and LifeWatch (www.lifewatch.eu). This provides a unified cross-reference system to all stakeholders optimising their taxonomic meta-data service functioning.

9.10 An Intelligent Wiki for Supporting Collaborative Research

Adam Eck, Derrick Lam, and Leen-Kiat Soh
Department of Computer Science and Engineering, University of Nebraska-Lincoln, USA, aeck[at]cse.unl.edu, dlam[at]cse.unl.edu, lksoh[at]cse.unl.edu

In the day-to-day operation of a scientific research group, members routinely perform collaborative tasks such as reporting the results of experiments, summarizing existing literature, refining descriptions of particular concepts and ideas, and analyzing the contributions of other members. One common tool gaining traction within the scientific community for supporting such collaborative activities is the wiki. However, traditional wiki software was designed to support tasks for a generalized audience and not the specific needs of researchers, such as semantic modeling and linking of content, support for publications and drafts of shared documents, advanced search functionality across multiple types and sources of information, and integration of wiki content with real-world data and experimental results. Provided within the Biofinity project at the University of Nebraska-Lincoln, we introduce the Biofinity Intelligent Wiki which aims to enhance the traditional wiki with intelligent support for the needs of collaborating scientists.

Specifically, the Biofinity Intelligent Wiki currently supports standard collaborative wiki activities such as creating, editing, and deleting pages; attaching media and other files to pages; and a search engine for finding content within pages. However, the basic features of the wiki have been augmented with emerging Web 2.0 collaborative technologies, including: 1) tagging pages with searchable keywords, forming semantic links between related pages; 2) page and peer ratings from system users, providing an explicit and easy-to-understand quality metric; 3) threaded discussions through comments, and 4) sharing pages both privately between users within the system, as well as publically through social networking tools such as Twitter and Facebook. Our wiki also supports multiple types of pages each with specialized user interface and content, including both standard wiki pages and shared documents (e.g., publications, notes, working drafts). Furthermore, we have implemented intelligent features designed to allow adaptive software agents to provide tailored support to the needs and activities of individual users, including: 1) user modeling, based on unobtrusive, implicit tracking of user activities and actions; 2) page modeling, based on the keyword content and user ratings of pages; and 3) automated recommendations of pages for users to view based on both the page itself and the past history of Wiki users.

In the near future, we plan to extend the novel functionalities of the Biofinity Intelligent Wiki to better support collaborative research, including 1) introducing two new page types – data pages and results pages – linked to and describing content stored in the Biofinity database containing data from a diverse range of scientific disciplines and the outputs of computational biology tools, respectively; 2) matchmaking users to form collaborative groups based on user activities, expertise, and tasks; 3) enhance page recommendations to match pages with users containing the expertise

and willingness to revise and improve existing pages; and 4) expand search to also use external sources used by the Biofinity project, such as Encyclopedia of Life and GBIF. In our longer term vision, we also want to use our intelligent wiki as a tool for automating ontology mining and refinement, as well as providing user training in both system usage and research practices.

9.11 Parallel processing, horizontal scaling and data restructuring – the architecture and infrastructure of LigerCat.

Anthony Goddard[1], Ryan Schenk, Catherine N. Norton[1], Holly Miller[1]
[1] Marine Biological Institute, Woods Hole, Massachusetts, USA, agoddard[at]mbl.edu

MEDLINE is a database of citations and abstracts in the diverse fields of medicine curated by the National Library of Medicine® (NLM). As part of the indexing process within MEDLINE, key words from a medical controlled vocabulary called MeSH descriptors are associated with each article. The web interface for accessing MEDLINE is PubMed (pubmed.org) developed by the National Center for Biotechnology Information (NCBI) at the NLM. In order to offer an intuitive way to explore the biomedical research in PubMed we created LigerCat (ligercat.ubio.org). This web application allows users to perform a keyword search for articles, then explore the resulting set via a tag cloud of the key words (MeSH descriptors). In addition to keywords, LigerCat's gene search allows users to enter a molecular sequence to search the PubMed database for articles related to that sequence. Lastly, users can search NLM's Journals database to find relevant journal titles then explore the keywords in articles published in those journals. In all cases, the search result set is displayed as an occurrence-weighted tag cloud, with the terms associated more articles presented in a larger font size. LigerCat can be used to identify the occurrence or relationship of taxonomic names to medical literature as demonstrated in Encyclopedia of Life (EOL) project; key words pertaining to a chosen species are displayed in the Biomedical Terms section of the relevant EOL webpage. Given the amount of data required to calculate tag clouds, a fast and scalable architecture is required. To facilitate this, LigerCat utilizes a number of open source tools to process data in parallel. By designing a modular, queue-based architecture, we were able to eliminate single points of failure in the system while also gaining the benefit of horizontal scalability. The modular nature of the architecture allowed us to replace a traditional relational database system with the Redis in-memory key-value store (code.google.com/p/redis) affording us a five-fold decrease in processing times, though at the cost of an increased memory requirement. By modularizing and virtualizing components of the architecture, we were able to scale processing across many processor cores of a single or many hosts or out onto public clouds where required. Scaling can occur in parallel and rapidly due to the small footprint of the processing components. LigerCat was developed using the Ruby on Rails framework, which allowed for rapid prototyping, development, testing and integration of the system's various components. The use of messaging queues, in-memory key-value stores and related tools offers informatics developers new solutions to existing data management and processing challenges. The modular nature enables components of this existing system to be reused in other research projects which are faced with similar challenges.

9.12 Scatter, Gather & Reconcile: assembling and annotating specimen data records.

Íñigo Granzow-de la Cerda[1], Rod Spears[2] and [2]James Beach
[1] Departament de Biologia Animal, Biologia Vegetal i Ecologia, Universitat Autònoma de Barcelona, Spain, inyigo.delacerda[at]uab.cat;[2] Biodiversity Institute, University of Kansas, Lawrence, USA, rods[at]ku.edu; beach[at]ku.edu

The efficient acquisition of complete and accurate botanical specimen data records is a universal objective of herbarium computerization initiatives. As part of a project to computerize the

Mexican plant specimens held by the University of Michigan Herbarium ("Mex@MICH"), we developed semi-automated workflows which include: imaging plant specimens and labels, minimal specimen label data entry for indexing and sorting, efficient full data entry methods, searching remote databases for specimen duplicate records, and finally reconciling matching specimen information from external sources to assemble complete, authoritative records.

Analyzing discrete steps of herbarium specimen data entry workflows enabled us to optimize specimen handling logistics, image acquisition and validation of data. Using a 'minimal data entry method' we executed string searches that matched partial Mex@MICH database records with database entries in SNIB (Sistema Nacional de Información sobre Biodiversidad de México) and GBIF. That analysis revealed a number of properties about variation in the level of completeness and usefulness of specimen information in GBIF. We intend to generalize the Mex@MICH botanical workflow—starting with image acquisition and ending with the reconciliation of data from online duplicate specimen records—and then to embed those methods into Specify 6 Software.

Network connectivity, federated query protocols, and community specimen caches have greatly advanced the accessibility of biological specimen data for discovery and retrieval. With Specify 6 we will use that infrastructure to bring botanical specimen data integration to a new level in order to benefit data acquisition, curation, and annotation methods within Specify collections.

This project is supported by U.S. NSF Research Collections Program grants 0138621 and 0646301 to the University of Michigan and University of Kansas.

9.13 The Cost of taking short-cuts in Data Entry – Finding a compromise between minimal data and time availability

Henry Engledow
henry.engledow[at]br.fgov.be, Belgium

Taking short-cuts during data entry can be costly to an institution in terms of time and money in the long term. This is due to the creation of duplicate and erroneous records mainly resulting from entering minimal data. This results in unforced errors due to the lack of information. Records most at risk are those found in subordinate tables as opposed to more central tables. The bibliographic references table was used as a test case. 1500 primary records (records that have never been edited) were examined with respect to their percentage completeness against a basic list of fields. The results varied depending on the type of bibliographic reference, e.g., books versus journals. For example, only 8 % of the book references were complete, 54 % could be tracked down using their date and abbreviation, while 38 % would need to be traced back to related taxa and / or specimen records to ascertain the specific reference. The latter exercise is extremely time consuming and could be avoided by entering in the original data more completely. To analyse more specifically the percentage and type of data errors and their probable causes, 500 journal records containing abbreviations, but lacking a link to the related journal record, were examined. These records were cleaned-up and basic information entered dependent on the type of reference material. The data in the original and cleaned datasets were compared: 6 % of the records were duplicates; 52 % had wrongly designated reference types; 50 % had wrong or non-standard abbreviations; while errors in the remaining fields varied between 3 and 6 %. Most of these errors could have been avoided by entering more complete basic information and by using related tables. The primary reason given for lack of completeness was insufficient time, as data entry is often in the framework of projects with deadlines. The time needed to enter information in these related tables is often not taken into account when setting achievable goals for project proposals. Taking the latter into account would lead to more reliable data, and save the institution time and money in the long run.

9.14 The GBIF Community Site: a social networking platform for Biodiversity Informaticians

Alberto González-Talaván
Global Biodiversity Information Facility Secretariat, Copenhagen, Denmark atalavan[at]gbif.org

The Global Biodiversity Information Facility (GBIF) launched in August 2010 an online collaborative environment called the 'GBIF Community Site' (community.gbif.org) aimed to promote and improve communication within the Biodiversity Informatics Community in general, and among the GBIF stakeholders in particular.

The site serves as a platform for collaborative projects, discussions, sharing of information and expertise, making announcements, requesting/providing guidance, etc.

A basic set of tools has been made available for the launch: collaborative groups, a news system, online chat, forums, messaging, (micro-)blogging and file/image/bookmark sharing. OpenID-based login, email notifications and direct publishing to Twitter are additional features that allow integration with other systems that will be expanded in the future.

But, apart from providing a selection of tools, the greatest potential of social platforms is their capacity to enable networking and make possible distant collaborative efforts difficult to achieve with previous systems: personal profiles permit finding of potential partners and services; public posts and the user comment system facilitate external contributions; the 'followers' system and the RSS feeds make it easy to keep up-to-date on what is new in the field of biodiversity informatics. Everything is integrated in a single, professional environment focused on the topics of your interest.

The software used to build the GBIF Community Site is ELGG (www.elgg.org), an Open Source, PHP-based platform with strong support from a large community of users and developers. This system and the additional tools developed for GBIF are available for others willing to deploy their own social platforms.

The GBIF Secretariat would like to invite all members of the TDWG community to visit the site, create accounts and make the most of this excellent networking opportunity!

