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

# Biodiversity Information Standards

# The Proceedings of TDWG

# Provisional Abstracts of the 2008 Annual Conference of the Taxonomic Databases Working Group

19-24 October 2008 Fremantle, Australia (Hosted at the Maritime Museum of Western Australia)

Edited by Anna Weitzman and Lee Belbin

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

Fremantle, 2008

TDWG 2008 was sponsored by:



© Taxonomic Databases Working Group, 2008

© Missouri Botanical Garden, St. Louis, Missouri, U.S.A., 2007

To be cited as:

# Weitzman, A.L., and Belbin, L. (eds.). Proceedings of TDWG (2008), Fremantle, Australia.

This book contains abstracts of the papers, posters and computer demonstrations presented at the Annual Conference of the Taxonomic Databases Working Group held 17-25 October 2008 at the Maritime Museum of Western Australia, hosted by the Department of Environment and Conservation, Western Australia.

The meeting attracted more than 160 participants from 25 countries and 109 prestigious scientific research institutions, museums and companies.

**The editors** gratefully acknowledge with thanks the contributions of Gail Kampmeier, Arturo Ariño, Donald Hobern, Adrian Rissoné, Stan Blum, Hilmar Lapp, Greg Whitbread, Éamonn Ó Tuama, Bob Morris, Annie Simpson, Steve Kelling, Tim Robertson and Markus towards the editing of this publication. We also appreciate the numerous independent peer reviewers who have ensured the quality of the submissions.

We also appreciate the work of Renato de Giovanni, Roger Hyam, Wouter Addink and Donat Agosti for preparing tutorials for this meeting.

Published and distributed as an Adobe<sup>®</sup> Portable Document Format (PDF) document for free download from the Conference web site at <u>http://www.tdwg.org/conference2008/</u>

# **Proceedings of TDWG**

### **Table of Contents**

Table of Contents	
<ol> <li>Introduction to TDWG</li> <li>1.1. TDWG – Looking Back and Looking Forward</li> </ol>	
Donald Hobern	10
1.2. TDWG Standards Roadmap	
Roger Hyam 2. From Small to Big: How to Capitalize on Small Contributions to the Big Dreams of Sha	10
Biodiversity Information	mg
2.1. Introduction: From Small to Big	
Gail E. Kampmeier	11
2.2. Every datum counts! - Capitalising on small contributions to the big dreams of mobilizing biodiversity information	
Vishwas Chavan, Éamonn Ó Tuama, Samy Gaiji, David Remsen, Nicholas King	12
2.3. Experiences in mobilizing biodiversity data – the story of HermesLite	
Piers Higgs, Tim Carpenter	12
2.4. Small pieces loosely joined: Building scientific web communities with Scratchpads Simon Rycroft, Dave Roberts, Kehan Harman, Vince Smith	13
2.5. Designing a global network to accommodate contributions from all sources and tech	
abilities	10
Tim Robertson 3. Introduction to TDWG Standards	13
3.1. An Introduction to Scientific Names of Organisms	
Richard Pyle	14
3.2. An Introduction to Data Exchange Protocols in TDWG Renato De Giovanni	15
3.3. An introduction to biological descriptions	15
Kevin Thiele	15
3.4. Globally Unique Identifiers in Biodiversity Informatics	16
<ul><li>Kevin James Richards</li><li>4. Biodiversity Informatics: Key Organizations and Projects</li></ul>	16
4.1. GBIF: From Prototype towards Full Operation	
Éamonn Ó Tuama, Vishwas Chavan, Markus Döring, Samy Gaiji , David Remsen, Ti	
Robertson 4.2. DataNetONE: A Distributed Environmental Data Archive	16
Matthew B. Jones	17
4.3. European Distributed Institute of Taxonomy (EDIT)	
Walter G. Berendsohn	18
4.4. The Biodiversity Collections Index – Linking 140 Million Records? Roger Hyam	18
4.5. KeyToNature Species Identification e-Tools for Education	10
Dr Peter Schalk <sup>1</sup> , Professor Pier Luigi Nimis <sup>2</sup> , Ir Wouter Addink <sup>1</sup>	19
5. Biodiversity Informatics Literature and More	
5.1. Plazi: The technical set up Guido Sautter <sup>1</sup> , Donat Agosti <sup>2</sup>	20
5.2. Plazi: Prospects for Mark Up of Legacy and Prospective Publications	_0
Terry Catapano	21
5.3. An evaluation of taxonomic name finding and next steps in BHL developments.	

	Chris Freeland	22
	5.4. Precision in accessing the descriptions, keys and other contents of digitized taxonomic	2
	literature: the INOTAXA prototype	
	Chris Lyal <sup>1</sup> , Anna Weitzman <sup>2</sup>	22
6.	Ontologies	
	6.1. Ontologies and interoperability in evolutionary comparative analysis	
	Francisco Prosdocimi <sup>1</sup> , Brandon Chisham <sup>2</sup> , Enrico Pontelli <sup>2</sup> , Julie D Thompson <sup>1</sup> , Arlin	
	Stoltzfus <sup>3</sup>	23
	6.2. Taxonomic ontologies: Bridging phylogenetic and taxonomic history	
	Peter E. Midford	23
	6.3. Lessons learned from semantic web prototyping in ecology	
	Cynthia Sims Parr <sup>1</sup> , Joel Sachs <sup>2</sup> , Tim Finin <sup>2</sup>	24
	6.4. SERONTO: a Socio-Ecological Research and Observation oNTOlogy	
	D.C. (Bert) van der werf <sup>1</sup> , Mihai Adamescu <sup>2</sup> , Minu Ayromlou <sup>3</sup> , Nicolas Bertrand <sup>4</sup> , Jakul	
	Borovec <sup>5</sup> , Hugues Boussard <sup>6</sup> , Constatin Cazacu <sup>2</sup> , Toon van Daele <sup>7</sup> , Sabina Datcu <sup>2</sup> , Mark	
	Frenzel <sup>8</sup> , Volker Hammen <sup>8</sup> , Helena Karasti <sup>9</sup> , Miklos Kertesz <sup>10</sup> , Pirjo Kuitunen <sup>11</sup> , Mandy	
	Lane <sup>4</sup> , Juraj Lieskovsky <sup>12</sup> , Barbara Magagna <sup>3</sup> , Johannes Peterseil <sup>3</sup> , Sue Rennie <sup>4</sup> , Herbert	
	Schentz <sup>3</sup> , Katharina Schleidt <sup>3</sup> , Liisa Tuominen <sup>13</sup>	25
	6.5. SONet (Scientific Observations Network) and OBOE (Extensible Observation	
	Ontology): facilitating data interoperability within the environmental and ecological science	es
	through advanced semantic approaches	
	Mark Schildhauer <sup>1</sup> , Shawn Bowers <sup>2</sup> , Matt Jones <sup>1</sup> , Josh Madin <sup>3</sup>	26
	6.6. A standard information transfer model scopes the ontologies required for observations	
	Simon J.D. Cox <sup>1</sup> , Laurent Lefort <sup>2</sup>	27
7.	Phylogenetics Standards	
	7.1. Obstacles and Avenues Facing Phyloinformatic Research	
	William H Piel	28
	7.2. Integrating space, time and form; the challenge of developing standards for	
	geophylogenies.	•
	David Kidd	29
	7.3. Data standards in phylogenetics: the NeXML project	20
	Rutger Vos 7.4. Emercing Data Standards for Phylogenomics Research	30
	7.4. Emerging Data Standards for Phylogenomics Research Christian M Zmasek	30
		50
	7.5. Emerging standards in phylogenetic nomenclature Nico Cellinese	31
	7.6. Towards a Simple, Standards-Compliant, and Generic Phylogenetic Database Module	
	Hilmar Lapp <sup>1</sup> , Todd Vision <sup>2</sup>	32
Q	Atlas of Living Australia - Australian Biodiversity Informatics	52
0.	8.1. Introducing Australia's Virtual Herbarium (AVH)	
	Ben Richardson	33
	8.2. CMKb - integrating Australian customary medicinal plant knowledge with the	55
	framework of Atlas of Living Australia	
	Jitendra Gaikwad <sup>1</sup> , Shoba Ranganathan <sup>2</sup>	34
	8.3. TAXAMATCH, a "fuzzy" matching algorithm for taxon names, and potential	54
	applications in taxonomic databases	
	Tony Rees	35
	8.4. Making effective use of the Atlas of Living Australia in conservation assessment through	
	advances in spatial analysis and modelling	-811
	Simon Ferrier	36
	8.5. Linking phylogenetic trees, taxonomy & geography to map phylogeography using	20
	Biodiverse	
	Dan Rosauer, Shawn Laffan	36
9.	Atlas of Living Australia - Architecture and deliverables	
	9.1. The Atlas of Living Australia – an introduction	

Donald Hobern, Nick dos Remedios, Dave Martin, Wolf Wanjura, Lynette Woodburn 9.2. Taxonomy Research & Information Network (TRIN): seeking biodiversity data	37
standards	
E. Margaret Cawsey <sup>1</sup> , Jim Croft <sup>2</sup> , Garry Jolley-Rogers <sup>2</sup>	38
9.3. Australian Faunal Directory (AFD) and Australian Plant Census (APC): Content,	
Architecture and Services	•
Greg Whitbread <sup>1</sup> , Helen Thompson <sup>2</sup> , Matthew Hand <sup>3</sup>	39
9.4. Providing Annotation Services for the Atlas of Living Australia	10
Ron Chernich, Stephen Crawley, Jane Hunter	40
10. Atlas of Living Australia - Exploring User Scenarios	
10.1. Uses for biodiversity data – the Atlas of Living Australia user needs analysis	41
Paul Flemons <sup>1</sup> , John Tann <sup>1</sup> , Lynda Kelly <sup>1</sup> , Donald Hobern <sup>2</sup> 11. Plenary - Invited Speakers	41
11.1. The Value of Data and its Management: Confessions of an Applied Environmental	
Scientist	
Neil Burrows	42
11.2. Biodiversity Information Standards: Are We Going Wrong, or Just Not Quite Right?	74
Jim Croft	42
12. Biodiversity-Related networks and Tools	
12.1. speciesLink & openModeller: Network and Tool Enabling Biodiversity Research	
Renato De Giovanni, Alexandre Marino, Sidnei de Souza, Dora Ann Lange Canhos,	
Marinez Ferreira de Siqueira, Vanderlei Perez Canhos	43
12.2. EDIT Platform for Cybertaxonomy – An Overview	
Andreas Müller, Andreas Kohlbecker, Pepe Ciardelli, Niels Hoffmann, Patricia Kelbert	44
12.3. Data Integration Services for Biodiversity Informatics: An Example from Plant	
Collections and Morphbank	
Greg Riccardi <sup>1</sup> , Boyce Tankersley <sup>2</sup>	45
12.4. Community Building and Collaborative Georeferencing using GEOLocate	
Nelson Rios, Henry L. Bart	46
12.5. Halt the loss of biodiversity data! – Towards a new archiving infrastructure for	
biodiversity information	
Anton Güntsch, Walter G. Berendsohn	46
13. Species Profile Model	
13.1. A gentle stroll through the Species Profile Model	47
Robert A. Morris	47
13.2. Plazi: Implementing SPM	48
Terry Catapano 13.3. Experiences in consuming SPM marked-up data	48
Patrick Leary	49
13.4. The Species Profile Model from an Avian Perspective.	49
Jeff Gerbracht, Steve Kelling	49
14. Multimedia in Biodiversity Informatics	77
14.1. Introduction to the session: The relationship of the GBIF Multimedia Resource Task	
Group to the presentations and to the aims of the TDWG Image Interest Group	
Robert A. Morris <sup>1</sup> , Vishwas Chavan <sup>2</sup>	50
14.2. The NBII's Digital Image Library – bringing together diverse images means quality	
control and standards	
Annette L. Olson	50
14.3. Creative Commons – Have Your Copyright and Share It Too	
Chuck Miller	51
14.4. Using Fedora Commons to create a persistent archive for digital objects	
Phil Cryer	52
14.5. Using the JPEG2000 image format for storage and access in biodiversity collections	
Chris Freeland	52