9.15 The Identity of Things – Why we need transparent delimitations of biodiversity objects

Anton Güntsch[1], Walter G. Berendsohn[1], Marc Geoffroy, Eckhard von Raab-Straube, Andreas Müller, Ward Appeltans, Yde de Jong
[1] Botanic Garden and Botanical Museum Berlin-Dahlem, Germany, a.guentsch[at]bgbm.org

The assignment of Globally Unique Identifiers (GUIDs) to information items networked in the emerging Biodiversity Informatics Infrastructure is largely regarded as a technical issue to be solved by database managers and developers. In fact, several facets of GUID creation, management, and resolving need sound technical solutions and standardization, and TDWG has taken a leading role in this process for some years (wiki.tdwg.org/GUID). However, we believe that the deployment of GUIDs in an increasingly web-based environment has aspects that go beyond the technical implementation and need attention by the communities creating, maintaining, and owning the data.

In particular, the question of which operations on a given object (e.g., taxon, scientific name, author, collection site, specimen, vernacular name, multimedia object) turn it into a new object so that a new GUID has to be assigned has never been sufficiently discussed. This is due to the complexity of the problem and the diversity of notions of the “correct” delimitation of objects. At the same time, biodiversity data providers issue GUIDs without a clear delimitation strategy, and

consumers of the associated services will not know whether an object bound to a given GUID does change over time and whether a statement about the relation between objects maintains its validity.

The Pan-European Species directories Infrastructure (PESI, www.eu-nomen.eu/pesi) is a European Union 7th framework project building an integrated infrastructure for the major European taxonomic checklists. PESI recognized the importance of a joint strategy for the assignment of GUIDs to its primary checklist elements (names and taxa) and implements rudimentary mechanisms for propagating identifiers from the initial data provider to the PESI portal and web services. This includes a documented set of rules supporting the decision on whether or not new objects and GUIDs should be created from existing ones (www.eu-nomen.eu/pesi/index.php?option=com_remository&Itemid=56&func=fileinfo&id=767).

The PESI example demonstrates that comparably little effort is needed to increase the stability and value of GUID systems for biodiversity sciences as long as the underlying delimitation strategies are properly documented.

9.16 Key-value stores: a nontraditional approach to managing metadata

Chuck Ha, Holly Miller, and Catherine N. Norton
Marine Biological Laboratory, Woods Hole, Massachusetts, USA, cha[at]mbl.edu

Many scientists struggle with managing disorganized data and metadata. Metadata are often in a format that is difficult to use, forcing many people to resort to manual editing in order to comply with today's data standards. This task is tedious, rigid, time consuming and when the standards change, the metadata becomes obsolete unless updated.

Most metadata are organized in one of two ways: spreadsheets or traditional databases. While spreadsheets are flexible, they often contain many header columns and blank values, which make it difficult to access the raw data in a meaningful way. Databases provide a centralized location, but are not flexible enough to accommodate the multiple spreadsheets and varying headers that often comprise sets of metadata.

Key-value stores are a more flexible style of database that allow for the many headers that are present in metadata spreadsheets. They not only automate the process of managing metadata, thus requiring far fewer human resources, but are also an excellent synthesis of these two organizational systems, providing both the flexibility of a spreadsheet and the permanence of a database. Storing the raw data in an easily-accessible place allows generic programs to be written and reused and creates a centralized place to access the data instead of files on a server.

The advantages of a flexible key-value store are demonstrated in the use of CouchDB (couchdb.apache.org) to manage the data for 19,000 mouse strains collected from web-accessible structured databases. The goal of the project is to create a resource for scientists studying the biology of aging. However data gathered from more than 5 different sources are highly variable and user requirements are still vague. A traditional database would be inefficient because of data manipulation required to enter values, but using a key-value store allowed us to enter raw data and store it without additional modification. The project in progress can be viewed at mouse.ubio.org.

The features of key-value store that were useful creating the Mouse Strain Database (mouse.ubio.org) would also be valuable in many biodiversity projects where a structured schema isn't possible, when data is being aggregated from many sources where standards differ, or at a stage during research and development that a structured schema isn't ready.

9.17 CollectionSpace: A community source collection management system for natural history collections and beyond

Christopher R. Hoffman
IST-Data Services, University of California, Berkeley, USA; chris.hoffman[at]berkeley.edu

CollectionSpace (www.collectionspace.org) is an open source, web-based software application for the description, management, and dissemination of museum collections information. UC Berkeley has selected CollectionSpace as the strategic platform for museum collections across campus due to its modular design, its capacity for customization and interoperability, its flexible hosting requirements, and most importantly its ability to support collections-based research, education, and public service. This poster will describe UC Berkeley's approach to data migrations, customizations, and full deployments of the system, featuring our deployment work for two members of the Berkeley Natural History Museum consortium (the University & Jepson Herbaria and the Phoebe A. Hearst Museum of Anthropology). As the project team approaches the release of version 1.0 of CollectionSpace, UC Berkeley and the other CollectionSpace partners are reaching out to other institutions and individuals in order to build a strong, collaborative community that can sustain CollectionSpace into the future. The Berkeley Natural History Museums are working together with the CollectionSpace project to build a set of templates and documentation that will help other natural history collections deploy CollectionSpace.

9.18 OGC Web Services for GBIF-Mediated Occurrence Data

Jörg Holetschek[1], Tim Robertson, Éamonn Ó Tuama
[1] Biodiversity Informatics and Laboratories, Botanic Garden & Botanical Museum Berlin-Dahlem, Germany; j.holetschek[at]bgbm.org

International networks and initiatives such as GBIF, BioCASE (www.biocase.org), and SYNTHESYS (www.synthesys.info) share the vision of free and open access to the world's primary biodiversity data, stored in a large number of databases worldwide. Observations and specimen data – from living collections as well as preserved specimens – are linked together, forming a huge number of occurrence records, each documenting the occurrence of one specimen at a given location at a certain point in time. Currently, the GBIF index lists 203 million occurrence records (2010-08-23).

The GBIF data portal can be used by scientists to find records of interest, for example for a certain taxonomic group or geographic region. However, the sheer number of records can make this a cumbersome task. Visualising the occurrences' geospatial information, i.e., the coordinates, offers another view on the data. Despite the multitude of data, browsing becomes feasible again by using maps; for taxonomic groups with a sufficient record pool, distribution maps can be created and the underlying data be examined.

Yet the number of records remains a demanding challenge. At present, 168 million GBIF records are georeferenced. Recent geospatial visualisation tools such as GeoServer (geoserver.org) and MapServer (mapserver.org) are swamped by such numbers; even if run on superior hardware and restricted to a certain area, on-the-fly creation of maps would be too slow to allow for convenient browsing. Different views of the data are required, depending on the user's map scale. Drilling down into the data details can then be achieved by zooming into a certain area of the map.

To enable this, we created cluster maps with different resolutions, using cluster cells (i.e., grid cells) ranging from one degree to 0.01 degrees in which individual occurrences are aggregated. We linked these cluster maps to zoom-based style descriptors and combined them into layer groups together with detailed views of the occurrence records. Thus, we were able to create Open Geospatial Consortium (OGC) web services (Web Feature Service, WFS, and Web Map Service, WMS)

that offer zoom-based views of the data; at large scales, they provide a coarse-grained view, at high zoom levels the full record details are presented. The layer groups can be used in conjunction with Google Maps, Google Earth or OpenLayers and can also be integrated into any other GIS application (Geographic Information System).

To allow taxonomic filtering, a second set of web services was implemented. It accepts a CQL (Common Query Language) filter on any of the ranks of kingdom, phylum, class, order, family, genus and species. In order to sustain fast map creation for user convenience, the clusters are precalculated for all taxonomic groups. Clustered Indexes were used on the precalculation results to ensure low access times despite the huge amount of data.

The initial web services were set up to deliver GBIF mediated African biodiversity data to a GIS client developed as part of the GEO (Group on Earth Observations) Protected Areas Assessment and Monitoring Pilot (GPAAMP), and also as a contribution to EuroGEOSS (www.eurogeoss.eu) which is developing an initial operating capacity for biodiversity as a European contribution to GEOSS (Global Earth Observation System of Systems). However, we think that the devised schema can be used for global occurrence data and also cope with growing record numbers in the future.

9.19 Transforming Citizen Science with the Biofinity Project

Mary Liz Jameson[1] and Bill Welch

[1] Department of Biological Sciences, Wichita State University, KS, USA; maryliz.jameson[at]gmail.com

The Biofinity Project (biofinity.unl.edu) is a free web-based repository for biodiversity data and tools designed to support research in the biological sciences. This project aims to empower investigation and discovery across the sciences and to transform the way that we access and analyze biodiversity data.

Tools provided by The Biofinity Project, such as mobile iPhone data-integration, My Labs collaborations, social networking applications, geo-tagging, and mapping have the capability of greatly advancing citizen science. We apply the applications and tools provided by The Biofinity Project in high school biodiversity studies that are aimed at engaging students in the biological sciences.

High school biodiversity surveys conducted by students and mentors included sampling a riparian habitat for fish and turtles, comparison of terrestrial habitats for arthropods, and survey of plants on restored versus virgin prairies. Organisms were collected, counted, measured/weighed, identified to the lowest possible level, imaged, and locality coordinates recorded. Using My Labs in The Biofinity Project, students and mentors entered and processed specimen records and associated data. Specimens that were imaged using The Biofinity Project's BioBlitz mobile application for smart phones were instantly geo-tagged using the Twitter location feature and data imported into The Biofinity Project web site. Using mobile applications, specimen data are mapped onto Google Maps, taxon lists are generated, and on-the-fly "field guides" created. Students and mentors can verify identifications, thus providing immediate feedback and an enhanced learning activity. Survey data are available through The Biofinity Project website and can be exported as Excel or XML archives.

Future endeavors will focus on use of "microchips" or PIT (Passive Integrated Transponder) tags for identification and monitoring of individual turtles in order to examine dispersal patterns, longevity, and recruitment. Data access and mapping capabilities in The Biofinity Project afford new ways of analyzing these data.

Functionality and tools provided by The Biofinity Project empower students and mentors to learn about local biodiversity, assist in identifying sites where invasive species may need to be monitored or controlled, and in identifying sites where rare or unique native species are found.

9.20 Hymenoptera Online iOS application for taxonomy

Norman F. Johnson, Joe Cora, Luciana Musetti
Department of Evolution, Ecology and Organismal Biology, The Ohio State University, USA, johnson.2[at]osu.edu

We present the first mobile taxonomic resource focusing on insect taxonomy for the Apple iOS operating system. This works for both the iPhone and iPad in a single app. Cellular data networks can be found in some of the most remote places on Earth, and by tapping into this wireless conduit, the lab and the field are merged into a single, more effective mobile research platform. The app provides untethered access to the Hymenoptera Online (HOL) database with an intuitive user interface. Resources available include a complete annotated taxonomic catalog, pdf library of taxonomic publications, primary occurrence data for specimens, mapping functionality, biological associations, and images. It avoids the inconsistent and time consuming need to load web pages in inadequate mobile web browsers. With the aid of location services present on the device, the HOL app presents nearby collecting localities as well as driving directions to the location to facilitate collecting efforts. Specimen images are accessible and can be enlarged with a simple pinch gesture to allow for real-time specimen identification. Habitat and biological association information aid the collector in pin-pointing a spot for focusing his/her time in an efficient manner.

9.21 Modeling and Publishing Biological Names and Classifications on the Semantic Web

Nina Laurenne, Jouni Tuominen, Mikko Koho and Eero Hyvönen
Semantic Computing Research Group (SeCo), Aalto University, Department of Media Technology, Finland, laurenne[at]cc.helsinki.fi

Periodic changes characterize the scientific naming system. As a result, the biggest challenge lies in ascertaining the actual meaning of names, when multiple taxonomic concepts are associated with them. This makes it hard to integrate biological information from different sources, such as publications, online databases, and museum collections, and search for it. On the Semantic Web, the problem can be approached by representing taxa, checklists, and their relations as ontologies that are decipherable for machines. Our goal is to establish a centralized ontology repository of biological names and classifications in Finland.

We have developed an ontology model for biological taxon names and the ONKI Ontology Service for publishing it as a service for humans and machines. The aim is to achieve a practical and maintainable name system for researchers, environmental authorities, and amateurs to find biological names to use, index content correctly and cost-efficiently using ontology services, and to lay out a foundation for making heterogeneous biological content interoperable in applications.

The model consists of three parts on the basis of the elaborateness of taxonomic information and the needs of the users. The parts are maintainable independently, but associations between them are possible. (1) Scientific names and taxonomic concepts are treated separately and detailed taxonomic information can be associated with them. The processes of systematic research that are relevant to changes of taxon names or concepts such as descriptions of new taxa, splitting and lumping of taxa are conceptualized. (2) Checklist-type information is supported; names occurring in different checklists, but referring to the same taxon can be linked. (3) Also, vernacular names in multiple languages including dialects are supported. The ontology model covers temporal dimensions, which make taxon names traceable to reveal conflicting taxonomies and competing views. The possibility to connect imprecise taxonomic knowledge to precise information allows for versatile and flexible name management for users with different needs. Results of queries do not only return a currently valid/accepted name but lead the user to the source of the information.

Ontology-based queries enable the retrieval of relevant contradictory information, which is an important feature for scientists.

We have two use cases regarding beetles to demonstrate the usage of the name ontology. (1) Cerambycid beetle names of five Finnish checklists from the years 1936-2010 are linked. The ontology is applied to the observational data approximately from the same period. Geographic information on the observational data is then disambiguated using Finnish Spatio-temporal Ontology (SAPO). By applying name ontology and SAPO we can explore the distribution of cerambycid beetles in the time-scale without extensive data harmonisation. (2) The other use case is nine genera of Afro-tropical eucnemid beetles and their chaotic classification. The model will be tested with the pilot group, in which splitting and lumping are common and names have changed for various reasons. For example, at least eight taxonomic concepts are associated with the genus *Pterotarsus*. This data is challenging as it includes study results, mistakes, and various nomenclatural changes.

This work is part of the national FinnONTO program and is funded by the Finnish Funding Agency for Technology and Innovation (Tekes) and a consortium of 38 public organizations and companies. Fruitful co-operation with Jyrki Muona, Hans Silfverberg, Hannu Saarenmaa, Leo Junikka, and the Finnish Museum of Natural History is acknowledged.

Semantic Computing Research Group: www.seco.tkk.fi/
National Semantic Web Ontology Project in Finland (FinnONTO), 2003-2012: www.seco.tkk.fi/projects/finnonto/
Biological Ontologies and Vocabularies: www.seco.tkk.fi/ontologies/biology/
ONKI Ontology Service: www.onki.fi
Finnish Spatio-temporal Ontology SAPO: www.seco.tkk.fi/ontologies/sapo/

9.22 United States Virtual Herbarium: Providing integrated, digital access to specimen data from U.S. herbaria.

Ben S. Legler, Mary E. Barkworth, Zack E. Murrell
University of Washington, WA, USA; blegler[at]u.washington.edu

U.S. herbaria are working together to provide integrated access to information about all of the 73.4 million specimens in the nation's over 811 herbaria, which range in size from approximately 250 specimens to over 7,000,000. Achieving this goal means that each herbarium needs to image each of its specimens, capture their label information into a database, and georeference the collecting localities when possible. Information from these herbaria will be aggregated at regional levels using internationally recognized data management standards. These regional aggregations will then be integrated into what will seem, to the user, like a single national database accessible through online interfaces and web services. The U.S. Virtual Herbarium (USVH; usvirtualherbarium.org) will coordinate with other national efforts, including the Network Integrated Biocollections Alliance and the U.S. node of the Global Biodiversity Information Facility. Since its formal inception in 2007, the USVH project has been working to identify all the herbaria in the US, including those not listed in *Index Herbariorum*, to obtain data on their holdings, and to determine how many have already been databased (approximately 13.3 million) and imaged (very few), almost all of which are vascular plants. An NSF-sponsored workshop in February 2010 helped identify several issues that need to be addressed in order to accomplish the project's goal. Various task forces are now working in these areas.

9.23 The Plant List: A New Widely Accessible Working List of All Plant Species

Chuck Miller[1], Robert Magill[1], Chris Freeland[1], Alan Paton[2], Eimear Nic Lughada[2], and Robert Allkin[2]

[1] Missouri Botanical Garden, St. Louis, USA, Chuck.Miller[at]mobot.org; [2] Royal Botanic Gardens, Kew, Richmond, UK

In response to Target 1 of the Global Strategy for Plant Conservation, Missouri Botanical Garden (MBG) and Royal Botanic Gardens (RBG), Kew have embarked upon a joint effort to create a widely accessible working list of all plant species. Recognizing the reality of the taxonomic impediment caused by a world-wide shortage of plant taxonomists, new techniques were conceived to combine existing taxonomist curated global checklists with available regional and monographic data. The World Checklist of Selected Plant Families (WCSPF) at RBG Kew containing 374,000 globally synonymized plant names was used as a foundation. The Tropicos® database at MBG containing many regional checklists was synthesized with WCSPF and other data from RBG Kew, the new Compositae checklist, and ILDIS (International Legume Database and Information Service, www.ildis.org) in a novel heuristically-driven computerized process. The result is a consensus synonymized list of 1.3 million plant names, including angiosperms, gymnosperms, pteridophytes and bryophytes, named The Plant List (TPL). TPL will be accessible online by the end of 2010.

9.24 An Event Model for Herbarium Specimen Data in XML

William E. Moen[1], Amanda K. Neill[2], and Jason Best [2]

[1] Texas Center for Digital Knowledge, University of North Texas, william.moen[at]unt.edu;
[2] Botanical Research Institute of Texas, aneill[at]brit.org, jbest[at]brit.org

The Apiary Project (www.apiaryproject.org), a collaboration of the Texas Center for Digital Knowledge at the University of North Texas and the Botanical Research Institute of Texas, is building a framework and web-based workflow for the extraction and parsing of herbarium specimen data. The workflow will support the transformation of written or printed specimen data into a high-quality machine-processable XML format. This poster describes an event model that informed the development of the Apiary XML Application Schema.

The Apiary Project is not developing an overall reference or data model for biological data. It has a more narrow focus, namely modeling the data on a herbarium specimen sheet to inform an extension to the Generic Darwin Core (DwC) XML Schema. The extension will need to accommodate all events that change specimen object metadata elements over time, including new identifications, ownership changes, and other non-taxonomic annotations. DwC is the foundational vocabulary for the Apiary Project, and we have defined additional Apiary-specific vocabulary terms to address the requirements of herbarium specimen sheets. The latter is now represented in a Generic Apiary XML Schema (www.apiaryproject.org/documents). Modeling the specimen data, however, was necessary to inform the development of the Apiary Application XML Schema, which imports the Generic DwC and Generic Apiary schemas and defines the structure of the Apiary XML record. Schemas are available at: www.apiaryproject.org/documents.

It is appropriate to view the information associated with the herbarium specimen object, including the sheet to which it is attached, as dynamic over time. While some information is unchanging (e.g., Collection and Occurrence information), other information can change after curatorial actions and research use—and the changes are traditionally manifest as additional marks or labels on the specimen sheet. The implication for the Apiary XML Schema is that it must accommodate the addition of information over time in a normative way.

For purposes of modeling, we propose the following two events:

- Collection event: data about this comes from the primary label and includes information about the collection (who/where/when) of the plant glued to the sheet.

This event and associated data have a one-to-one relationship with the Specimen Object.

- Annotation event: data about these come from identifications (determinations), curatorial actions related to processing and ownership changes, and other additions of information modifying what we know about the plant glued to the sheet, or what we know about the sheet itself. The Specimen Object and Annotation Event have a one-to-many relationship.

The resulting Apiary Application XML Schema reflects primarily the focus on annotations. We define annotation very broadly and Annotation Event accommodates multiple annotation events related to adding information to specimens (and in some cases information about the object that may not be present on the specimen sheet). Since there is only one collection event for a specimen object, the collection event *qua* event does not need to be represented in the XML Schema.

Our conclusion is that an event model accommodates the dynamic nature of information associated with a specimen object over time and may have utility in consideration of filtered-push. The Apiary Application XML Schema provides the structure to capture the creation, evolution, and transition of specimen objects.

A Framework and Workflow for Extraction and Parsing of Herbarium Specimen Data: www.tdwg.org/proceedings/article/view/567

An Application Profile Using Darwin Core Rendered in the New Dublin Core Application Profile Framework: www.tdwg.org/proceedings/article/view/537

The Apiary Project is funded by a National Leadership Grant (LG-06-08-0079) from the U.S. Federal Institute of Museum and Library Services.