<ul> <li>14.6. OntoMorphBankSter: Image-driven Ontology and/or Ontology-driven Image Annotation Greg Riccardi<sup>1</sup>, Austin Mast<sup>1</sup>, Daniel Miranker<sup>2</sup>, Ferner Cilloniz<sup>3</sup></li> <li>15. The Global Invasive Species Information Network 15.1. Developing a cyberinfrastructure: Experiences from a regional network of plant towonomists</li> </ul>	52
<ul> <li>taxonomists</li> <li>Zack Murrell, Derick Poindexter, Michael Denslow</li> <li>15.2. Building a TAPIRLite Toolkit for the Global Invasive Species Information Network (GISIN)</li> </ul>	55
Jim Graham <sup>1</sup> , Catherine Jarnevich <sup>2</sup> 15.3. Invasive Species Information Management and Exchange in the Americas: IABIN Invasive Information Network (I3N)	56
Christine Fournier 15.4. Ensuring the relevance of invasive species information management tools for policy development and effective management Judy Fisher	56 57
15.5. Data Models to share Invasive Sspecies Information through the Global Invasive Species Information Network Annie Simpson <sup>1</sup> , Jim Graham <sup>2</sup> , Elizabeth Sellers <sup>3</sup> , Michael Browne <sup>4</sup>	57
<ul> <li>16. Bringing Organism Observations into Bioinformatics Networks</li> <li>16.1. The Significance of Observations to Biodiversity Studies. Steve Kelling</li> </ul>	59
16.2. Towards Global Observation Networks: GBIF's plans for mobilizing primary biodiversity data Vishwas Chavan, Éamonn Ó Tuama	60
<ul> <li>16.3. Directions in observational data organization: from schemas to ontologies Matthew B. Jones<sup>1</sup>, Chad Berkley<sup>2</sup>, Shawn Bowers<sup>3</sup>, Josh Madin<sup>4</sup>, Mark Schildhauer<sup>2</sup></li> <li>16.4. Dealing with Sensitive Data: managing access, issues of data quality, and data outparties</li> </ul>	60
ownership Arthur David Chapman 16.5. Domain-centric observation networks: Experiences gained from the Avian Knowled Network	61 ge
Denis Lepage 17. Wild Ideas!!	61
17.1. A Hadoop-based Prototype for the Filtered-Push project Zhimin Wang <sup>1</sup> , Hui Dong <sup>2</sup> , Maureen Kelly <sup>3</sup> , James A. Macklin <sup>3</sup> , Paul J. Morris <sup>4</sup> , Rober A. Morris <sup>2</sup>	t 62
17.2. Data Exchange Standards – The Case for Being Stupidly Simple Chuck Miller,	63
<ul><li>17.3. 3D and beyond. What graphics cards can do for you Bart Meganck, Patricia Mergen</li></ul>	63
17.4. SpeciesIndex: A Practical Alternative to Fantasy Mashups? Roger Hyam	64
<ol> <li>Computer Demonstration</li> <li>18.1. Introducing the EDIT Desktop Taxonomic Editor</li> <li>Pepe Ciardelli, Andreas Müller, Anton Guentsch, Walter Berendsohn</li> </ol>	65
<ul><li>18.2. SilverCollection: A Standardized Web Portal for Biological Collections Michael Giddens</li></ul>	66
18.3. Mandala 8: A Database System that Can Start Out Small and End Up Big Gail E. Kampmeier	67
18.4. The new EDIT Specimen and Observation Explorer for Taxonomists Patricia Kelbert, Niels Hoffmann, Jörg Holetschek, Anton Güntsch, Walter G. Berendso	
18.5. EDIT Data Portal – A flexible publishing tool for biodiversity information networks Andreas Kohlbecker <sup>1</sup> , Andreas Müller <sup>1</sup> , Niels Hoffmann <sup>2</sup>	68 69

18.6. NatureGate: An Online Service to Promote Education for Sustainable Development a Biodiversity	ınd
Eija Lehnuskallio <sup>1</sup> , Arja Kaasinen <sup>2</sup> , Mauri Åhlberg <sup>2</sup>	70
18.7. NBII Clearinghouse (Mercury)-Distributed Metadata Management, Data Discovery a	ina
Access System Giri Palanisamy, Vivian Hutchison, Mike Frame	71
18.8. IRMNG – The Interim Register of Marine and Nonmarine Genera	/1
Tony Rees	72
18.9. Using AquaMaps for biodiversity assessment, including a prototype MPA (Marine	
Protected Area) network design tool	
Tony Rees	73
18.10. Open Access Data Policy and Tools for its Implementation at the Finnish Museum of	of
Natural History	74
Hannu Saarenmaa <sup>1</sup> , Miruna Badescu <sup>2</sup> , Alex Morega <sup>2</sup>	74
18.11. Feature recognition for herbarium specimens (Herbar-Digital) Karl-Heinz Steinke, Robert Dzido, Martin Gehrke, Klaus Prätel	74
18.12. INOTAXA implemented - a digital future for taxonomic literature	/+
Anna Weitzman <sup>1</sup> , Chris Lyal <sup>2</sup>	75
19. Poster	
19.1. ETI BioPortals: customizable common access system to distributed heterogeneous da	ata
sources	
Ir Wouter Addink, Dr Peter Schalk	76
19.2. Sampling biodiversity sampling	
Arturo H. Ariño, Javier Otegui	77
19.3. A Global Information System for the Conservation and Sustainable Use of Plant Genetic Resources for Food and Agriculture (PGRFA)	
Elizabeth Arnaud <sup>1</sup> , Michael Mackay <sup>1</sup> , Dag Terje Filip Endresen <sup>1</sup> , Sonia Dias <sup>1</sup> , Rajesh	
Sood <sup>1</sup> , Kiran Viparthi <sup>1</sup> , Milko Skofic <sup>1</sup> , Adriana Alercia <sup>1</sup> , Tito Franco <sup>1</sup> , Frederick Atieno <sup>1</sup>	1
Xavier Scheldeman <sup>1</sup> , Anzar Shamsie <sup>2</sup> , Selim Louafi <sup>3</sup> , Pete Cyr <sup>4</sup>	, 78
19.4. LifeWatch – e-Science and Technology Infrastructure for Biodiversity Research	
Walter G. Berendsohn, Marie Gebhardt	79
19.5. Proof of concept study of the Socio-Ecological Research and Observation oNTOlogy	7
(SERONTO) for integrating multiple ecological databases.	
Nicolas Bertrand <sup>1</sup> , Herbert Schentz <sup>2</sup> , Bert Van der Werf <sup>3</sup> , Barbara Magagna <sup>2</sup> , Johannes	00
Peterseil <sup>2</sup> , Sue Rennie <sup>1</sup>	80
19.6. The Register of Antarctic Marine Species: towards a comprehensive inventory of the Southern Ocean biodiversity	
Bruno Danis	81
19.7. SCAR-MarBIN: free and open access to Antarctic Marine Biodiversity Data	01
Bruno Danis	81
19.8. Development of the Integrated Publishing Toolkit by the Global Biodiversity	
Information Facility (GBIF)	
Markus Döring, Tim Robertson	82
19.9. The DNA Bank Network – How GBIF Technology Enables a New Generation of	
Communicating DNA Repositories	82
Gabriele Droege, Holger Zetzsche, Anton Guentsch, Birgit Gemeinholzer 19.10. Invasive Species Tools for Risk Assessment in the Americas: IABIN Invasive	02
Information Network (I3N)	
Christine Fournier	83
19.11. Taxonomic data exchange facilities for descriptive tools	
Hélène Fradin, Elise Kuntzelmann, Régine Vignes Lebbe	83
19.12. A blueprint of the GBIF decentralisation strategy	
Samy Gaiji <sup>1</sup> , Éamonn O'Tuama <sup>2</sup> , Vishwas Chavan <sup>2</sup> , David Remsen <sup>2</sup>	84
19.13. BioGeobif: Automated Georeferencing Using BioGeomancer	

Andrew William Hill <sup>1</sup> , Paul Flemons <sup>2</sup> , Reed Beaman <sup>3</sup> , Ajay Ranipeta <sup>2</sup> , Philip Goldstein <sup>1</sup> , Robert P Guralnick <sup>1</sup>	85
19.14. How to Reduce Your Coffee Consumption – Performance Investigations on Large	
Occurrence Databases	
Jörg Holetschek	85
19.15. A Pan-European Species-directories Infrastructure (PESI)	
Yde de Jong <sup>1</sup> , Charles Hussey <sup>2</sup> , Mark Costello <sup>3</sup> , Thierry Bourgin <sup>4</sup> , Anton Guentsch <sup>5</sup> , War	rd
	86
19.16. CABIN: Implementing a Biodiversity Information Network in Central Africa	
Charles Kahindo <sup>1</sup> , Franck Theeten <sup>2</sup> , Patricia Mergen <sup>2</sup> , Bart Meganck <sup>2</sup> , Kim Jacobsen <sup>2</sup> ,	
	87
19.17. Building and Management of the Korean Biodiversity Information System	
	88
19.18. Information Integration in Documentation of Biodiversity	00
	88
19.19. Using LSIDs to Link Taxonomic Concepts to Scientific Names for Efficient Data	00
Integration	
	89
19.20. The Collaborative Ontology Creation Process Within The ALTER-Net Community	07
For the Socio-Ecological Research and Observation oNTOlogy (SERONTO)	
	90
19.21. Kansas State University's Biodiversity Information System: Combining institutional	/0
expertise to enhance access to collections information	
Mark Mayfield, Carolyn Ferguson, Gregory Zolnerowich, David Allen, Harish Maringan	ti
	91
19.22. The roots and branches of the GeoPlatform tree.	1
	91
19.23. The roots and branches of the EDIT GeoPlatform tree	1
Bart Meganck <sup>1</sup> , Pere Roca Ristol <sup>2</sup> , Franck Theeten <sup>1</sup> , Patricia Mergen <sup>1</sup> , Garin Cael <sup>1</sup> , Kim	
	92
19.24. RDF based Reference Network Architecture for a distributed Digital Library system.	
The Royal Museum for Central Africa's use case on African Bird information (STERNA	,
project)	
	92
19.25. tropicos.org - A New Look	12
	93
19.26. Remote Annotation of Specimens: Use cases and messages of a Filtered Push network	
Paul J Morris <sup>1</sup> , Hui Dong <sup>2</sup> , Maureen Kelly <sup>3</sup> , James A Macklin <sup>3</sup> , Robert A Morris <sup>2</sup> , Zhimin	
	n 94
19.27. Standardizing Mansfeld's World Database of Agricultural and Horticultural Crops by	
mplementing a concept-based data model	y
	95
	95
19.28. Unlocking Biodiversity Data: Semantic Cyberinfrastructure for Investigation and	
Discovery (SCID)	96
	90
19.29. The GBIF Work Programme 2009-2010: From Prototype towards Full Operation $(f_{12}, f_{23})^{-1}$ Alberta Change $(f_{12}, f_{23})^{-1}$ Alberta Change $(f_{12}, f_{23})^{-1}$ Alberta Change $(f_{12}, f_{23})^{-1}$	
Éamonn Ó Tuama <sup>1</sup> , Vishwas Chavan <sup>1</sup> , Samy Gaiji <sup>2</sup> , Alberto Gonzalez-Talaván <sup>1</sup> , Nichol	
	96
19.30. GBIF's Global Biodiversity Resources Discovery System	~ <b>7</b>
	97
19.31. The Encyclopedia of Life: Status report on species pages, contributions, and curators	
	98
19.32. The Global Names Architecture: an integrated and federated approach to enabling	
discovery and access to biodiversity information.	00
David P Remsen	99

19.33. Ten Years of FloraBase	
Ben Richardson <sup>1</sup> , Alex Chapman <sup>2</sup>	99
19.34. EDIT mapViewer: a Gegraphic Web Application for Taxonomists	
Pere Roca Ristol <sup>1</sup> , Pablo Sastre <sup>1</sup> , Jorge Lobo <sup>1</sup> , Franck Theeten <sup>2</sup> , Bart Meganck <sup>2</sup> , Patrici	ia
Mergen <sup>2</sup>	100
19.35. e-Biosphere 09: International Conference on Biodiversity Informatics	
David E. Schindel, Cynthia Parr	100
19.36. Development of the Generation Change Program Domain Models and Ontology	
Management Embedded in the GCP Platform	
Rosemary Shrestha <sup>1</sup> , Jeffrey Detraas <sup>2</sup> , Thomas Metz <sup>2</sup> , Martin Senger <sup>2</sup> , Graham McLare	
Elizabeth Arnaud <sup>3</sup> , Tom Hazekamp <sup>3</sup> , Adriana Alercia <sup>3</sup> , Guy Davenport <sup>1</sup> , Reinhard Sim	ion <sup>4</sup> ,
Jayashree Balaji <sup>5</sup> , Richard Bruskiewich <sup>2</sup>	101
19.37. Data Models To Share Invasive Species Information Through The Global Invasive	e
Species Information Network	
Annie Simpson <sup>1</sup> , Elizabeth Sellers <sup>2</sup>	102
19.38. Xper <sup>2</sup> : an open tool to manage descriptive data	
Visotheary Ung <sup>1</sup> , Régine Vignes-Lebbe <sup>2</sup>	103
19.39. The SCAR Antarctic Biodiversity Database of terrestrial and limnetic data	
Dave Watts	103
19.40. An Evaluation of Taxonomic Name Recognition (TNR) in the Biodiversity Herita	ge
Library	
Qin Wei <sup>1</sup> , Chris Freeland <sup>2</sup> , P. Bryan Heidorn <sup>1</sup>	104
19.41. The Global Compositae Checklist – integration of nomenclature and taxonomic	
concepts from multiple data sources	
Aaron Wilton <sup>1</sup> , Kevin Richards <sup>1</sup> , Jerry Cooper <sup>1</sup> , Christina Flann <sup>2</sup>	105
20. Contributed Abstracts	
20.1. Managing Biodiversity Data and Species File Software	
David C. Eades	106
20.2. PlantCollections – A Community Solution	
Boyce Tankersley <sup>1</sup> , Greg Riccardi <sup>2</sup> , David Vieglais <sup>3</sup> , Min Cai Henderson <sup>1</sup> , Christopher	
Dunn <sup>4</sup> , Dan Stark <sup>5</sup> , Pam Allenstein <sup>6</sup>	107

### 1. Introduction to TDWG

#### 1.1. TDWG – Looking Back and Looking Forward

Donald Hobern CSIRO Entomology

TDWG has been meeting for 23 years. Over this time it has achieved many things and made significant contributions to the development of biodiversity informatics to become the field it is today.

TDWG started with botanists and biologists seeking the best approaches to managing their institutional databases. Technologies have opened up increasingly rich opportunities to expand this focus. In recent years, the key focus of TDWG has been on developing tools and standards to facilitate integration of biodiversity data to support scientific research, policy development and the needs of the public.

TDWG has evolved to focus more and more on the IT aspects of its work. People who identify themselves as software engineers or information technologists are increasingly involved with TDWG. However, the long-term vitality of the organisation depends critically on the continuing involvement of practicing taxonomists, collection managers, ecologists and other domain experts who understand the origins and use of biodiversity information. TDWG's strength depends on the expertise of both domains.

To achieve this goal, TDWG needs to continue focusing on some basic requirements:

- 1. Improved outreach to biologists and users of biodiversity information. TDWG needs to articulate its goals to these wider communities and to improve its understanding of the practical issues hindering these users from sharing, discovering and using relevant data.
- 2. Simplifying the tools and methods offered to data providers and users of biodiversity data. Making use of TDWG data standards should not depend on our users being prepared to master XML technologies or web server configuration.
- 3. Separation of concerns. TDWG should build on its work in recent years developing a common architecture and ontology for biodiversity data. This approach assists TDWG members to contribute in the area of their personal expertise. Biologists should be able to discuss and define the data items they need to be able to exchange. Software engineers should be able to develop tools and protocols for transporting these elements from provider to user without exposing the complexities of their solutions. TDWG should be the meeting place where this happens.

I urge all participants at this year's conference to consider how we can all contribute to TDWG being the venue for productive cross-fertilisation between biological science and information technology.

#### **1.2. TDWG Standards Roadmap**

Roger Hyam

#### TDWG/Royal Botanic Garden Edinburgh

It is a goal of biodiversity research to combine data of different kinds from multiple sources. The providers of data do not know who will use their data or how it will be combined with data from other sources. The consumer needs some level of commonality across all the data received so that it can be combined for analysis without the need to write bespoke computer software for every new data combination. This commonality needs to seamlessly extend to new types of data as they are made available. An architecture is required to provide this commonality.

A degree of commonality could be achieved simply by specifying how data should be serialised (converted to files). If all suppliers passed data as well-formed XML, for example, it would provide a degree of interoperability, but clients would still not know how the elements within one XML

document relate to those in another or how the items described in those documents were related. At the other extreme, the architecture could provide a detailed data type library which described the way in which each kind of data should be serialised at a fine level of granularity – which XML elements must be present and what they should contain. It is, however, highly unlikely that a single set of serialisations would meet all needs any more than a single federation schema would. It remains a requirement of some networks that they have well defined data types to ensure that the data passed is valid and fit for purpose.

The architecture therefore has to meet two needs. It has to allow generic interoperability and a restricted validation of data for some networks. It does this by taking a three pronged approach:

- 1. An ontology is used to express the shared semantics of the data but not to define the validity of those data. Concepts within the ontology are represented as URIs (Universal Resource Identifiers).
- 2. Exchange protocols use formats defined in XML Schema (or other technologies) that exploit the URIs from the ontology concepts.
- 3. Objects about which data are exchanged are identified using Globally Unique Identifiers (GUIDs).

This means that (although exchanges between data producers and clients may make use of different XML formats) the items the data are about and the meaning of the data elements is common across all formats.

Progress has been made since this architecture was proposed two years ago. A roundup of past developments is summarised in the TDWG 2008 Roadmap that will be distributed at the conference.

# 2. From Small to Big: How to Capitalize on Small Contributions to the Big Dreams of Sharing Biodiversity Information

#### 2.1. Introduction: From Small to Big

Gail E. Kampmeier Illinois Natural History Survey

Amidst the grand initiatives and prospects of quickly gathering the "low-hanging fruit" represented by the big players and data providers, the needs of individual and small group contributors of biodiversity information have been neglected. Ignoring these small contributors, who often represent taxon specialists, producers of agricultural datasets, or compilers of geographically-based faunal inventories, misses rich content that could be used for a variety of high quality analyses, *e.g.*, contributing towards informed decisions for sustainable use of our biotic resources. This neglect poses a great challenge for the global biodiversity informatics community to find a way to balance the needs of the smaller contributors whose immediate goals are more focused on the specimens on their bench than sharing their findings over the internet.

This session will address issues of content mobilization; attempts to meet infrastructural and technical challenges to contributing and sharing biodiversity information; and facing the realities of biodiversity informatics in the developing and mega-biodiversity world.

Support is acknowledged from: U.S. National Science Foundation; University of Illinois

## 2.2. Every datum counts! - Capitalising on small contributions to the big dreams of mobilizing biodiversity information

#### Vishwas Chavan, Éamonn Ó Tuama, Samy Gaiji, David Remsen, Nicholas King Global Biodiversity Information Facility

Access to, and sharing of, authentic, appropriate, up-to-date, and required (sufficient-to-analyse) volumes of data are essential for biodiversity conservation and sustainable utilization of biotic resources. However, the fact is that both biodiversity and its data are unevenly distributed and unequally accessible across the globe. An untapped wealth of data exists in isolated and non-interoperable formats, which makes it impossible for stakeholders to gain access when the data are most needed. These datasets are also most often in danger of being orphaned or lost. While one of the major challenges to realizing big dreams of mobilizing biodiversity information is to overcome socio-political barriers to sharing information, another often neglected challenge is to develop a standards-based and technical infrastructure that is hassle-free for both big and small contributors to share and exchange data on biodiversity.

Through major international and national initiatives, attempts are being made to address the problems faced by major data providers. However, amid these major initiatives, concerns and problems faced by smaller contributors are often ignored. Yet every "datum" counts when conservation and sustainable use of biotic resources is the goal. One of the reasons for the exclusion of smaller contributors is the inadequate, inappropriate (to them), and complex set of existing standards.

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

Therefore, TDWG, as a global standards community, together with the big biodiversity informatics projects and players, and national and international science funding agencies, are obliged to address impediments faced by the small and mid-sized contributors. This requires identification of (i) current standards/protocol barriers, (ii) what short-term and long-term standards/protocols are required, (iii) data sets that may become lost or orphaned because of perceived barriers in the implementation of standards (or, how can standards help prevent datasets from becoming orphaned?), (iv) mechanisms for contributor-friendly tools for working with standards/protocols within the research data management chain, and (v) concerns for data sensitivity through standards implementation.

This presentation while highlighting the concerns and realities faced by small yet valuable contributors, will also investigate the need for hassle-free and easy-to-implement standards and protocols for mobilising "fit-for-use" primary biodiversity data.

Support is acknowledged from: Global Biodiversity Information Facility

#### 2.3. Experiences in mobilizing biodiversity data - the story of HermesLite

Piers Higgs, Tim Carpenter Gaia Resources

While distributed networks of biodiversity information are being developed and are starting to deliver on their goals, there remain many hurdles to be overcome for individual institutions in delivering data to these networks. Organisations face hurdles in the form of organizational policies, missing or unavailable infrastructure, a lack of skills and resources, and a range of issues relating to managing people and their expectations.

HermesLite is an open source tool set created as a result of a project for the Western Australian Museum to assist in the mobilization of data outside of the institution. HermesLite's validation framework performs customizable checks for correct types and rational values for data, *e.g.*, a

longitude of "401" or "Four hundred and 1" is likely to be incorrect. HermesLite integrates with a variety of data serving tools such as TAPIRLink and MapServer to make the data accessible to a range of audiences and requires minimal changes to an organisational infrastructure that already delivers web content.

HermesLite was developed as an open source product that is freely available for other organizations to use to assist with their own challenges, and is available as a Sourceforge project as well as from the Gaia Resources web site (<u>http://www.gaiaresources.com.au/hermeslite/</u>) The authors will present their experiences in specifying, funding, developing, and implementing HermesLite and will also share some insight into mobilizing data from institutions of various sizes.

Support is acknowledged from: Western Australian Museum, Australian Biological Resources Study

#### 2.4. Small pieces loosely joined: Building scientific web communities with Scratchpads

#### Simon Rycroft, Dave Roberts, Kehan Harman, Vince Smith

Scientists drove early development of the World Wide Web, primarily as a means of rapid communication, document sharing and data access. They have been slower to adopt the web as a medium for building research communities. Yet web-based communities hold great potential for accelerating the pace of scientific research. In this presentation we describe our work on the Scratchpad project – a tool for developing taxonomic research communities to build, share and manage biodiversity data on the Web. In less than two years the Scratchpad project has enabled the self-assembly of more than 70 research communities with over 700 registered users. Collectively these scientists have built more than 130 thousand pages of content. By making it easier for groups of taxonomists to self assemble around a common interest, and for individuals to contribute to group effort without formal management, Scratchpads are removing the traditional barriers to the size, sophistication and scope of unsupervised efforts in biodiversity science. These barriers are: 1) That knowledge about natural history is difficult to aggregate, *i.e.*, the transition from biodiversity data to knowledge is difficult to negotiate. Scratchpads provide a flexible infrastructure that supports alternative ways of making this transition. 2) That taxonomists work independently and informally they do not like being told how to work or what to do. Scratchpads provide the freedom for taxonomists to work in different ways, at their own pace, without necessitating consensus. 3) Scientific work across countries and institutions is complicated by legal and organizational barriers, principally having to do with intellectual property. Scratchpads assign ownership to the scientists generating the content, but enforce a (Creative Commons) licensing framework through which others can reuse this output.