9.25 Making Biological observing data "talk" amongst the data sets and making it accessible and reproducible

Hassan Moustahfid[1], Philip Goldstein[2], Charles Alexander[1]

[1] Hassan Moustahfid, NOAA/IOOS, Silver Spring, MD, USA, Hassan.Moustahfid[at]noaa.gov; [2] Philip Goldstein, OBIS-USA, University of Colorado Boulder, CO, USA, Philip.Goldstein[at]Colorado.EDU

Datasets generated in scientific surveys conducted by fisheries agencies or via other surveys (e.g., fishery-independent data) are important information for agencies-specific stock/population assessment, fisheries management and research. Dissemination of these types of data has often been limited to the organizations conducting the surveys. Effort has been made to provide informatics services and expose these data to a wide community (e.g., Ocean Biogeographic Information System (OBIS), and Global Biodiversity Information Facility (GBIF)). Despite accomplishments in technology and community, and the increased accessibility of fishery-independent data, their scientific acceptability is often limited by a lack of reproducibility in data analyses and because many applications require richer data than currently supported by community services. For example surveys data are highly heterogeneous and the variety of formats, logical structures, and sampling methods in fishery independent data create significant challenges. Here we describe an informatics framework for fishery-independent data that will expand information content and reconcile standards for the representation and integration of these biological observations for users to maximize the value of these observing data. We further propose that the approach described can be applied to other datasets generated in scientific surveys and will provide a vehicle for wider dissemination of biological observing data. We propose to employ data definition conventions that are well understood in NOAA/IOOS (U.S. National Oceanic and Atmospheric Administration, Integrated Ocean Observing System) and to combine these with ratified Darwin Core terminology, policies and guidelines.

9.26 Reusable biodiversity informatics tools

Dmitry Mozzherin, Patrick Leary, Anna Shipunov, Alexey Shipunov
Encyclopedia of Life, <dmozzherin[at]eol.org

Many biodiversity informatics web-based projects provide Application Programming Interfaces (APIs), which allow outside users to take advantage of the projects' functionality. However when operating on large datasets, network communication often becomes a bottleneck. In such cases, using software installed on the client side can significantly improve processing times.

During the development of some Global Names Architecture (GNA) applications such as Global Names Index (GNI) or Global Names Integrated Taxonomic Editor (GNITE) the authors have created a set of components for processing scientific names. These components are created in the Ruby programming language (they are called 'gems' in Ruby) and are very easy to install on various platforms and operating systems. We also provide wrappers that allow these components to be accessed from other programming languages.

Scientific Name Parser (biodiversity gem, rubygems.org/gems/biodiversity): It is unthinkable to start any automatic treatment of biodiversity information without first finding the semantic elements of scientific names. Scientific names often have a complex structure that can be hard to parse in an automated fashion. This component is able to evaluate scientific names with complex structure including multiple authorships, names of hybrids, etc. The underlying technology used in the parser could be adapted to strictly enforce the codes in order to create a nomenclatural code verification tool to support submission of newly described species. In addition to the standard Ruby interface, there are also command line and socket interfaces which allow it to be used with other languages.

Taxonomic matcher (taxamatch_rb gem, rubygems.org/gems/taxamatch_rb): This component uses algorithms developed by Tony Rees to decide if two name strings are variants of the same scientific name. The core of this gem is written in the C programming language to increase performance. This component can either be optimized for convenience (rapid development) or for performance (processing large datasets and caching the results).

Darwin Core Archive Tool (dwc-archive gem, rubygems.org/gems/dwc-archive): A component to handle creation of new or reading of existing Darwin Core Archive files.

Developers working on biodiversity related projects can use flexible, reusable, open source components such as these to quickly recombine the functionality in various meaningful ways. These components can be published in online software repositories, allowing public discovery of and access to these tools. Such 'small' tools are being developed by many biodiversity groups in several programming languages. One downside of this approach is that currently finding such tools is not a trivial task. The authors are interested in helping to develop a shared registry or repository of such tools for the benefit of the broader biodiversity informatics community.

This work was sponsored by grants for the Encyclopedia of Life (MacArthur & Sloan) and the Data Conservancy (NSF).

An example of reusing the parser tool: www.marinespecies.org/aphia.php?p=match
biodiversity parser online: gni.globalnames.org/parsers/new
taxamatch original: www.cmar.csiro.au/datacentre/taxamatch.htm
taxamatch-webservice by GBIF: code.google.com/p/taxamatch-webservice/

9.27 The Biofinity Project: Providing Next-generation Tools for Biodiversity Research

Federico Ocampo[1], Ma. Celeste Alvarez-Bohle[2], and Belén Maldonado[1]

[1] Instituto de Investigaciones de Zonas Áridas, Mendoza, Argentina, federico.ocampo[at]gmail.com; [2] Laboratorio de Entomología, Facultad de Ciencias Naturales, Universidad Nacional del Nordeste, Corrientes, Argentina.

The Biofinity Project (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 biodiversity information and genomics and provides support for inclusion of external data regardless of format. The repository allows scientists and, to certain level, the public) 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).

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 Biofinity Project features My Labs (http://biofinity.unl.edu/biofinity/lab/info) an integrated application that permits collection of data, classify results, and map located specimens. My Labs is a unique feature of The Biofinity Project that provides online hosting of biological systems databases and tools for collaborative research across multiples users and/or institutions. Using My Labs, users can easily enter, manipulate, publish, map, integrate, and process their data in an easy-to-use, secure web environment. This system features integrated Google Maps for geo-coded specimen distribution mapping, and data can be shared via KML export.

We showcase our research on scarab beetle biodiversity using Biofinity and My Labs.

9.28 Validation, analysis and aggregation of long-term trait observation data of genebank material

Markus Oppermann[1], Jens Keilwagen, Helmut Knüpffer, Swetlana Friedel and Andreas Börner

[1] Institute of Plant Genetics and Crop Plant Research (IPK), Germany, markus.oppermann[at]ipk-gatersleben.de

The Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany, has, like many other collections of plant genetic resources (gene banks), over many years carried out observations of characterisation and evaluation traits of its accessions during seed multiplication of its accessions. The electronically recorded data are, however, very heterogeneous and may contain errors due to their manual registration. The aim of the present research is to make these data suitable for further processing (e.g., statistical analysis) and to identify potentially erroneous data. For this aim, a suitable workflow was developed, which is presented here for the traits “flowering time” and “thousand grain weight” in barley germplasm.

The pre-processing of data by a syntactic error correction and the subsequent application of statistical methods to detect outliers and check plausibility leads directly to an improvement of the data quality. Since the accessions grown in a particular year do not necessarily represent a statistical

sample of the complete collection, and since the cultivation is carried out without repetitions, the statistical evaluation methods usually applied in plant breeding cannot be applied here.

Therefore the observation data of a trait within a cultivation year are transformed into ranks and then standardized to a 0-to-1 scale. The resulting normalized ranked data are then utilised for further analysis. Through this transformation, also multi-dimensional analyses can be performed to reach a small number of hits for combined search criteria. In particular, statements about the "best fit" or "worst fit" for single traits or any combination thereof can be made. It is expected that, when including weather data of the respective years of cultivation, the conclusions about the genebank accessions can be improved.

Based on this study, a project is being initiated, aimed at developing algorithms for aggregating long-term observation data on genebank material, and at integrating such analysis functionalities into IPK's online Genebank Information System.

9.29 A Georeferencing Tool to Improve Biodiversity Data Quality

Allan Koch Veiga, Antonio Mauro Saraiva, Etienne Américo Cartolano Júnior
Universidade de São Paulo, Escola Politécnica, Computing Engineering Dept., Agriculture Automation Laboratory, São Paulo, Brazil, saraiva[at]usp.br, etienne.cartolano[at]gmail.com

Data Quality plays an important role in the field of Biodiversity Informatics. Low-quality data can negatively impact models and analyses used to support strategic decisions aimed at biodiversity conservation and sustainable use.

In the context of biodiversity data quality, geospatial data are quite important. They refer to the geospatial location of occurrences, which are characterized by observations or specimen collections and can be utilized, for example, to create distribution models and density maps.

Due to the relevance of geospatial data in Biodiversity Informatics, the development of policies, methods and tools to improve quality in the digitization of geospatial data is of utmost importance. Keeping this necessity in mind, the Biodiversity Data Digitizer (BDD) tool [1] utilizes a georeferencing resource to aid in the digitization of geospatial data focusing on quality.

This web-based resource, referred to as the Biodiversity Georeferencing Tool (BGT), allows users, by means of an interactive three-dimensional map based on the Google Earth API (code.google.com/apis/earth) to obtain the names of the country, state, city latitude, longitude and altitude of a location by clicking on its position on the map. The BGT allows improvement of geospatial data quality in at least three aspects: completeness, accuracy and consistency.

The completeness of data tends to grow due to the fact that the user does not necessarily need to have the exact geographical coordinates to fill out the latitude, longitude and altitude fields. In other words, the user can locate a known region on the map, such as a park or a mountain, and utilize a more specific reference, like a river or a road, to obtain approximate geographical coordinates.

The accuracy of data may also be enhanced in some cases. When the user does not possess the exact geographical coordinates of an occurrence – obtained with a Global Positioning System (GPS) receiver – one frequent solution is to utilize the geographical coordinates of the center of mass of the city where the occurrence was registered. In this case the user can obtain more accurate coordinates if he uses a reference in the field for the occurrence.

Another aspect of Data Quality that can be enhanced with this tool is data consistency. Since geospatial data is completed automatically, BGT ensures that the user cannot enter conflicting or erroneous data, such as "Country = Brezil, State = Rio de Janeiro, City = London, Latitude = 0, Longitude = 0".

BGT was developed using PHP and JavaScript programming languages, which facilitate its integration with BDD. For map rendering and geospatial data obtainment, BGT uses the Google

Earth API and the Google Maps API Web Services, therefore requiring the installation of the Google Earth plug-in on the user's web browser.

In summary, the integration of BGT into BDD allows for improvement in biodiversity data quality within the geospatial data domain. It raises the confidence level in the results of analyses and models that will guide biodiversity conservation and sustainable use.

[1] Cartolano, E.A.J., Saraiva A.M., Teles, J.A.,Krobath, D.B., Correa, P.L.P. 2009. The Biodiversity Data Digitizer (BDD) tool. TDWG Annual Conference. Montpellier, France.

9.30 Thirty months of progress: Encyclopedia of Life Content Status, September 2010

Katja Schulz, Jennifer Hammock, and Cynthia Parr

Encyclopedia of Life, Species Pages Group, National Museum of Natural History, Smithsonian Institution, USA, schulzk[at]si.edu, hammock[at]si.edu, parrc[at]si.edu

Since its launch in February 2008 with 30 000 founding species pages, the Encyclopedia of Life (EOL) has connected with many distinguished online partners and enriched and multiplied its content. Here we summarize current content status and major recent and upcoming additions.

Presently there is vetted content available for 331 000 taxa on EOL. 1 823 000 taxa have at least a link to a specialist project available. There are images available for 77 000 taxa. One year into the EOL Marine Theme, there is vetted content available for 80 000 marine species and approximately 100 000 marine taxa altogether. There are images for 18 000 marine taxa, and text for103 000 taxa. Increasingly, newly connected content partners are enriching EOL pages with a wider variety of content. 159 000 taxa have vetted content in at least two different sections of the page. 47 000 taxa have both images and text.