Scratchpads do not remove all the barriers to sharing rich contextualized data and knowledge on natural history. Issues of complexity loom large when aggregating heterogeneous content across different Scratchpad sites. However, as we will demonstrate, data mining tools provide alternative strategies for keeping this complexity under control. Ultimately the amplification of group effort enabled by the social networking tools like the Scratchpads has the potential to transform the way we document and describe biodiversity. Scratchpads are a working example of how small contributions by taxonomists can be loosely joined to help realize the big dreams of sharing biodiversity knowledge.

### 2.5. Designing a global network to accommodate contributions from all sources and technical abilities

Tim Robertson GBIF

Overview

Networks such as the Global Biodiversity Information Facility (GBIF) have managed to mobilise vast quantities of data by deployment of tools based on standards and protocols defined by the Biodiversity Information Standards (TDWG) group. While these standards and the protocols used to provide the data to GBIF have matured over time, it remains a challenge for the ordinary bench scientist and smaller collections and institutions with limited technical capacity to contribute their data. While important for the standards to evolve, it remains imperative that there exists a simple mechanism to contribute data.

#### Today

This presentation provides an overview of the GBIF network, illustrating the minimum 'entry requirements' for a data custodian to contribute today. The data flow to the global index will be explained, highlighting some of the bottlenecks and problematic areas that currently exist. Proposals for change will be given for standard and tool development that would allow for an easier contribution path for those with limited technical capacity. Furthermore, a streamlined approach, under development with the forthcoming GBIF provider tool, will demonstrate how providers could share an entire dataset.

#### Looking to the future

Some of the challenges faced by GBIF relate to:

- o inclusion of additional types of data elements;
- o broadening the scope of datasets contributed;
- o increasing value to users; and
- o capitalising on advancing internet technologies.

All of these require a continual evolution of standards and tool development. During this development we must ensure that the bridge between the scientists' data modeling capability and technical advancements is managed sufficiently so that all may contribute.

The definition of standards could be done in a manner that is easier to digest for those who are not technically focused. Current TDWG vocabulary definitions will be compared with those published by the International Standards Organisation (ISO).

An invitation will be extended to join a highly focused task group committed to delivering to TDWG in 2009 a proposal for a transfer schema for occurrence data. This schema will go beyond the capabilities of existing schemas to extend the scope of data that can be shared, and the group will aim to deliver the schema with a reference implementation (a working example illustrating the successful use of the schema).

Support is acknowledged from: Global Biodiversity Information Facility

### 3. Introduction to TDWG Standards

#### 3.1. An Introduction to Scientific Names of Organisms

Richard Pyle Bishop Museum

Although most people (and certainly most scientists) are familiar with the practice of assigning scientific names to organisms, many are not aware of the history of the practice, the current rules and regulatory bodies for assigning scientific names to organisms, the subtle but important distinctions between "taxonomic names" and "taxonomic concepts" (and the corresponding implications to biodiversity informatics), or the minefield of potential pitfalls surrounding the ambiguous and inconsistent terminology used in associated discussions.

There are five major realms of scientific names for organisms, each associated with a particular "Code" of practice (Botanical, Zoological, Bacteriological, Viral, and plant cultivars). Although these five realms are reasonably well-delineated, there is some area of overlap and/or uncertainty about which Code certain kinds of organisms fall under. Moreover, each Code has a different starting date, different regulatory bodies and mechanisms, and different rules in force for assigning new names and treating existing names.

From an informatics perspective, there has been a great deal of discussion on the difference between scientific names of organisms, and the taxonomic concepts they are intended to represent. Although these issues have long been recognized, the biodiversity informatics community has paid particular attention, in terms of database object definitions and information management approaches. Although much progress has been made addressing these issues, there is still some confusion about contexts and priorities, and semantic ambiguities remain.

#### **3.2.** An Introduction to Data Exchange Protocols in TDWG

#### Renato De Giovanni

Historically, most of TDWG's activities have been dedicated to creating standards to represent biodiversity data. Standards were created to represent biological descriptions, taxon concepts & names, specimen & observation records. Standards are currently being created for collections descriptions, literature data and invasive species, among others.

All these standards use a common approach based on XML (Extensible Markup Language) to structure data into a widely accepted format. Standards are increasingly important with the growing need to access, integrate and analyze the volume and variety of biodiversity data becoming available. Using such standards leads to reduced effort, cost, time and risk requirements when implementing new software applications.

During the last years, protocols for mobilizing biodiversity data have also been discussed in TDWG. Initially, this happened in the context of federated networks consisting of independent data providers sharing data using predefined data formats and protocols. Such networks needed a standard way to discover, search and retrieve distributed data over the Internet, allowing new data portals and applications to be developed. TAPIR, the TDWG Access Protocol for Information Retrieval, was created for this purpose by unifying two existing similar protocols – BioCASe and DiGIR.

As the existing data infrastructure grows, integration between different federated networks combining different kinds of biodiversity data becomes a common requirement for carrying out more complex analysis and research. In order to deal with this scenario, the TDWG Technical Architecture Group created an ontology defining semantics and relationships for some of the most relevant classes of objects that are found in TDWG data standards. The strategy also includes recommendations for using Global Unique Identifiers and data exchange protocols.

This presentation provides an overview of how to exchange data over the Internet using TDWG standards, with particular emphasis on TAPIR.

BioCASe: <u>http://www.biocase.org/products/protocols/</u> DiGIR: <u>http://digir.net/</u> TAPIR: http://www.tdwg.org/activities/tapir/executive-summary/

#### **3.3.** An introduction to biological descriptions

Kevin Thiele Western Australian Herbarium

Capturing biological descriptions in standard format is a hard problem; in fact, a claim will be made that descriptions are an order of magnitude harder than most other areas of biological standardization.

This is because any standard must include both the descriptive data and an ontology (the descriptors) from which the description gains meaning. Descriptors are domain-specific – descriptors needed for beetles will necessarily be different from those for orchids (in fact, two beetle families may well require very different descriptors). Hence, an ontology for life is impossible – or is it? The principal challenge in descriptive standardization, as in all TDWG activities, is to create a standard which will allow descriptive data from a variety of sources to be combined, and to allow a bounded inference that "yellow" and "rounded" in descriptions from two sources mean the same thing. This in turn requires agreement on normativity and the capture of normativity statements. Once descriptive data are captured in standard form they can be used for many purposes, including the generation of structured natural language, interactive identification keys and research into the distributions and ecology of features of interest. The TDWG Structure of Descriptive Data standard goes a long way to solving the difficult issue of capturing descriptive data, but more work is needed.

#### 3.4. Globally Unique Identifiers in Biodiversity Informatics

Kevin James Richards Landcare Research

The successful development of interoperable networks depends on the ability of clients to uniquely identify and locate data items provided by multiple sources. To address this need, TDWG has adopted Life Science Identifiers (LSID) for use as stable identifiers for data items in Biodiversity Informatics data networks. This presentation provides an introduction to Globally Unique Identifiers (GUIDs), in particular LSIDs, what they are, and how they are used.

"Why do we need GUIDs", "to what objects we should apply GUIDs", "how do we resolve a GUID", "how do we reuse GUIDs" are common questions that will be addressed in this talk. The presentation is intended to provide an introduction for people who have had limited exposure to these technologies.

### 4. Biodiversity Informatics: Key Organizations and Projects

#### 4.1. GBIF: From Prototype towards Full Operation

Éamonn Ó Tuama, Vishwas Chavan, Markus Döring, Samy Gaiji , David Remsen, Tim Robertson

Global Biodiversity Information Facility

This presentation provides an overview of the core informatics activities in the 2009-2010 GBIF work programme necessary to implement the GBIF Strategic Plan 2007-2011. In advancing "from prototype to full operation", GBIF recognises the need to move to a more distributed network based on the active engagement of more autonomous nodes, thereby enabling scalability and diversification beyond what a centralised infrastructure dependent on the limited capacity of the Secretariat can provide. Acting as a catalyst for change, GBIF will draw on the experience of its more advanced nodes and will seek to capacitate all participants who recognise that biodiversity knows no boundaries, and that working together towards the common good of a successful network shall benefit all.

Informatics and Participation comprise the two thematic areas in the 2009-2010 work programme. Both are closely aligned, with Participation undertaking the crucial role of conveying to Informatics the needs of Participants for products and services, and, once developed, channeling them back to Participants, in a continual process of improvement. GBIF can thus play a central role in TDWG, by representing the informatics requirements and experiences of the broader biodiversity community, thereby helping to set priorities around standards and interoperability issues for "real world" practitioners.

GBIF's informatics programme now consists of three interconnected work areas:

- 1. Inventory, Discovery, Access (IDA); focusing broadly on standards for interoperability, and, in particular, on a metadata strategy for GBIF that includes the development of dataset inventory tools to enable users to discover and access biodiversity resources.
- 2. Content; consisting of the DIGIT and ECAT work areas, the former driving the mechanisms, both technical and sociological, to mobilise primary biodiversity data, and the latter developing a comprehensive Global Names Architecture; together, both work areas will inform the development of data provider tools and web services that allow Participants to assess and improve the quality and fitness-for-use of the data they serve and use.
- 3. Infrastructure and Data Portal; the GBIF data portal will continue to be developed and enhanced as the focal point through which most users interact with the GBIF network, providing the metadata and data index for the whole network, facilitating data providers in getting data online, and users accessing data from across the network.

The main informatics components of the GBIF work programme, planned, or already underway with the input of dedicated task groups, are briefly described and cross-referenced to other GBIF events and presentations in this conference. These components include 1) the Global Biodiversity Resources Discovery System, 2) the Global Names Architecture, 3) the GBIF Data Portal, 4) the Provider Toolkit, 5) the Harvesting/Indexing Tool, and 6) the Node Portal Toolkit.

Support is acknowledged from: Global Biodiversity Information Facility

#### 4.2. DataNetONE: A Distributed Environmental Data Archive

Matthew B. Jones NCEAS, UC Santa Barbara

Ecological, environmental, and earth science studies that elucidate the processes driving biodiversity are fundamentally important to science-based management of the natural world. Understanding the patterns and processes that control biodiversity is critical to the long-term management of our world's ecosystems. Scientific programs collect data on the distribution and abundance of organisms and the natural environment in which they exist. Currently these data are collected by a wide variety of academic and governmental agencies distributed throughout the world. However, there is no mechanism for broadly federating them to achieve accessibility. We have proposed to build the DataNet Observation Network for Earth (DataNetONE), a global, distributed, environmental data archive spanning the environmental sciences. DateNetONE will make the wealth of scientific observations of the earth available for scientific studies. DataNetONE has been proposed by a consortium of institutions that are active in scientific data archiving, including ecological research centers, field station networks, government agencies, and libraries.

DataNetONE will consist of a large number of geographically distributed Member Nodes that house data archives and metadata describing those data. Member Nodes will be linked together by contributing metadata to a series of replicated Coordinating Nodes that provide valuable services to Member Nodes. The Coordinating Nodes provide a common infrastructure to handle, for example, distributed authentication, fault tolerance, geographic, taxonomic, and temporal search services, and data replication services. The Coordinating Nodes will monitor the health and accessibility of the Member Nodes in order to create a carefully architected replication service by which Coordinating Nodes will ensure that all data are replicated among geographically separated institutions. This replication service will ensure the long-term preservation of data and make the data accessible in different areas of the world.

DataNetONE is an ambitious project, and will only succeed through careful planning and widespread support for its mission. A major focus will be on the sustainability of DataNetONE, with the goal that the network has both the financial and technical means to be self-sustaining after ten years. This sustainability goal is critical to the preservation of data that can be used to understand biodiversity and our natural world.

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

#### **4.3.** European Distributed Institute of Taxonomy (EDIT)

Walter G. Berendsohn Botanic Garden and Botanical Museum Berlin-Dahlem (BGBM)

The European Distributed Institute of Taxonomy (EDIT) was introduced to the TDWG audience in 2006 and 2007. It is a 5-year project of 26 leading natural history institutions and organisations (started in March 2006) financed by the European Commission.

Much of the project's resources are devoted to issues in biodiversity informatics, specifically addressing the needs of taxonomists and their work at member institutions. Workpackage 5 of the project is coordinated at the BGBM in Berlin and attempts, among other tasks, to enhance cooperation of the participating institution's IT departments. The central effort is aimed at the creation of an "Internet Platform for Cybertaxonomy". This is envisioned as a distributed computing platform that will directly assist taxonomists in their daily work, increase the efficiency of producing taxonomic revisions, checklists, inventories, etc., and make the results of the work accessible without delay.

The first year of EDIT was largely spent at setting up the organisation and defining the aims of the Platform project. The actual software development phase is now nearing halftime. An international team of 16 developers employed by the project and actively supported by institutional staff presently works on realising the vision. With the team distributed among nine institutions, coordination of development becomes a serious challenge. Adding to this is the close cooperation with other projects, such as the Pan-European Species directories Infrastructure (PESI, <u>http://www.eu-nomen.eu/pesi/</u>), and CATE (Creating a Taxonomic E-science, <u>http://www.cate-project.org/</u>). Taxonomists from EDIT institutions and from these projects form the audience EDIT products will be delivered to.

Considerable progress has been made over the past year, and a series of aspects of the platform will be presented in contributions during this TDWG meeting. The "Platform version 1", to be presented by the end of November 2008, will, for the first time, present an integrated view of platform components. EDIT's software development for the platform is based on a Common Data Model (CDM), firmly grounded in TDWG standards, and the CDM library, an open source shared domain model library in Java that can be used as a foundation for many other Java based biodiversity projects.

We hope that once EDIT's software is in use, other taxonomic software developers will integrate their products as interoperable components of the platform. More information is available at <u>http://wp5.e-taxonomy.eu/</u>.

Support is acknowledged from: European Commission

#### 4.4. The Biodiversity Collections Index – Linking 140 Million Records?

Roger Hyam TDWG

There are approximately 140 million Darwin Core and ABCD specimen records available through the GBIF data portal and possibly more in circulation. They all make use of the CollectionCode field. This field is used to indicate the location of the specimen. Unfortunately there is no controlled vocabulary for the data in this field. This means that a data consumer cannot be certain whether any two records represent specimens from the same reference collection or different reference collections.

The Biodiversity Collections Index (BCI) has been built as a central index to all collections of materials used in biodiversity research. It provides a Globally Unique Identifier (GUID) for every natural history collection on earth. These GUIDs can act as a controlled vocabulary for the CollectionCode field.

The advent of a single, controlled vocabulary will allow all specimen records to be related to each other on the basis of the locations of the specimens they represent. In addition:

- Researchers can use the IDs provided by BCI to unambiguously indicate the location of the voucher specimens they have used in their published works.
- Collection Curators can use BCI IDs to track the use of materials from their collections. BCI also helps curators tag individual specimens with their own GUID by providing shared use of its GUID resolution services.

The data in BCI comes from two sources. Any member of the biodiversity research community can register and contribute to the data held in the index. In addition to this, authoritative data that has been curated by established sources is displayed in a non-editable form alongside the community data.

BCI is a collaborative effort. It was instigated in October 2007 as a response to the work on Globally Unique Identifiers and Natural Collections Descriptions that was going on within TDWG. It was formalised as a project on 5th December 2007 with the signing of a Memorandum of Co-operation between the Global Biodiversity Information Facility (GBIF), TDWG and the Royal Botanic Garden Edinburgh.

The initial authoritative data available in the system comes from four principle sources. Index Herbariorum (IH), itself a collaborative effort supported by The New York Botanical Garden (NYBG), has provided seed data about herbaria. IH data are presented as authoritative, non-editable data within BCI and will be periodically updated from the official IH database maintained by NYBG. Insect and Spider Collections of the World (ISCW) has provided seed data on entomological and arthropod collections and is treated in a similar way to IH. Biorepositories.org is a Barcode of Life Initiative. BCI reflects the confirmed records from the Biorepositories database as an authoritative source of data.

BCI is accessible on the web at <u>http://www.biodiversitycollectionsindex.org</u>. For technicians BCI also exposes data via a series of web services, including a Life Science Identifier (LSID) Authority and associated HTTP proxy service, a REST service, OAI-PMH syndication, CSV downloads and an experimental JSON service. Details of these services are given on the website.

All members of the biodiversity informatics community are invited to make use of the BCI website and web services, to register, contribute data and suggest improvements. BCI is here to serve you.

Support is acknowledged from: GBIF, Royal Botanic Garden Edinburgh, TDWG

#### 4.5. KeyToNature Species Identification e-Tools for Education

Dr Peter Schalk<sup>1</sup>, Professor Pier Luigi Nimis<sup>2</sup>, Ir Wouter Addink<sup>1</sup> <sup>1</sup> ETI BioInformatics, <sup>2</sup> Università di Trieste

The identification of organisms is fundamental to biology. Accurate identification provides a correct name through which detailed information can be unlocked. Such information includes taxonomic descriptions, ecological relations, economic values, conservation status, legislation status and genetic code.

The EC funded KeyToNature (K2N) network (4.8 million Euro, 14 partners in 11 countries) integrates the approach to digital taxonomic identification systems in Europe. Standardization of data, exchange protocols, algorithms, software applications and intellectual property rights will provide a basis for collaboration in the K2N network. In turn, this will allow the output of innovative 'custom' identification services via the Internet and mobile phone for a wide range of target groups. Initial demonstrations of the network have been developed.

The main objectives of K2N are to:

- o Increase access to, and simplify the use of eLearning tools for identifying species;
- Address educational content interoperability;

- Optimise the effectiveness and the quality (including provision of multilingual access) of educational content; and
- Suggest best practices to address barriers to educational use, production, exposure, discovery and acquisition.

K2N adopted the TDWG standard for descriptive data (SDD) as one of the solutions to achieve its goals. A challenge in K2N will be to adapt SDD to 'atomized' identification data and to standardize species identification data sets so they can be combined and re-used in packages adhering to the standard. SDD will be used at a large scale in K2N to exchange descriptive data in production quality tools.

K2N mobilises a critical mass of content: Less than a year after K2N started 1,320 identification keys in eleven languages are listed at <u>http://www.keytonature.eu</u>. Many of these keys are custom made versions derived from a central K2N data set.

The K2N network focuses on collaborative identification services using shared data resources and tools, with a particular focus on education. Multilingual access will enhance delivery and integration in e-learning environments and curricula. The long-term goal is to enhance the knowledge of biodiversity at all educational levels in an innovative way. The target audiences are primary and secondary school students and teachers and university students and lecturers.

Apart from primary identification data (characters and states), K2N will also unlock a wealth of secondary data (illustrations, descriptions) available in Europe, using the same shared mechanisms. A survey of secondary data identified 241,421 items, mainly still images (172,426) and taxon pages (59,004). Both primary data and secondary data will be made available using a new integrated approach that permits automatic production of identification tools for different media: the Internet, mobile and desktop applications.

K2N follows a user-centric approach. More than 200 teachers from primary school to university participated in focus groups and interviews conducted in all partner countries. This survey resulted in a pedagogical and educational needs analysis report. In 2008/2009 the K2N project will conduct several experiments with web-based and mobile identification services for educational applications. For example, the 'Mobile Nature Guide' for schools in The Netherlands provides access to identification data and user-defined keys covering more than 5,000 plants, animals and fungi.

Support is acknowledged from: EC, eContentPlus, KeyToNature

### 5. Biodiversity Informatics Literature and More

#### 5.1. Plazi: The technical set up

Guido Sautter<sup>1</sup>, Donat Agosti<sup>2</sup>

<sup>1</sup> Institut für Programmstrukturen und Datenorganisation, Universität Karlsruhe, <sup>2</sup> Plazi

Plazi is a Swiss non-profit whose goal is to digitize and OCR biosystematics publications, mark them up with XML reflecting both logical structure and important details, and then make them accessible through both a public search portal. The structural markup includes taxonomic treatments, some inner structure and paragraphs. The detail markup comprises taxonomic names (resolved to their full forms and attributed with LSIDs) and geo-referenced collecting locations.

Anyone with a login for Plazi can upload documents and contribute to the treatment collection, thus a community of users can work together to extend the collection hosted at Plazi.

ABBYY FineReader software is used to transform page images into text containing HTML markup reflecting the printed layout. GoldenGATE is used to create XML documents with semantic markup from the OCR results. GoldenGATE is an editor designed for users who are not XML experts. It

allows for editing XML documents at the element level rather than the syntax level, allowing users to work with the data without understanding the details of XML syntax.

Even with this assistance, doing the complete markup manually would require prohibitive effort. Thus, we use Natural Language Processing (NLP) components to create most of the markup automatically. Users verify the NLP results and correct errors. GoldenGATE seamlessly integrates automated markup components through a slim plug-in interface. To simplify checking and correcting NLP results, further plug-ins provide specialized views on the documents, *e.g.*, a list view of the XML elements of a given type, helpful for finding false positives after the automated markup of taxonomic names. Additional plug-ins fetch data from external sources and store them in XML attributes, including LSIDs for taxonomic names (from Hymenoptera Name Server, Index Fungorum and (soon) ZooBank) and geographic coordinates for collecting locations (from a variety of web services).

Once the documents are marked up, we store them in the Search & Retrieval Server (SRS) at http://plazi.org:8080/GgSRS/search. This works directly through a connector plug-in in the GoldenGATE Editor. The SRS extracts individual treatments from documents and stores them in its archive. The treatments themselves are stored in the file system; management data goes into a PostgreSQL database. Then the SRS creates index entries from marked-up taxonomic names, collecting locations, and bibliographic metadata. The indices also reside in the database.

The web front-end of the SRS makes the marked-up documents publicly accessible. For humans it provides searching and browsing. The search form is generated directly from the indexing tables in the backend database. Taxonomic names offer links to additional descriptions in other sources (based on LSIDs). A link to GoogleMaps allows users to view geo-referenced collecting locations. Users can list compilations of index data without looking at complete treatments, *e.g.*, a list of all taxa of a given genus with their collecting locations.

For machine access, the web front-end offers the treatments as TaxonX XML documents. The data are also accessible through a TAPIR service (<u>http://plazi.org:8080/GgSRS/tapir</u>), which provides taxa, materials citations (which are pulled and indexed by GBIF), and whole treatments. An RSS feed for treatments exists to simplify drawing data from the server for applications unaware of TAPIR, like Google web crawlers.

#### 5.2. Plazi: Prospects for Mark Up of Legacy and Prospective Publications

#### Terry Catapano Plazi/Columbia University

Plazi has been engaged in the digitization of taxonomic literature for several years. As we have discussed at prior TDWG conferences, we have developed an XML schema, taxonX, for the markup of treatments in which to date we have encoded roughly 5000 treatments. We will discuss the significant challenges encountered and make some recommendations on approaches further work.

Perhaps the most important consideration in digitization of taxonomic literature is the recognition of the different problems presented by legacy and prospective literature. Existing historical literature presents a high degree of formal variation which makes modeling in an XML schema quite difficult. Any resulting schema ought to necessarily refrain from strict specification of document structures in order to accommodate the many and various ways in which treatments have been presented over the past two centuries. There will also likely be need to include robust mechanisms for the normalization of data in the source documents. Another peculiar problem is determining a policy for transcription of the original text. Should corrections be made? If so, with what indications of such "editorial" activity should be included? The variability of the form, as well as quantity and quality of data will limit applications aiming to use XML encoded legacy treatments. Only so much can be done at the level of the markup. The greater share of the burden will fall on "downstream" processes to achieve interoperability.

New born-digital literature presents a different set of considerations and possibilities. Document modeling certainly is easier. Authors can be given a fairly strict set of options in creating new treatments. The strict structures, being defined and predictable, are far more easily processed. Plazi has had some experience this year preparing XML versions of new articles for Zootaxa and PLOS One. The experience taught us that in order to benefit from the advantages to be gained from electronic publication it is preferable to create an XML version from the start, rather than converting from a manuscript in a conventional word processing format as we did. Otherwise, one replicates the same process and encounters the same problems and limitations of encoding legacy materials. If richly encoded at an early stage, the data in the treatment can be repurposed much more readily and for a wider range of applications. To more fully exploit the possibilities of digital literature, we advocate changes to nomenclatural codes to shift the concept of the taxonomic treatment from that of a simple textual publication object to a complex digital object containing both text and data. In doing so, consideration should be given emerging standards for the modeling complex digital objects such as the Metadata Encoding and Transmission Standard (METS; http://www.loc.gov/mets), Open Archives Initiative - Object Reuse and Exchange (OAI-ORE; http://www.openarchives.org/ore/), and MPEG-21/Digital Item Declaration Language (DIDL;

http://www.iso.org/iso/iso\_catalogue/catalogue\_tc/catalogue\_detail.htm?csnumber=35366).