Important recent and imminent content partners include the Smithsonian Institution, which has published their invertebrate zoology image collection on EOL already, with the vertebrate and entomology collections to follow shortly, and the Ocean Biogeographic Information System (OBIS) who have shared depth range data for 68 000 marine species. In addition, the International Union for Conservation of Nature (IUCN), who previously had been sharing Red List status with us for 44 000 species, will shortly be publishing a comprehensive array of text fields also.

Next steps:

- As part of our development roadmap for the coming year, we will investigate the possibility of propagating appropriate content from higher taxa on EOL to their child taxa; for instance, a general descriptive paragraph from a family page might be copied into the appropriate chapter of the genera and species within the family. Content multiplication of this kind will be particularly important for very species-rich groups where expertise and manpower is only sufficient to populate a limited number of pages, or where the only available content at the species level is highly technical. This is expected to be a significant fraction of not-yet populated species pages, possibly more than half.
- Now that connecting a museum image database has been successfully demonstrated, this should be more easily replicable at many institutions with similarly organized multimedia collections. The code which transferred the data from the Smithsonian image collection to EOL should be adaptable to other institutions using the same cataloguing software (Electronic Museum, or EMu, a very widely used system among museums internationally. If the vendor can be convinced to incorporate the EOL export as a standard feature, connecting would become very easy for a large number of institutions not yet connected.)

9.31 Veg-X – An exchange standard for plot-based vegetation data

Nick Spencer[1], Miquel De Cáceres[4], Susan Wiser[1], Robert Peet[3], Brad Boyle[5] and Martin Kleikamp[2]

[1] Landcare Research Ltd, New Zealand, SpencerN[at]landcareresearch.co.nz; [2] Bergisch-Gladbach, Germany; [3] University of North Carolina, USA; [4] Universitat de Barcelona, Spain; [5] University of Arizona, USA

Collaborative research initiatives and synthetic vegetation analysis are often limited by difficulties in sharing or combining datasets. Can we facilitate these activities by means of an exchange schema for plot-based vegetation data?

In 2003, the Ecoinformatics Working Group and the Governing Council of the International Association for Vegetation Science (IAVS) endorsed the development of a standard exchange schema for vegetation plot data. In 2007 and 2008 two workshops were held at the National Evolutionary Synthesis Center (NESCent) in Durham, NC. The first workshop was held to formulate a common set of goals, concepts, and terminology for plot-based vegetation data. At the second workshop this ontology was developed into an XML schema representation incorporating elements from a number of other schema (e.g., Ecological Metadata Language, Taxon Concepts Schema, DarwinCore v2). Early drafts were subsequently presented at the 2008 TDWG conference and also at a workshop at the National Center for Ecological Analysis and Synthesis (NCEAS) in 2008.

The exchange schema for plot-based vegetation data (Veg-X) allows for observations of vegetation at both individual plant and aggregated observation levels. It ensures that observations are fixed to physical sample plots at specific points in time, and makes a distinction between the entity of interest (e.g., an individual tree) and the observational act (i.e., a measurement). The schema supports repeated measurements of both individual organisms and plots, allows observations of entities to be grouped following predefined or user-defined criteria, and ensures that the connection between the entity observed and taxonomic concept associated with that observation are maintained. Veg-X has now been adopted by the New Zealand National Vegetation Survey databank (NVS) and The Botanical Information and Ecology Network (BIEN) coordinated by NCEAS in the USA.

Exchange standards followed by the development of ecoinformatics tools built around those standards should allow scientists to efficiently combine plot data over extensive spatial and temporal gradients in order to perform analyses and make predictions of vegetation change and dynamics at local and global scales.The draft exchange standard can be viewed and discussed via its Wiki at wiki.tdwg.org/twiki/bin/view/Vegetation/WebHome.

9.32 Twelve Years of Migratory Fish Counting: Evolving Information Strategies for a Citizen Science project

R.D. Stevenson

University of Massachusetts, Boston, USA, robert.stevenson[at]umb.edu

For the last12 years in April and May, the Parker River Clean Watershed Association (PRCWA) has coordinated a survey of the migratory adult alewives (*Alosa pseudoharengus*, a herring species) using citizen volunteers. Alewives enter the Parker River in 2-6 bursts each lasting 1-3 days to reproduce. Each burst is called a run. Alewife runs occur in many rivers along the Atlantic Coast of the United States and Canada. The Parker is a small watershed (60 km^2) with relatively small runs (maximum 10 minute counts of 80 fish, maximum annual run size from 500 to 30,000). Fish are counted visually as they move through fish ladders built to allow passage around dams. Volunteers sign up for one to several hour-long observation slots each week during the migration season.

The information exchange needs for the Alewife count are similar to those needed by other citizen science projects: These include 1) project management 2) working with volunteers

(recruitment, training, coordination of observation effort, feedback on the project), 3) generation of data and results (data recording and reporting, data formatting and checking, data plotting and statistical analysis, summarizing results) and 4) communication of survey results with volunteers, partners and the press. There are a diverse range of people and organizations with whom information is exchanged including the management team (PRCWA leader, site coordinators, data manager), the PRCWA board, local landowners, volunteers, several local and state conservation groups, local, state and federal government agencies, academic scientists and the press.

Communication methods have gone through distinct four phases during project's history. I. From 1998 to 2003 most of the communication was by traditional means including person to person meetings, mail, newspapers, and telephone although already some aspects of all four of our information exchange needs were accomplished using email. Data organization and analysis was done using spreadsheet software on personal computers. II. During the 2004 season we tried using a database backed website designed by a graduate student team of software engineers from UMass Boston. The software had the capability to help with all of our information exchange needs. In addition the software could be used to scale up monitor projects across sites within a river and across watersheds by giving local coordinators the ability to describe site characteristics and control scheduling and analyze data. However, the resources needed to debug the software, improve the user interface, and refine the data analysis could not be sustained. III. From 2005 to 2007 we went back to the older methods but with the improvement that most of the volunteers sent in their results electronically. IV. Starting in 2008 the project has used Google Spreadsheets as a way for observers to report their observations, significantly reducing the work of the data coordinator. The advancement has been made possible because of the availability of Google software and the increasing savvy of the volunteers about the internet. For a small, mostly volunteer organization such as the PRCWA, the key to increasing the use of digital communications to monitor the fish migrations, whether it be YouTube videos to help with training or twitter feeds to report a run has started, is the availability of open source or ubiquitous digital tools and a volunteer group that has learned how use the technology.

9.33 Lifemapper 3: Geospatial Data and Computational Tools for Biodiversity Research

Aimee M. Stewart, C.J. Grady, James H. Beach

University of Kansas, Biodiversity Institute, USA, astewart[at]ku.edu, cjgrady[at]ku.edu, beach[at]ku.edu

Lifemapper 3 (LM3, www.lifemapper.org), supported by NSF EPSCoR 0919443, is an archive of biodiversity data and a set of computational tools provided through web services. Lifemapper synthesizes the known distributions of terrestrial plants and animals, and predicts future distributions based on various climate scenarios. It is built around the openModeller (openmodeller.sourceforge.net) ecological niche modeling (ENM) platform and uses web services to provide data and analysis using open-source tools. GBIF provides a monthly cache of their specimen occurrence database, which contains specimen records from over 283 data providers.

Data products LM3 offers include specimen records, projected species distribution maps, and environmental data, such as predicted climate data from the International Panel on Climate Change Fourth Assessment Report. Computational tools include ENM services, which allow users to request ecological niche modeling experiments computed with user-uploaded or LM archive input data, and user-specified modeling parameters.

The ChangeThinking project, supported by NSF Grant 0918590, uses the LM archive to build curricula and activities for teaching principles of ecology, biodiversity and climate change to middle school students.

LM3 will advance the repeatability and reproducibility of ecological research as part of the NSF-funded "A CyberCommons for Ecological Forecasting". VisTrails (www.vistrails.org), an open-source workflow and provenance management system developed at the University of Utah, will allow researchers to access LM3 web services in a workflow interface. Complete metadata for workflows will be saved in Ecological Metadata Language (EML, knb.ecoinformatics.org/software/eml), a standard schema for ecological metadata developed at the U.S. National Center for Ecological Analysis and Synthesis (NCEAS, www.nceas.ucsb.edu). EML documents will capture information about data inputs, processing parameters, outputs, and annotations explaining the reasoning behind decisions made in the research. The experiment metadata will be saved in a CyberCommons EML metadata catalog, and in a Data Observation Network for Earth repository (DataONE, https://dataone.org) to allow for discovery, query, and persistence of data and research.

The Lifemapper 3-SAM (LM3-SAM) project, supported by NSF 0851290, integrates LM3 and the widely used desktop application SAM, Spatial Analysis in Macroecology (www.ecoevol.ufg.br/sam). LM3-SAM will enable creation and analysis of multispecies macroecological grids. A Specify (specifysoftware.org) plug-in will submit experiments using local collection data and catalog results, while a QuantumGIS (www.qgis.org) plug-in will allow grid deconstruction to visually identify patterns. All LM3-SAM experiments can be created and executed in the VisTrails environment and will produce EML metadata.

In collaboration with the Map of Life Project (www.mapoflife.org), supported by NSF 0960549, the LM3 workflow will become more generic to support flexible species distribution modeling and additional analyses. Data archives will be accessible between projects through shared Application Programming Interfaces (APIs) and described in a complementary way in EML.

10. Computer Demonstrations (and Posters)

10.1 Biodiversity Data Digitizer – BDD

Etienne Américo Cartolano Júnior, Antonio Mauro Saraiva, Allan Koch Veiga, Diogo Borges Krobath, Luiz Guilherme Pilar Saraiva, Guilhermo Tavares

Universidade de São Paulo, Escola Politécnica, Computing Engineering Dept., Agricultural Automation Laboratory, São Paulo, SP, 05508-900, Brazil, etienne.cartolano[at]gmail.com, saraiva[at]usp.br

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. BDD is based on the Darwin Core standard (DwC), published by TDWG, that is centered on taxa, their occurrence in nature as documented by observations, specimens, and samples, and related information. Standards such as the Multimedia Resources Metadata Group Schema (MRTG Schema), currently submitted as TDWG draft standard, for collation, management and dissemination of multimedia resources relevant to biodiversity, and the Dublin Core, an interoperable metadata standard that support a broad range of purposes and business models, are used as complements to DwC. Other draft standards, developed for specific purposes and communities, are also being implemented as modules in BDD: interactions between specimens with focus on pollinators, pollinator monitoring and pollination deficit. 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 is the registration and handling (update, delete, and search) of species occurrences (specimens) data following Darwin Core, and of specimen interaction data, following the Interaction Extension. Data can be displayed on maps or table records and can be published to other systems via the TDWG Access Protocol for Information Retrieval (TAPIR). BDD

helps users improve and maintain data quality. Where relevant, users are prompted with 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 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 with suggestions, enhancing and completing the data record and decreasing the chance of entry errors. New features, always keeping data quality in mind, 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 and export data with spreadsheets. 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, Postgre database, and the Yii framework. For the future, it can evolve to accommodate the new Pollinator Interaction Extension under development with support from FAO. A prototype version can be visited at bdd.pcs.usp.br.