#### 5.3. An evaluation of taxonomic name finding and next steps in BHL developments.

#### Chris Freeland Missouri Botanical Garden

The Biodiversity Heritage Library (BHL) is an international consortium of natural history libraries that are actively scanning scientific literature from their collections and publishing the digitized content online at <u>http://www.biodiversitylibrary.org</u>. The presentation will cover updates on BHL progress thus far, including a review of the BHL portal application and services developed for bibliographic data exchange, as well as a roadmap of features scheduled for design and development. Of interest to taxonomists and bioinformaticians, the presentation will also cover a recently completed evaluation of taxonomic name finding algorithms, including TaxonFinder and FAT. Discussion will include a review of successes and failures in extracting taxonomic name strings from the BHL corpus, as well as suggestions for improvement.

Support is acknowledged from: MacArthur Foundation, Moore Foundation, Sloan Foundation

# 5.4. Precision in accessing the descriptions, keys and other contents of digitized taxonomic literature: the INOTAXA prototype

Chris Lyal<sup>1</sup>, Anna Weitzman<sup>2</sup> <sup>1</sup> The Natural History Museum, <sup>2</sup> Smithsonian Institution

Access to legacy and new taxonomic literature through digital means is becoming the norm. Against a background of developing digital internet search power, this brings with it the potential and userdriven requirement to locate and extract information and data with increasing precision and speed. This in turn requires both highly atomized content and a powerful and flexible interface. The system described here, INtegrated Open TAXonomic Access (INOTAXA), is now in prototype form, and shows the potential of such an interface. Searches take place across all content or a specified subset. Any term can be used for a search, and the context in which it is found restricted. Alternatively, Boolean logic can be used across 53 indexed fields, or the full content browsed taxonomically, geographically or by person. Thus INOTAXA allows extraction of parsed data on names, authors, places of publication, places of use; specimens cited, nomenclatural types, relationships (taxonomic, nomenclatural, phylogenetic and ecological) with other taxa, etc. Taxon names may be restricted to valid (accepted) names only, synonyms or, of course, all names may be returned. Specimen data, extracted from the literature according to user-set conditions, may be viewed and downloaded. In addition to fine-detail content treatments and keys can be retrieved, again according to more or less complex criteria and restrictions, according to user needs. Following testing by a selected panel from different taxonomic and other backgrounds, the prototype is now available on the web for testing by the wider community at http://www.inotaxa.org. Feedback is requested to allow further development and refinement.

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

### 6. Ontologies

#### 6.1. Ontologies and interoperability in evolutionary comparative analysis

Francisco Prosdocimi<sup>1</sup>, Brandon Chisham<sup>2</sup>, Enrico Pontelli<sup>2</sup>, Julie D Thompson<sup>1</sup>, Arlin Stoltzfus<sup>3</sup>

<sup>1</sup> Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), Department of Structural Biology and Genomics (IGBMC), Illkirch, <sup>2</sup> New Mexico State University, <sup>3</sup> National Institute of Standards and Technology

The methodology for evolutionary comparative analysis has developed over several decades. The methodology now includes familiar and powerful software tools for inferring phylogenetic trees, reconstructing ancestral states and comparing models of evolution. Lagging far behind however, are the informatics resources to facilitate integrative, large-scale, or automated analyses, especially for data other than molecular sequences. Because the informatics challenges of evolutionary analysis are similar to the challenges of interoperability and data integration experienced in other areas of bioinformatics, they may benefit from ontologies and data standards used in these other areas.

In order to facilitate development of interoperable systems, members of the NESCent Evolutionary Informatics working group (<u>http://www.nescent.org/wg/evoinfo</u>) have developed a Comparative Data Analysis Ontology or CDAO (Prosdocimi, et al., <u>http://www.evolutionaryontology.org/cdao</u>). CDAO uses the Description Logics (DL) dialect of the Web Ontology Language (OWL) to formalize key concepts and relations in evolutionary analysis, focusing on phylogenetic trees, character data, Operational Taxonomic Units (OTUs) and evolutionary transitions. CDAO has been subjected to simple tests of representation and reasoning. The next stage in the development of CDAO will involve its use in projects such as interoperability of sequence family data resources and one on developing software support for a data standard for phylogenetics (MIAPA, Minimal Information for a Phylogenetic Analysis).

Support is acknowledged from: U.S. National Science Foundation and ANR (France).

#### 6.2. Taxonomic ontologies: Bridging phylogenetic and taxonomic history

#### Peter E. Midford

University of Kansas

An ontology is a tool for representing our understanding of the fundamental concepts and relations in a domain of knowledge. It can also serve as a bridge to map concepts between domains. Most taxonomic ontologies reflect the tradition dating back to Aristotle where taxonomic groups are, metaphysically speaking, classes whereas organisms are treated metaphysically as individuals. More recently a number of authors including Ghiselin and Hull have argued that taxonomic groups, particularly species, but generally including more inclusive taxa are more appropriately seen as individuals ordered by a part\_of relation rather than member\_of or instance\_of relation. Apart from the metaphysical arguments on the nature of taxonomic groups, treating groups as individuals ordered by part\_of may facilitate appropriate inferences (*e.g.*, reasoning about species ranges) and discourage the sort of inference that leads to problems with character state reversals (*e.g.*, flightless birds). Despite inferential and representational issues, most taxonomic ontologies still arrange taxa in the traditional Aristotelian class hierarchy.

The Teleost Taxonomy Ontology (TTO) was developed for, and is actively used by, the Phenoscape project (<u>http://phenoscape.org/</u>). Phenoscape is developing methods for comparing species that combine genomics and morphology by extending an ontology of zebrafish morphology to a larger group of fish (Ostariophysi). The TTO is used by phenoscape as a source of taxonomic names to construct statements using terms from several ontologies that describe characters observed in the fish taxonomy literature. Although Phenoscape only uses species to construct character statements, names of more inclusive taxa are used to make statements that characters are (or not) homologous. The TTO is generated by parsing and transforming an export of Eschmeyer's Catalog of Fishes, with minor modifications contributed by experts in particular groups within the Ostariophysi. TTO currently contains approximately 36,200 primary terms and 39,000 taxonomic synonyms.

In order to facilitate inclusion of the TTO in the repository of the Open Biomedical Ontologies (OBO) project, the initial structure of the TTO was modeled on an OBO ontology developed to capture the entire NCBI taxonomy though the TTO was never populated from the NCBI taxonomy. In addition to the class hierarchy, TTO and NCBI ontologies include terms for taxonomic ranks (*e.g.*, family) and a special relation type (has\_rank) that links each taxonomic name to its appropriate rank term.

The process of revising the TTO to represent taxa as individuals instead of classes arose from dissatisfaction with the semantics of taxonomic rank terms and the has\_rank relation. This revision required examination of the species concept literature and the decision that TTO should represent species and higher taxa as individuals, effectively viewing them as clades.

Treating the referents of taxon names as clades raises problems with names that were either not phylogenetically based, or were based on phylogenetic trees that are no longer valid. Such names may best be treated as referring to sets of disconnected clades not directly linked by phylogeny. This approach may allow the TTO to serve as a bridge between phylogenetic history (*e.g.*, clades) and taxonomic history (taxon concepts).

Support is acknowledged from: U.S. National Science Foundation and the National Evolutionary Synthesis Center (NESCent)

#### 6.3. Lessons learned from semantic web prototyping in ecology

Cynthia Sims Parr<sup>1</sup>, Joel Sachs<sup>2</sup>, Tim Finin<sup>2</sup> <sup>1</sup> Smithsonian Institution, <sup>2</sup> University of Maryland Baltimore County

We review lessons learned from the SPIRE project – Semantic Prototypes in Research Ecoinformatics (<u>http://spire.umbc.edu</u>) and the ongoing LepTree project (<u>http://www.leptree.net</u>). We include discussion of how ontologies are and could be used by the Encyclopedia of Life (EOL).

Ontologies are a means of storing and representing knowledge and have received increasing attention in recent years. OWL (the Web Ontology Language) and the related RDF (Resource Description Framework) are languages designed to specify the contexts for and logical relationships among terms. Once information in these formats is exposed on the web (the "semantic web"), specialized applications can easily find and integrate it with related information. For example, applications can determine whether "crow" in a web document refers a bird or a Native American tribe. Applications can merge data for "body mass" from different datasets using appropriate unit conversions or methodology adjustments, and they can fuse data from distributed, heterogeneous sources in response to query.

Large complex ontologies are difficult to understand and therefore difficult to re-use. Lightweight ontologies for ecological observations and invasive species enabled SPIRE to quickly prototype tools for ecoblogging and bioblitzes.

One approach to exposing and using semantic web data is to store and reason with data in traditional technologies such as relational databases and then expose it in RDF, rendering it accessible by

"semantic search" crawlers (such as Swoogle) and specialized online query tools (such as TripleShop). RDF is used primarily as an interchange format. SPIRE has taken this approach, and EOL is taking this approach via the TDWG Species Profile Model standard.

Another approach is to work with data in RDF triple stores and then export it as RDF to the web. The LepTree project takes this approach. Rather than a relational data table with hundreds of cells that may be empty for any given row which may be compounded when new fields are added, RDF triples can capture rich information on an ad hoc and extensible basis. This approach is especially worthwhile in distributed environments where schemas are expected to evolve. Widespread adoption will require investment in tools that make it easy to use ontologies and export the instances. EOL is interested in using this approach for atomized descriptive data.

Currently many researchers are exploring a "linked data" vision of the semantic web. In addition to a hyperlinked web of HTML documents, data providers produce a highly interlinked web of data. One uses a Semantic Web Browser such as Disco or Tabulator rather than a standard web browser. As a user follows links, data relevant to each page are gathered for visualization and analysis. Unless data are explicitly linked however, they may not be discoverable. An advantage of this approach is that web sites can simultaneously offer both human-readable and machine-readable information, with the nature of the browser determining which version is served to a user.

Of course the best way forward depends on the desired outcomes. In general, we would recommend development of lightweight ontologies and a strategy that supports both the "semantic search" and "linked data" approaches. Data should be exposed as RDF/OWL and links to other semantic documents should be maximized.

Support is acknowledged from: U.S. National Science Foundation, MacArthur Foundation, Sloan Foundation

#### 6.4. SERONTO: a Socio-Ecological Research and Observation oNTOlogy

D.C. (Bert) van der werf<sup>1</sup>, Mihai Adamescu<sup>2</sup>, Minu Ayromlou<sup>3</sup>, Nicolas Bertrand<sup>4</sup>, Jakub Borovec<sup>5</sup>, Hugues Boussard<sup>6</sup>, Constatin Cazacu<sup>2</sup>, Toon van Daele<sup>7</sup>, Sabina Datcu<sup>2</sup>, Mark Frenzel<sup>8</sup>, Volker Hammen<sup>8</sup>, Helena Karasti<sup>9</sup>, Miklos Kertesz<sup>10</sup>, Pirjo Kuitunen<sup>11</sup>, Mandy Lane<sup>4</sup>, Juraj Lieskovsky<sup>12</sup>, Barbara Magagna<sup>3</sup>, Johannes Peterseil<sup>3</sup>, Sue Rennie<sup>4</sup>, Herbert Schentz<sup>3</sup>, Katharina Schleidt<sup>3</sup>, Liisa Tuominen<sup>13</sup>

<sup>1</sup> ALTERRA Wageningen University and Research Centre, <sup>2</sup> Bucharest University - Department of Systems Ecology, <sup>3</sup> Umweltbundesamt (Federal Environment Agency – Austria) - UBA, <sup>4</sup> Centre for Ecology and Hydrology - CEH, <sup>5</sup> Centre of Academic Sciences - CAS, <sup>6</sup> French National Institute for Agricultural Research - INRA, <sup>7</sup> Research Institute for Nature and Forest - INBO, <sup>8</sup> Helmholtz Centre For Environmental Research - UFZ, <sup>9</sup> University of Oulu, <sup>10</sup> Institute of Ecology and Botany, <sup>11</sup> University of Jyväskylä, <sup>12</sup> Institute of Landscape Ecology Slovak Academy of Sciences - ILE-SAS, <sup>13</sup> Finish Environmental Institute - SYKE

SERONTO is an ontology developed within ALTER-Net, a Long Term Biodiversity, Ecosystem, and Awareness Research Network funded by the European Union. ALTER-Net addresses major biodiversity issues at a European scale. Within this framework SERONTO has been developed to solve the problem of integrating and managing data stored and collected at different locations within the European Union. SERONTO is a product of a group of people with diverse scientific backgrounds.