10.2 STERNA advanced semantic web tool about resources on birds

S. Cooleman[1], G. Geser[2], M. Louette[1], D. Meirte[1], P. Mergen[1], A. Mulrenin[2] and S.M. Pieterse[3]

[1] Royal Museum for Central Africa (RMCA) ; Tervuren, Belgium, stijn.cooleman[at]africamuseum.be, michel.louette[at]africamuseum.be, danny.meirte[at]africamuseum.be, patricia.mergen[at]africamuseum.be; [2] Salzburg Research Forschungsgesellschaft m.b.H., Salzburg, Austria, guntram.geser[at]salzburgresearch.at, andrea.mulrenin[at]salzburgresearch.at; [3] Netherlands Centre for Biodiversity Naturalis, Leiden, The Netherlands; sander.pieterse[at]ncbnaturalis.nl

STERNA (Semantic Web-based Thematic European Reference Network Application, www.sterna-net.eu) aims to develop a capacious bird information space following the guidelines of the European Digital Library (EDL). This distributed digital library solution will provide online public access to the rich European scientific and cultural heritage.

Since 2008, leading European organisations that collect and manage digital content on nature are participating in the STERNA consortium. They focus on improving the ability to search and access museum collections and audio-visual archives. The selected bird-related content comprises species, specimens and environmental associated information including material on birds such as photographs, drawings, 3D-images, sound recordings, movies, ancient books and ethnographic objects. Various contributors provide information services for users interested in birds, ranging from students, amateurs to land managers to scientists.

The project demonstrates how integrated access to heterogeneous collections of many institutions can be achieved based on Semantic Web technologies and standards such as Resource Description Framework (RDF) and Simple Knowledge Organisation System (SKOS).

As a content provider, the RMCA has established a plan to disseminate information regarding a selection of ornithological publications and unique specimens from its African Biology Department, in collaboration with its Human Sciences Department, which provides supplementary information on ethnographic objects composed of feathers or beaks. These data are linked to the scientific names.

An extensive quality check has been performed on these data by the curators and other experts whenever available access has been given to high quality images and metadata.

The semantic integration of content resources into one Web-accessible network is an ongoing challenge. STERNA uses a set of versatile tools that support various aspects of the content enrichment process (i.e., the RNA Toolset). 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.

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

A semantic search portal prototype (science.naturalis.nl/collections/sterna-birdwatchers-portal) has been available since November 2009. This search portal is still being validated and reformed by Trezorix and NCB Naturalis. An upgrade of the RNA Toolset and STERNA data model was recently realised in order to improve user-friendliness, searchability and updating of external data. These improvements include amongst others using a Common Register for taxonomic and vernacular bird names to enable multilinguality in the search interface.

The Access to Biological Collection Data (ABCD) Schema is used as a model for the RNA Toolset standard adaptation to natural history data, such as the RMCA bird type specimen data set via an online portal based on Biological Collection Access Services (BioCASE, biology.africamuseum.be/BiocaseProvider_2.4.2/www/querytool/main.cgi?dsa=STERNA). Tests have been made to get the STERNA prototype connected with BioCASE providers. The RMCA assesses interoperability between GBIF and STERNA, including usage of Semantic Web technologies.

Due to its participation in Biodiversity Information projects and its membership in Biodiversity Information Standards, the RMCA is involved in the project as technical advisor, and as work package leader for technology improvement, target user validation and evaluation of the STERNA approach. Researchers helped identify target users and their needs and create user certification tests. In order to spread the workload, evaluation methodology is now implemented by the Netherlands Institute for Sound and Vision.

STERNA is a showcase project for using semantic technologies and standards (like RDF and SKOS) and demonstrating the capability 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 GBIF.

Thanks to the STERNA project consortium partners from Salzburg Research Forschungsgesellschaft m.b.H. (Project Leader), Archipelagos, DOPPS Birdlife Slovenia, Heritage Malta, the Hungarian Natural History Museum, the Icelandic Institute of Natural History, the Natural History Museum of the Municipality of Amaroussion, the Natural History Museum of Luxembourg, Netherlands Centre for Biodiversity Naturalis, the Netherlands Institute of Sound and Vision, the Royal Museum for Central Africa (RMCA), the Teylers Museum, and Wildscreen/ARKive. STERNA's technical architecture and tools are provided by Trezorix. STERNA is a Best Practice Network funded under the EU *e*Content*plus* programme in the target area Digital Libraries: ec.europa.eu/information_society/activities/econtentplus/index_en.htm

10.3 Using the Vistrails Scientific Workflow Management System for Species Distribution Modeling

C.J. Grady, Aimee M. Stewart, James H. Beach

Biodiversity Institute, University of Kansas, Lawrence, KS, USA, cjgrady[at]ku.edu; astewart[at]ku.edu; beach[at]ku.edu

Lifemapper 3 (LM3, www.lifemapper.org) is an archive of species occurrence data and predictive distribution models and a set of geospatial computational tools provided through web services. LM3 synthesizes the known distributions of terrestrial plants and animals and predicts future distributions based on various climate scenarios. It is built around the openModeller (openmodeller.sourceforge.net) ecological niche modeling (ENM) platform deployed on a cluster and uses web services to provide data and analyses. LM3 ingests GBIF species occurrence records in monthly updates.

Vistrails (www.vistrails.com), developed at the University of Utah, is a scientific workflow management system allowing assembly and document exploratory computational tasks. Vistrails provides a graphical user interface for authoring workflows, parameterizing modules, and pipelining

data through computational steps and output visualizations. A distinguishing feature of Vistrails is its ability to generate comprehensive provenance metadata about complete workflows.

We are implementing a Vistrails interface for Lifemapper 3 that will allow biogeographers, macroecologists, and other interested researchers and students without programming backgrounds to easily assemble and run environmental niche models and explore a large number of experiments to visually assess the impacts of various climate change scenarios on species distributions.

In this demonstration, we will first illustrate the building of a simple ecological niche model experiment with one set of museum occurrence data points. Secondly, we will demonstrate Vistrails' ability to perform parameter sweeps by creating experiments for multiple species within a genus or family. Then we will expand these parameter sweeps to include changes in algorithm parameter values, resulting in batch submission of experiments for multiple species and multiple algorithm parameter sets. Finally, we will look at Ecological Markup Language generated for each experiment and how discuss it can be publicly cataloged to document the methods and transformations used to produce a particular experimental result, and also how it can be used to re-run niche modeling workflows at a future date.

This development work is supported by U.S. NSF EPSCoR Grant: 0919443

10.4 Tracking molecular processes of specimens in taxon-based biodiversity research

Gail E. Kampmeier[1] and Shelah Morita[2]

[1] Illinois Natural History Survey, Institute of Natural Resource Sustainability, University of Illinois at Urbana-Champaign, USA, gkamp[at]illinois.edu; [2] North Carolina State University, USA

The Mandala 8 database system www.inhs.illinois.edu/research/mandala/, which tracks and associates specimen data, taxonomic history, literature, and images, is being extended to document molecular processes associated with specimens.

Associating and tracking multiple molecular processes with the specimens to which they belong is often a challenge not met successfully, as seen by the incomplete specimen data registered for many sequences in GenBank (www.ncbi.nlm.nih.gov/genbank). As large genome projects are distributed in labs around the world, tracking the work done by various collaborators becomes a challenge. The lack of accurate specimen data, often including missing unique identifiers, locality and collecting event information, and taxon identification, diminishes the usefulness of obtaining molecular data and reporting it. In addition, without accurate recording of the processes used to obtain molecular data, time, effort, and materials can be wasted on processes that have been shown not to work in the past.

While these tools and processes are becoming increasingly commonplace, many students and researchers may not have yet optimized a routine for perfectly extracting the desired genome sequences, and the variables that contribute to refining this knowledge are numerous: what part of the organism was extracted and by what method? What buffer was used and in what concentration? How was the extracted material stored and what is the identifier used to associate it with a particular specimen? This process of extraction may be repeated many times for a single specimen, each needing a unique extraction identifier that can be associated with the unique specimen identifier. Each extraction may be linked to one or more attempts to amplify the extracted DNA (deoxyribonucleic acid) using a PCR (polymerase chain reaction) technique, specific primers (short DNA fragments involved in DNA synthesis), purification methods, and procedures to optimize the number of bands, their strength, and length. All of these variables can contribute to success or failure of attempts to obtain a sequence associated with a PCR identifier.

Ultimately information, including where the specimen is vouchered and the data associated with its collection, will be deposited in GenBank, published in a journal (literature citation), and

perhaps become part of TreeBase (www.treebase.org), with associated identifiers and URLs leading to online records reflecting the outcome of the research. The ability to document and retrieve this information from a database source available to individuals or team members working on distributed phylogenetics projects should improve the reliability and efficiency of project workflow.

"Taxonomic, Phylogenetic, and Evolutionary Studies of Horse flies (Diptera: Tabanidae): An Integrated Approach to Systematics Training." Partnerships for Enhancing Expertise in Taxonomy, NSF DEB 0731528 www.inhs.illinois.edu/research/tabanid

"Building the Dipteran Tree of Life: Cooperative Research in Phylogenetics and Bioinformatics of the True Flies (Insecta: Diptera)." Assembling the Tree of Life, NSF EF 0334948 www.inhs.illinois.edu/research/FLYTREE

10.5 Enterprise architecture for managing biodiversity data in Finland

Hanna Koivula, Mikko Heikkinen, Tapani Lahti and Hannu Saarenmaa

Finnish Museum of Natural History, Finland, hanna.koivula[at]helsinki.fi

Finnish Museum of Natural History is currently launching enterprise architecture for its data management. The aim of the project has been to support all data management processes at the museum carried out by both researchers and curators. The heart of this architecture is a data warehouse, where all data (collection, observational and literature based) are copied from various primary sources. This secondary data are parsed from the operative data sources thru an XML schema. This document-level schema (www.luomus.fi/fmnh2008) has been modified from the ABCD schema to better suit the needs for observational data and is backwards compatible with ABCD.

The structure of the data warehouse is based on a multidimensional star (or snowflake) schema, containing a few fact tables referencing any number of dimension tables. The facts that the data warehouse helps analyze are classified along different dimensions: the fact tables hold the main (stable) data, while the (slowly changing) dimension tables describe each value of a dimension and can be joined to fact tables as needed.