The ontology is a formal description of the concepts and relationships for the most important aspects of biodiversity data derived from monitoring, experiments and investigations. SERONTO is an ontology that enables seamless presentation of data from different origins in a similar conceptual manner.

With SERONTO, meta-analysis, data mining, and data presentation should be possible across datasets collected for different purposes. SERONTO consists of a core ontology and a separate unit and dimensions ontology. The core ontology is designed to be the basis for domain specific ontologies

(*e.g.*, species, geography, water, vegetation), which extend the concepts and relationships of the core for their specific needs and requirements. The concepts of the core are derived from scientific principles and lean heavily on statistical methodology. Important considerations in designing SERONTO were:

- 1. Repeatability: The ontology should be capable of holding enough meta-data that another person can repeat the experiment or observation at another place and time. It is not obligatory, however, to provide all information for all datasets; for instance, some information may be missing for old datasets.
- 2. Transparency: It must be possible to record and retrieve meta-data describing what actually happened. SERONTO includes concepts of things going wrong and documenting data collection under less than ideal conditions. If data and meta-data are available in this way, it will be clear what assumptions must be made to combine data and correctly interpret analyses.

Important concepts in the SERONTO core are:

- 1. Investigation item the research object or experimental unit;
- 2. Parameters the measurement, classification and treatment of the investigation item;
- 3. Value sets placeholders for time series and other complex data;
- 4. Reference lists nominal values, such as species lists;
- 5. Methods used for each parameter, including units, scale, and dimensions;
- 6. Sampling structure the origin of the research object or population, and the way it was chosen;
- 7. Groupings of objects, such as experimental blocks, on which observer, time or other aspects are assigned or related to;
- 8. Additional information, such as actors (observer, observer groups and institutions), project information, etc., can be attached to several different concepts.

Each subsequent analysis has to make assumptions. The assumptions of any particular analysis can be found in the deviation between how the data were obtained and the requirements of the analytical method.

The presentation will go deeper into the design considerations and the core concepts. Explanations of the concepts, their interrelationships, and their use in subsequent analysis will be given along with examples from different domains.

### Support is acknowledged from: ALTER-Net, a Network of Excellence funded by the EU within its 6th Framework Programme

# 6.5. SONet (Scientific Observations Network) and OBOE (Extensible Observation Ontology): facilitating data interoperability within the environmental and ecological sciences through advanced semantic approaches

Mark Schildhauer<sup>1</sup>, Shawn Bowers<sup>2</sup>, Matt Jones<sup>1</sup>, Josh Madin<sup>3</sup> <sup>1</sup> National Center for Ecological Analysis and Synthesis, U. California Santa Barbara, <sup>2</sup> University of California Davis, <sup>3</sup> Macquarie University, Australia

As the world's biodiversity resources increasingly come under threat, it is imperative to develop a deeper scientific understanding of the processes shaping community and ecosystem structure and function. Insights should emerge from broad-scale, integrative approaches that bring together disparate data from ecology, the biodiversity sciences, and other environmental and life science disciplines. These types of analyses are currently very difficult to conduct, however, because the required data lack interoperability because of their complexity and heterogeneity. While the Web has greatly elevated the potential for rapid discovery and access to information about the natural world, it

remains extremely difficult for scientists to comprehensively and precisely find the data they need, and to rigorously interpret these for synthetic analyses.

We report here on two developments that have direct relevance to the data interoperability challenges confronting biodiversity researchers. OBOE is an "Extensible Observation Ontology" written in the (Web Ontology Language- Description Logics) OWL-DL that provides a core model for generically describing scientific observations and measurements. Researchers on the SEEK project (http://seek.ecoinformatics.org) identified the "scientific observation" as a pervasive and powerful construct for characterizing the rich set of specific objects (entities or specimens) and measurements of their characteristics, that constitute the vast bulk of data in the ecological and biodiversity sciences. OBOE provides a base ontology that can be extended with domain ontologies to incorporate specialized terminologies and constructs that differ among scientific disciplines (*e.g.*, between ecological field science and hydrology). OBOE also provides a flexible mechanism for inter-relating data, such as clarifying the semantic relationships among the columns of a data table. OBOE is currently being used to facilitate smart search and semi-automated integration of heterogeneous data found in the Knowledge Network for Biocomplexity (KNB) repository (http://knb.ecoinformatics.org).

The Scientific Observations Network (SONet) is a new effort, funded by the U.S. National Science Foundation, to facilitate the development of a community-sanctioned, unified data model for observational data. SONet will convene a community of experts, including computer scientists, information managers, and environmental researchers, to develop open-source, standards-based approaches to the modeling of "observational data". SONet provides resources for working groups in several key areas to convene regularly to define and develop specifications and technologies to achieve multidisciplinary data interoperability, based on a semantically extensible framework for describing scientific observations. The SONet effort will identify a common, core model for scientific observations, as well as coordinate the construction of domain-specialized ontologies within the ecological and environmental sciences, that link to the core model. Prototype applications will be built that demonstrate and test the utility of this approach. A final goal for SONet will be to propose mechanisms for establishing a sustainable community to foster these ontology development activities in the future. This 3-year effort is just beginning and will involve the participation of members from the TDWG community. The SONet activity will likely have strong affinities with the Observations Task Group within TDWG's OSR Interest Group.

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

# 6.6. A standard information transfer model scopes the ontologies required for observations

Simon J.D. Cox<sup>1</sup>, Laurent Lefort<sup>2</sup>

<sup>1</sup> CSIRO Exploration and Mining, <sup>2</sup> CSIRO ICT Centre

A standard information model and data transfer encoding for observations and measurements (O&M) has been issued by the Open Geospatial Consortium (OGC), and is on track for publication by the International Standards Organisation (ISO). Based on an analysis by Fowler and Odell, an observation is modeled as an event whose result is an estimate of the value of some property of the feature of interest. The terminology matches the "feature" meta-model used in the geospatial community, but can be applied to non-spatial investigations, so provides general vocabulary for scientific observations. The core model has been validated in geology, geophysics, chemistry, water resources, climate science, environmental monitoring, security and intelligence and taxonomy. An auxiliary model provides for the description of the sample design, in terms of sampling points, curves, sections and specimens. Adoption of a standard terminology supports cross-discipline data interoperability - important for many studies in the natural sciences and for problem solving in the natural environment.

The model provides a structure for observation data. Domain specialization requires the development and management of vocabularies and ontologies to be used as values for elements within a data instance. These include:

- o Domain feature types (*e.g.*, organism, organism occurrence, ecosystem),
- Observable properties (*e.g.*, location, taxon, size, frequency),
- o Observation procedures,
- o Scales and reference systems for observation results (including taxonomies),
- o Sampling-feature relationship types (e.g., part-whole, manifolds, networks and topology), and
- o Specimen preparation procedures.

The first two are tightly linked: the observed property must be associated with the type of the feature of interest. Strict application of the model may require the domain ontology to include previously unrecognized feature types. For example, the notion of "organism occurrence", with observable properties place, time and taxon, is useful in the taxonomic community.

The other aspect of domain specialization is to establish patterns or extensions for sampling features. For example, ecosystem surveys may involve both spatial sampling in the field using quadrats and exsitu observations on specimens. The sampling feature model may be used as-is, but additional domain requirements have led to specialization in at least two applications: Climate Science Markup Language, and the GeoSciML Borehole model.

Ontology-strengthened profiles complement the standardization of the generic data structures and facilitate interoperability within as well as across communities. The challenge here is to coordinate the efforts of multiple communities and factor governance arrangements which can produce a consistent framework rather than a series of independently developed vocabularies. Adopting the O&M skeleton defines the scope of these ontologies. The skeleton can therefore facilitate the coordination of efforts at inter-community level and lead to the construction of more stable ontologies and vocabularies supporting the development of observational data services standards and of multi-disciplinary applications.

### 7. Phylogenetics Standards

#### 7.1. Obstacles and Avenues Facing Phyloinformatic Research

William H Piel Yale Peabody Museum

Phyloinformatics is anticipated to emerge as a new growth area in bioinformatics. Assembling the tree of life will remain at the core of this discipline, but the real expansion will be in the improved extension of phylogenetic knowledge into other areas of the life sciences, such as functional genomics, evo-devo, enzyme-inhibitor co-evolution, and pathogen-host co-evolution. Ultimately, phyloinformatics could become the centerpiece of a synthesis of all biological sciences into a single phylogenetic context that summarizes the complete history of life. The U.S. National Science Foundation sponsored iPlant Collaborative is organizing a Grand Challenge Workshop in fall 2008 to start the process of building a phyloinformatics infrastructure for the plant world. This effort will help identify the obstacles and avenues that need addressing for this to be done at a larger scale. I will highlight some of the informatics needs that we can anticipate as critically important.

Phyloinformatic research requires the following elements: (1) primary archival databases to capture raw phylogenetic data in a digital format; (2) secondary archival databases to assemble and store synthesized trees of life or different parts of the tree of life; (3) standards for exchange formats; (4) standards for universal identifiers and natural ontologies for objects that form the glue among phylogenetic data-sets and among phyloinformatics and other disciplines; (5) analytical databases and software designed to store, mine, and compute over large numbers of and large trees; (6) methods for

extracting generalized phylogenetic knowledge from sets of more specific phylogenetic patterns; (7) methods, services, and APIs for enhancing biological patterns with phylogenetic insight; (8) visualization of phylogenies and biological data patterned with phylogenetic knowledge.

I will focus on obstacles that face TreeBASE as a primary archival database; the problem of generalized topological querying; and the challenges facing large-scale tree visualization.

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

# 7.2. Integrating space, time and form; the challenge of developing standards for geophylogenies.

David Kidd CPB, Imperial College London

Organisms and their genes evolve across space as well as through time, yet the trees and reticulate graphs with which we represent evolutionary relationships do not currently incorporate the spatial component. Geophylogenies are spatially referenced phylogenetic graphs that explicitly represent spatiotemporal evolutionary hypotheses which can be stored, visualized, and analysed within geographical information systems and Earth browsers (*e.g.*, Google Earth and NASA WorldWind). Future data standards for geophylogenies thus require the integration of phylogenetic, geographic and temporal standards, with the added complication of continental drift decoupling the connection between geographical place and coordinate systems.

Simple geophylogenies can be created by attaching spatial coordinates to the nodes of a phylogenetic model and then placing the branches so that they follow the shortest geographical path between nodes. If the spatial coordinates refer to an orthogonal geographic datum (coordinate system) then the shortest path is a straight line. However, if the datum refers to a spheroid then the shortest distance between two points will follow a great-circle and the branches are also required if the shortest path is not the most appropriate path, for example where an impassable barrier separates a pair of sister taxa.

Geophylogenetic nodes may be appropriately represented as a simple single point (x,y,t), for example where graph nodes represent widely dispersed populations. However, such situations are perhaps the exception. Often entities are observed at more than one location or are hypothesised to occupy a certain geographical range, *e.g.*, species and gene ranges. In such cases, nodes may be better represented as a set of point locations, a polyline (*e.g.*, river reaches) or a polygonal geographic region (*e.g.*, species range). As ranges change through time independently of the distribution of phylogenetic nodes, inferred historic ranges may need to be attached to the vertices that separate branch segments as well as nodes.