Dimensions can be managed with semantic web techniques to preserve the history and quality of the data. The next phase of this project is to model and apply semantics for managing taxonomy. Finnish time-based locality names have also been modeled as an ontology and can be applied for managing the geographical dimension at a national scale.

One of the key primary data sources for the data warehouse is Fieldjournal.org – an English version of a citizen science project Hatikka, launched in 2005 in Finland, and now containing about 2 million records. Fieldjournal.org is a web-based tool for citizen scientists and non-governmental organizations for collecting and organizing observational data. It uses XML-based data structures in the storage and management of primary data, but all data retrieval is based on the data warehouse optimized for fast ad hoc queries. Modern Online Analytical Processing (OLAP) tools make data analysis and reporting very flexible for various purposes. The data warehouse also contains observational data collected in research projects, as well as specimen data, which gives observations a nice scientific comparison background.

The Fieldjournal.org site (www.fieldjournal.org) can be used to store observations during various events, including the BioBlitz of TDWG 2010 Conference. The system allows for pinpointing from maps the observation localities as points, lines and areas from anywhere in the world. Unlike most data portals that are limited to some organism group only, observations from any organism group can be entered at the same time. Data entry forms can be customised for various projects and purposes, and they can include multimedia.

10.6 The Atlas of Living Australia – Spatial Portal

Dave Martin, Ajay Ranipeta, Angus MacAulay, Adam Collins and Lee Belbin
Atlas of Living Australia, Australia, lee[at]blatantfabrications.com

The Atlas of Living Australia will be launched to the public in mid-October. This project will integrate the widest range of biological observations in the Australian region at one web site. The web site will cover all types of biological data, conservation and invasive status, identification keys, literature, images, videos, molecular data, citizen science including annotations services, species interactions and mapping and spatial analysis.

The Atlas is a part of Australia's National Collaborative Research Infrastructure, and is currently funded to $38 million until June 30, 2012 with partners making $26.5 million in-kind contributions. The Atlas of Living Australia (ALA) was established through the National Collaborative Research Infrastructure Strategy (NCRIS) program. The project is a collaborative partnership between Commonwealth and State organisations that have stewardship over biological data and expertise in bioinformatics. It seeks to create one or more portals for deploying the rich biological data stored by Australian biodiversity institutions in flexible, integrated and innovative ways, and to provide tools for synthesis and analysis of these data.

The demonstration will outline the Atlas and seek input and feedback. The main developed components of the Atlas are the Biodiversity Information Explorer, the Spatial Portal and an Annotations Service. The BIE provides species pages and details of observations. The Spatial Portal can map species to order-level, over 500,000 gazetteer features and 300 environmental and contextual layers. It can also list, plot and download species in an Active Area defined by six methods. Analysis methods include MaxEnt and Environmental Domains.

10.7Linking Specify 6 Databases with Morphbank Repositories and Morphster Ontologies

Tim Noble, Rod Spears and Andy Bentley
Specify Software Project (www.specifysoftware.org), Biodiversity Institute, University of Kansas, Lawrence, KS, USA.
timo[at]ku.edu; rods[at]ku.edu; abentley[at]ku.edu

Specimens in one or more preparation types are held in biodiversity collections. "Prep types" include: dried, frozen, pickled, skinned, cleared and stained specimens, as well as more specialized types such as frozen tissues, DNA extracts, 3D images of skeletons, and images of the whole organism and its parts. Images provide documentation for habitat, behavior, natural colors, or features or morphological views of organisms. At the same time, microphotographs of body parts or of serial tissue sections, SEMs, TEMs, X-ray computed tomography (CAT), animations of serialized stacks of CAT scan images, or images from Magnetic Resonance Imaging (MRI) and video recording have become increasing popular preparation types for particular research objectives.

The Morphbank Project (www.morphbank.net) operates as a community repository for specimen images and provides indexing, browsing, and manipulation tools to make a collaborative community repository highly useful as a biological image resource discovery, analysis and annotation environment. Although individual Specify database repositories provide safe, local archiving functions, the benefit of sharing multimedia specimen preparations in Morphbank is valuable for two reasons: 1) it makes the images more easily discoverable to a much broader scientific audience, and 2) it provides the opportunity for image database users to provide feedback and annotations to the museum data providers, particularly annotations on the taxonomic identity of the organism.

The Specify, Morphbank and Morphster (www.morphster.org) Projects are collaborating in the development of a network protocol for automated deposition of Specify specimen images in Morphbank. We are also developing a protocol to provide "write-back" annotation updates to

Specify servers so that collection data managers are able to see and evaluate new or revised annotations from Morphbank users. Additionally, we plan to develop a Morphster plug-in extension for Specify that will explicitly represent ontology annotations as an integral part of specimen metadata.

In this demonstration we will show how image attachments in Specify can be sent and registered in the Morphbank repository, how those images can be linked to character ontologies, and discuss future integration plans among the three projects.

This Project is supported by U.S. NSF Grant 0851278.

10.8 DarwinCore Archive Descriptor Utility

David Remsen[1] and Michael Giddens[2]

[1] GBIF Secretariat, Copenhagen, Denmark, dremsen[at]gbif.org; [2] SilverBiology, Baton Rouge, LA, USA

The Darwin Core is a body of standards, ratified by TDWG in 2009, that include a set of terms relating to taxa and their occurrence in nature, and a set of practices regarding the use of these terms in the publication of biodiversity data and information. GBIF has adopted a text-based solution for using Darwin Core that both simplifies and extends the publication of species and species-occurrence data. This format is referred to as a Darwin Core Archive (DWCA) and provides a relatively non-technical option for publishing biodiversity data that does not require complicated installations of data publication software. Darwin Core Archives can be published via a simple web address or URL.

Darwin Core Archives support the publication of enriched data types that extend the core terms while retaining the relatively simple, text-based data format. These extensions, however, require the inclusion of an XML descriptor file that serves as a map to the different files and data elements in the archive. Many biologists and data managers find working with XML challenging while otherwise finding the technical threshold for producing Darwin Core Archives quite low.

The DarwinCore Archive Descriptor Utility is a web application that presents a simple interface for describing the data elements a data publisher wishes to serve to the GBIF network as basic text files and composes the appropriate XML descriptor file to accompany them. It communicates with the GBIF registry to provide an up-to-date listing of all relevant Darwin Core terms and available extensions and presents these in a simple checklist format. It provides links to supporting documentation and web-cast tutorials that collectively make publishing primary biodiversity data and annotated species checklists simpler than ever. In this session we will demonstrate the utility, provide hands-on exercises and solicit comments and recommendations.

www.silverbiology.com/clients/gbif/metamaker/

10.9 Let Taxonomists do Taxonomists' Work – even in Legacy Literature Digitization & Markup

Guido Sautter[1,2], Donat Agosti[1,3], Bob Morris[1,4]

[1] Plazi, Switzerland; [2] KIT, sautter[at]ira.uka.de; [3] American Museum of Natural History, New York, USA; [4] University of Massachusetts, Boston, USA

In recent years, projects like BHL (Biodiversity Heritage Library) have scanned and OCRed (Optical Character Recognition) large amounts of taxonomic legacy literature. However, using this data in mash-ups and semantic web applications like EoL (Encyclopedia of Life) requires the documents to be enhanced with semantic XML markup, which is considerable effort to create. The tools developed by Plazi have mitigated this effort considerably, but enhancing a document still takes about a minute of manual work per page. In several markup projects, we have observed that cleaning the document from print layout artifacts and OCR errors, a prerequisite for semantic

enhancement, accounts for about half of this effort. Doing this cleanup does not require any knowledge about taxonomy or biodiversity. Thus, taxonomists enhancing documents waste half of their time on markup activities that do not correspond to their qualifications – and payment.

As projects like Distributed Proofreaders (OCR proofreading), FoldIt (protein folding), and others have shown, activities that do not require thorough domain knowledge can be delegated to online communities. To activate this potential for the semantic enhancement of digitized taxonomic legacy literature, we have developed a Facebook Application that lets Facebook users clean documents, so taxonomists can concentrate their efforts on those steps of the semantic enhancement process that requires their specific knowledge.

To motivate Facebook users to participate, the application models cleaning documents as a competitive game. Users score points for their contributions, depending on both their number and their correctness. To foster disseminating the application, users can invite their friends to participate, and they score if the invitees contribute. Further, the application awards users with honorary badges for certain achievements, incrementally in bronze, silver, and gold. It honors three different types of achievement: the total number of points scored, streaks of error free contributions, and dissemination to friends who make contributions themselves.

A preliminary study pointed out the problem that users tend to simply commit the documents to rake in the score without actually checking and correcting it. To discourage this behavior, the application takes two measures: (a) it lets several users vote on each individual decision, and (b) it samples controversial decisions and uses them as actual CAPTCHAs. If too many users cheat, however, voting and CAPTCHAs cannot prevent its impact on data quality. A possible solution to this problem (currently not in use) is to individually alter the initial state in which a contribution request is presented to each voting user, preventing a couple of cheating users from reaching a conclusive vote.

To mitigate the impact the voting has on throughput – it multiplies the user time each decision takes – the weight of a user's vote increases with each error free contribution. This also benefits the user himself, as he then rakes in a larger cut of the score for some contribution because he has to share with fewer co-voters. Any error sets his vote's weight back to its starting value.

In the proposed demo, we plan to show the Facebook application to other people it might be useful to. This includes not only the actual application user interface inside Facebook, but also its architecture, and the tools for importing to-be-processed documents into the application backend.

10.10 Publishing Biological Classifications as SKOS Vocabulary Services on the Semantic Web

Jouni Tuominen, Matias Frosterus, [1]Nina Laurenne and Eero Hyvönen

Semantic Computing Research Group (SeCo), Aalto University, Department of Media Technology, Aalto, Finland; University of Helsinki, Department of Computer Science, [1] laurenne[at]cc.helsinki.fi

Taxon names may refer to more than one taxon and a taxon may have multiple names, which makes information retrieval and data integration problematic. On the Semantic Web, taxon names with unambiguous URIs can be collected into controlled vocabularies or ontologies, which enable the sharing of information in an interoperable way. For example, the observational data of birds can be annotated using these vocabularies. The vocabularies may contain relations between taxa, which can be used for further enhancing information retrieval. For instance, a user interested in the ecology of carnivores is possibly interested in the ecology of cats, too.

We have used the SKOS (Simple Knowledge Organization System) data model to represent a taxonomic hierarchy in RDF (Resource Description Framework). The basic unit of the SKOS model is a concept, which is used for representing taxa that are ordered into a single classification. The

vocabulary contains information about each taxon, e.g., their taxonomic ranks, scientific names, and common names.

The taxonomic hierarchy is modelled by using the hierarchical *skos:broader* relation. For example, the genus *Felis* is included in the subfamily Felinae. The preferred scientific and common names of the taxa are represented with the property *skos:prefLabel* and alternative names with *skos:altLabel*. The authorship information of a taxon is defined with the property *skos:note,* and the property *rdf:type* is used to indicate the taxonomic rank. We have extended the SKOS data model by introducing the property *creator*, which states the organization that has created the data, and *linkToWikipedia*, which provides the user with additional information about a taxon in Wikipedia. Taxa are referred to by using unique URIs that point to the location of the information describing the object on the web.

The model is demonstrated with the worldwide checklists of mammals (4,629 species) and birds (9,300 species). These checklists are extensive and contain the vernacular names in Finnish, Swedish and English, making them useful for a wide audience.

Once a taxonomic checklist has been represented in SKOS, it can be published instantly in the ONKI Ontology Service. The ONKI service provides a SKOS vocabulary browser for the human user and ready-to-use web widgets, and application interfaces (API) for applications. These components enable browsing, querying and visualizing of vocabularies, thus supporting use cases such as content indexing, taxon name disambiguation, searching, and query expansion.

The ONKI SKOS browser consists of three main components: 1) taxon name search with semantic autocompletion, 2) hierarchy and 3) properties of taxa. When a user types text in the search field, a query is performed to match taxon names. The result list shows matching names that can be selected for further examination. When a name is selected, the classification is visualized, and the properties are shown.

Taxonomic data can be maintained and edited with standard tools supporting the SKOS data model, such as the ontology editor Protegé 4 with the SKOSEd plugin and the SAHA metadata editor.

At the moment, new taxonomic checklists (over 80,000 species) of the Finnish Museum of Natural History are being published in ONKI. These vocabularies are integrated with other ontologies using the national Semantic Web ontology infrastructure FinnONTO.

This project is part of the national FinnONTO program funded by the Finnish Funding Agency for Technology and Innovation (Tekes) and a consortium of 38 public organizations and companies.

Semantic Computing Research Group: www.seco.tkk.fi/
National Semantic Web Ontology Project in Finland (FinnONTO), 2003-2012: www.seco.tkk.fi/projects/finnonto/
Biological Ontologies and Vocabularies: www.seco.tkk.fi/ontologies/biology/
ONKI Ontology Service: www.onki.fi/
SAHA – Browser-based Semantic Annotation Tool: www.seco.tkk.fi/services/saha/

10.11 Xper²: from names to expertises

Visotheary Ung, Florian Causse and Régine Vignes Lebbe
Laboratoire Informatique et Systématique, Paris, France, visotheary.riviere-ung[at]snv.jussieu.fr

Inventories, monitoring and protecting biodiversity gather together most biodiversity actors towards a common aim: a better knowledge of our environment. To succeed, they need user-friendly and efficient tools, in order, firstly, to name the taxa they are examining and secondly, to perform an efficient taxonomic work (Polaszek, 2005, Godfray, 2002). Xper² is a versatile software for storing, managing, editing and on-line publishing of taxonomic knowledge and free access keys. Written in Java, it is available on Windows™ Mac™ or Linux in French, English or Spanish versions and we extend our linguistic skills to Asian languages as we are proud to present our latest Chinese version. With its intuitive interface Xper² is aimed at professional taxonomists as well as naturalists

who merely want to identify specimens using a ready-made application. Xper² is free of charge and can be downloaded at: lis-upmc.snv.jussieu.fr/lis/?q=en/resources/softwares/xper2.

Xper² version 2.1 focuses on interoperability between systems and can import and export into Structured Descriptive Data format (Hagedorn *et al*, 2006), the SDD standard emerged from a TDWG initiative. The knowledge bases can be exported to HTML files for an easy publishing and to Nexus formats for phylogenetic analysis.

Xper²'s users are taxonomists, teachers, fauna and flora experts and ecologists. We show here, some results and projects in order to browse many basic functionalities of Xper² that assist our users in their daily work. Among others we have chosen to illustrate pure taxonomic work, with the creation of flora; an extension to phylogenetic studies with an example of Lagomorph's parasites; epidemiologic monitoring examples and case studies of animal welfare.

lis-upmc.snv.jussieu.fr/lis/?q=en/resources/softwares/xper2.
www.tdwg.org
pubmlst.org/software/analysis/start/manual/nexus_format.shtml
Godfray H.C.J., 2002. Challenges for Taxonomy. *Nature* 417:17—19.
Hagedorn, G., Thiele, K., Morris, R. & Heidorn, P.B. 2006. The Strutured Descriptive Data (SDD) w3c-xml-schema, version 1.1.
Polaszek A., 2005. A universal register for animal names. *Nature* 437: 477.

10.12 RHS Orchard: harvesting a different sort of fruit

Rupert G. Wilson and Janet J. Cubey
Royal Horticultural Society Garden Wisley, UK, rupertwilson[at]rhs.org.uk

The Royal Horticultural Society (RHS, www.rhs.org.uk) is the UK's leading gardening charity dedicated to advancing horticulture and promoting good gardening. Our goal is to help people share a passion for plants, to encourage excellence in horticulture and inspire those with an interest in gardening.

The RHS manages several unique datasets including:

- *RHS Plant Finder* (>70,000 cultivated plants available in the trade)
- in excess of 283,000 plant names
- 9 International Cultivar Registers > 300,000 names
- more than 40,000 English common names
- more than 450 plant advice profiles
- more than 25,000 plant portraits

RHS data are primarily managed in the popular *BG-BASE* system as well as in image management software and several in-house developed Microsoft SQL server databases with bespoke front ends.

The RHS has long been aware of the value in information "locked into" our existing databases. Our goal is to make that information accessible, and add to its value through integration and interoperability. Fundamental to the realisation of this goal is Project Rubus; a key cross-cutting programme within the RHS led by the Science & Advice team in partnership with the Online and IT Development teams.

The vision of Rubus is "*To utilise to the full our unique depth of knowledge in the field of cultivated plants by creating a knowledge management system for all our horticultural data capable of delivering this to our audiences*".

Rubus encompasses two interdependent streams of activity:

- Development of RHS Orchard; a tool kit capable of uniting & handling the diverse plant and horticultural data held by the RHS;
- Development of online interfaces, driven by RHS Orchard, to deliver these data and make our knowledge accessible to members and the public.

Products delivered so far include:

- RHS Orchard: a web based platform/tool kit for managing the centrally held data
- Plant Name Checker: a simple web tool designed to help any RHS employee who wants to know the correct name of a plant and/or how to style the name
- A suite of web services to deliver data to the completely re-developed online Plant Selector search

The RHS has experience of engaging with a wide range of different audiences including children and families, educators, and the complete spectrum of gardeners from novices to experienced amateurs, professional horticulturalists and plantsmen. Through Project Rubus we aim to improve access to our information for all our audiences, while also increasing awareness of cultivated plants in the cultural and biodiversity domains.

11. Contributed Abstract

11.1 Taxon Meta-Ontology TaxMeOn – Towards an Ontology Model for Managing Changing Scientific Names in Time

Nina Laurenne, Jouni Tuominen, Mikko Koho and Eero Hyvönen
Semantic Computing Research Group (SeCo), Aalto University, Department of Media Technology, Aalto, Finland; University of Helsinki, Department of Computer Science, laurenne[at]cc.helsinki.fi

The Semantic Web is based on machine-processable meanings, i.e., metadata and ontologies providing a new promising approach for representing and managing the scientific name system. A scientific name and the related taxonomic concept(s) form a unit in which one or the other can change. Therefore, names are unreliable identifiers for taxa. We present an ontology-based method for representing scientific names and taxonomic information that change in time. The practical goal of the system is to facilitate more accurate taxonomic metadata descriptions about names, integration of heterogeneous scientific data about organisms in different times and from different sources (publications, data bases, observations, museum collections), and to enable semantic searches, and linking in applications.

Conceptualized taxonomic information is expressed as *classes* that represent sets of individual *instances*. For example, *Cerambyx* is an *instance* of the *class* genus. *Properties* defined as relations between *classes* tell how individuals are related to each other. For example, genus and author are *classes* connected by the *property* 'is described by' (e.g., *Cerambyx* is described by Linnaeus). This ongoing research introduces an ontology schema, i.e., a meta-ontology (currently ca. 20 *classes* and 70 *properties*) to model information about scientific names, taxon concepts, authorships and taxonomic statuses. The focus is not on the characters defining taxa or on physical specimens. Terms of Darwin Core were applied when possible, but additional and more elaborate terms were needed too. The system is based on the Web Ontology Language OWL standard.

Taxonomic ranks in the model are represented as an OWL subclass hierarchy of *classes,* and individual taxa (e.g., *Cerambyx*) as *instances* of them. The taxonomic hierarchy is based on the 'part of' relation between *instances* of taxa. The 'subclass of' relation is not used, because a taxon hierarchy semantically defines memberships of organisms in groups rather than the subclass of relation (of OWL). If subclasses were used, an *instance* of a genus would be falsely an *instance* of taxa of higher levels.

The taxonomic concepts of differing views can be mapped to each other in multiple ways. The connection between taxa can be very specific (congruent, is included/includes, overlaps with; all determined by the characters or a membership of a group) or loosely defined, which leaves the choice to add incomplete or lacking information. The taxon names and the concepts are referred to

using Universal Resource Identifiers (URI) that also act as pointers to the location of the information that describe the objects on the web. A change in a taxon name or in a concept leads to the creation of a new concept and URI that is related to the former concept(s) and the time of the change. For example, when a species is shifted into another genus, a new species *instance* is created without changing the old, and the two views can be managed in time.

The ontology editor Protegé was used for constructing the *classes* and the *properties* of the model. The model is being applied to develop an ontology of nine genera of Afro-tropical eucnemid beetles. The taxonomy of the group is challenging and for instance, the genus *Pterotarsus* has gone through 22 taxonomic changes. The result will be published on the ONKI Ontology Service.

This work is part of the national FinnONTO program and is funded by the Finnish Funding Agency for Technology and Innovation (Tekes) and a consortium of 38 public organizations and companies. Fruitful co-operation with Jyrki Muona, Hans Silfverberg, Hannu Saarenmaa, and the Finnish Museum of Natural History is acknowledged.

Semantic Computing Research Group: www.seco.tkk.fi/
National Semantic Web Ontology Project in Finland (FinnONTO), 2003-2012: www.seco.tkk.fi/projects/finnonto/
Biological Ontologies and Vocabularies: www.seco.tkk.fi/ontologies/biology/
ONKI Ontology Service: www.onki.fi/

Index to Contributors