Modern geographic datums are suitable for shallow-time but as time deepens and continental drift breaks the one-to-one mapping between place and datum coordinates current coordinates will need to be transformed to palaeo-coordinates. The palaeo-model on which such transformations are made, as well as the temporal datum (stratigraphic or radionucleotide) from which dates have been obtained, must be part of any standard.

The geophylogenetic data model is a new development and as such can take advantage of existing and emerging standards, *e.g.*, phylogenetic data in XML (NexML), PhyloXML and the Geography Markup Language (GML). If embraced by the scientific community, geophylogenies have considerable potential to reinvigorate historical biogeography and will underlie any future Map of Life in which the Tree of Life is threaded through earth history.

Support is acknowledged from: National Evolutionary Synthesis Centre and Encyclopedia of Life

#### 7.3. Data standards in phylogenetics: the NeXML project

Rutger Vos University of British Columbia

Phylogenetic data involve operational taxonomic units, graph representations of their evolutionary relationships and comparative data from which the relationships are inferred. To represent phylogenetic data, different types of entities - lists, trees, matrices and their intersections, need to be modeled. These representations are not just sequential or recursive record sets, but rather more complex data models.

The "nexus" flat file format implements such a model using blocks of "taxa", characters and trees. The relationships between entities - operational taxonomic units, character state sequences and nodes in trees within different blocks are defined by labels and implicit orderings of the entities. By virtue of its conciseness and readability, the nexus format has been widely adopted as the standard for many applications, both offline and online.

Unfortunately, the format lacks some formal mechanisms to ensure its continuing robustness in the face of advances in the field of phylogenetics. Principally, it lacks a formal versioning scheme by which new features can be adopted or bugs can be deprecated. It lacks a formal grammar or schema by which its syntax can be validated and it lacks a formal definition of its semantics by which subtle differences in interpretation can be exposed (such as implicit or explicit rootedness of trees, uncertainty versus polymorphism). This has led to mutually incompatible dialects of the standard that hamper the interoperability necessary for more complex analyses and work flows.

Two examples of this problem are the TreeBase project where data submissions need to be curated by hand and are not accessible at a level of granularity that promotes data re-use, and the CIPRES portal where around 20% of submitted data sets contain unrecoverable errors. The success of the standard also has led the phylogenetics community to lag behind in the adoption and standardization around new technologies (*e.g.*, standard markup languages, ontologies, web services) compared to related fields (for example genetics, where data standards are more strict, and where ontologies are becoming highly useful resources).

To address these issues, the Evolutionary Informatics Working Group at the National Evolutionary Synthesis Center has undertaken several projects aimed at improving interoperability at the level of data syntax (NeXML), data semantics (CDAO) and data exchange (PhyloWS). The aim of NeXML is to develop a next generation data standard for phylogenetics; a standard that is based conceptually on the nexus format but with strictly defined syntax and semantics and implemented in XML.

Using the W3C stacks for XML (schema, namespaces, etc.) and ontologies (the Web Ontology Language, OWL and SAWSDL), members of the working group have developed a semantically annotated schema that models Operational Taxonomic Units (OTUs), character state matrices of various types, trees and networks. The schema is versioned and designed for extensibility. Application and web service support is growing, and a developer meeting to implement NeXML support on a variety of online data resources is planned for December 2008.

Support is acknowledged from: National Evolutionary Synthesis Center (NESCent), Durham, NC, USA

#### 7.4. Emerging Data Standards for Phylogenomics Research

Christian M Zmasek Burnham Institute for Medical Research

The term phylogenomics was initially used to describe the application of phylogenetic information for gene function analysis (Eisen, 1998). More recently, the expression has also been employed to describe attempts to reconstruct the evolutionary history of species based on whole genome analyses

(for example, Dunn et al., 2008), as well as various types of studies involving the intersection of genomics and phylogenetics.

A common feature of phylogenomic analyses is the requirement to annotate biological entities such as molecular sequences, phylogenetic trees with data fields such as sequence identifiers, taxonomic data, and, possibly multiple, support values. Simple examples of this are gene trees which have been reconciled with species trees. Such trees have nodes which are at least associated with sequence identifiers as well as taxonomic data. Furthermore, nodes might contain information about whether they represent gene duplications or speciation events. Very similar examples are phylogeographic studies which involve phylogenetic trees with nodes associated with taxonomic as well as geographical information. Currently, there is no widely accepted data standard for phylogenomic information. In general, individual research groups develop their own ad hoc approaches to manage their data. This practice is problematic for data storage, submission, retrieval, exchange, as well as (meta-) analyses.

Recently, several standards have been proposed to meet the data exchange, annotation, and metadata vocabulary needs of phylogenomics research. Following an overview of these, the application of one such proposed standard (phyloXML; <u>http://www.phyloxml.org</u>) in a phylogenomics study relating phylogenetic data, genomic data, as well as protein architectures, is discussed as a practical example and model.

References:

- Eisen J.A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Research, 8, 163-167.
- Dunn C.W. et al. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. Nature, 452, 745-749.

#### 7.5. Emerging standards in phylogenetic nomenclature

Nico Cellinese Florida Museum of Natural History

Recent advances in systematic biology have featured the development of methods for inferring phylogenetic trees. Trees are used to investigate a vast array of biological phenomena and to answer fundamental questions in systematics, evolutionary biology, biogeography, ecology and other domains. The past 20 years have seen the development of phylogenetic nomenclature that extends the tree-thinking approach to biological nomenclature. While traditional nomenclatural approaches apply names to taxa using methods based on taxonomic ranks, phylogenetic nomenclature applies names based on explicit statements of taxon relationships that are validated by phylogenetic trees.

Because a name is explicitly linked to a clade (an ancestor and all of its descendants), it is associated with a taxon concept rather than a rank. Therefore, phylogenetic nomenclature promotes nomenclatural stability and avoids unnecessary name changes when taxa shift between clades. In contrast, a taxon shift in rank results in a consequential cascade of name changes in traditional approaches. Monophyly is widely used to indentify groups that are worth naming, and therefore we are redefining traditional taxon concepts. Reconciling taxon names and concepts represents a challenge from a theoretical and informatics perspective. The PhyloCode represents an application of phylogenetic nomenclature and provides a set of rules for generating phylogenetic definitions. These are node-, branch-, and apomorphy-based definitions according to whether they refer to node-, branch-, and apomorphy-based clades. Every definition requires one or more specifiers, which are reference points that serve to indicate the clade to which the name applies. Specifiers can be species, specimens or apomorphies.

Phylogenetic definitions can be stored as mathematical transformations. Therefore they can be applied unambiguously, and could be easily managed and processed as standards-based statements by computational applications. Although the PhyloCode is not yet officially published, phylogenetic

definitions are currently being published and clade names are increasingly used in the literature. Clade names will have to be stored, managed and potentially linked to sources of trees and matrices, e.g., TreeBASE. A prototype database, RegNum, is currently being developed, and it is due to be completed in 2009.

#### 7.6. Towards a Simple, Standards-Compliant, and Generic Phylogenetic Database Module

Hilmar Lapp<sup>1</sup>, Todd Vision<sup>2</sup> <sup>1</sup> National Evolutionary Synthesis Center (NESCent), <sup>2</sup> Department of Biology, CB 3280, University of North Carolina, Chapel Hill, NC 27599, USA

Despite a diversity of online data repositories in evolutionary biology and highly sophisticated tools for data analysis, most phylogenetic information is not available from any online resource. This lack of information is primarily because individual researchers lack the necessary informatics infrastructure. It is rare to find an online data repository in evolutionary biology that provides its data in a unified exchange format with explicit machine-interpretable semantics and a Application Programming Interface (API) that is predictable. This lack of phyloinformatics standardization contributes to the slow pace of online knowledge accumulation by hindering the ability of individual scientists and the field as a whole to efficiently disseminate, aggregate and mash up phylogenetic data.

A software infrastructure that could overcome some of these obstacles and empower phylogenetic data producers to easily integrate their data into an online network of linked data would involve: i) a simple data model with full support for ontologies and arbitrary metadata so that it can adapt to the needs of individual data producers; ii) a service-oriented architecture with a standard data format, computable semantics and a standard API; and iii) bindings to popular programming languages and toolkits.

While this is an ambitious vision, it can be achieved by a coordinated effort among existing and emerging online data providers and power users. In recent years, a number of the components of this infrastructure have begun to emerge through a process of collaborative software and standards development. The PhyloDB module for the BioSQL schema, which is closely integrated into the Bio\* programming toolkits, together with a first generation of tools to manage trees within PhyloDB, represents a simple yet extensible data model that supports arbitrary metadata with semantics tied to ontologies.

An emerging specification for a standard phylogenetic data provider web-services API called PhyloWS aims to provide a programmable and predictable way to browse, query, and retrieve data from an online phylogenetic database. The recently proposed NeXML format and CDAO ontology can supply the unified exchange format with explicit and computable semantics in which phylogenetic data can be consistently aggregated and mashed up from online data providers. Small widget-like tools that can be embedded into websites or applications such as the tree visualization tool PhyloWidget can serve to demonstrate and exploit the value of rapid online data dissemination that the infrastructure components facilitate.

While standards specifications and ontologies are key and necessary ingredients towards enabling interoperability and large-scale integration, it is the availability of reusable, flexible and communityowned software supporting those standards that ultimately and most effectively promotes their wide adoption. Towards that end, an upcoming hackathon is being planned at NESCent to help facilitate interoperability among evolutionary data providers by focusing on support for standards in the area of syntax, semantics and webservices. Further nurturing the community of data providers, software developers and end users that are interested in phyloinformatics infrastructure can lead to significant contributions to problems in biodiversity management, agriculture, human health and basic research.

Support is acknowledged from: Japanese Database Center for Life Sciences (DBCLS), the Computational Biology Research Center (CBRC), Google Summer of Code(TM), and the U.S. National Evolutionary Synthesis Center (NESCent, U.S. National Science Foundation #EF-0423641)

### 8. Atlas of Living Australia - Australian Biodiversity Informatics

#### 8.1. Introducing Australia's Virtual Herbarium (AVH)

Ben Richardson WA Herbarium, Dept of Env. and Conservation

Australia's Virtual Herbarium (AVH; <u>www.chah.gov.au/avh</u>) is a web application that provides access to about 6 million specimen records from the six State, Territory and Commonwealth herbaria in Australia. The data are retrieved from web service access points at each member institution using TDWG standards, namely the BioCASE and TAPIR protocols and the ABCD specimen data standard. The user is presented with an interface that allows them to download data in a number of formats, including ABCD. The geocodes for the collection localities may also be placed on a map alongside rainfall, temperature and other layers and downloaded as an image for reuse in publications.

The site has a number of use cases:

- At its simplest, it allows clients of Australian herbaria to gain access to specimen data as one dataset.
- In "extended" mode, queries can be made from a large number of available fields for curatorial use as a customisable report tool, as well as to supply custom datasets to clients.

While merging the 6 datasets for use in AVH we discovered a number of data consistency issues. For example, there are a number of ways to specify the rank of the final epithet of a botanical name, *e.g.*, "subspecies", "subsp.", and "ssp.". Each of these values is used in common situations, so we needed to agree on a terminology to increase the reliability of the query results, and the utility of the data in subsequent analyses.

Another issue was how to handle the situation where a taxon name was used differently across the member institutions. The bloodwood tree, Marri, is known in some jurisdictions as *Eucalyptus calophylla* and in others as *Corymbia calophylla*. The interface needed to be able to discover that these were synonyms and display both in the query result. A separate project, the Australian Plant Census, was initiated to answer this problem.

What sets AVH apart?

- Provides an authoritative national view of plant specimen databases. Value has been added to the data as a result of the extra data consistency checking.
- o AVH itself is a TDWG standards compatible web service.
- Simplifies access to herbarium specimen data for both human and software clients of Australia's major herbaria; clients use a single web interface to all Australian records rather than interact with six separate institutions and deal with the data consistency issues themselves.
- The terminology used is understandable by clients of herbaria. Clients don't need to understand ABCD to query the data.
- Provides access to the data through a large number of query fields, many of which are normally only available to database operators in each institution.
- o Catalyst for the adoption of TDWG web service standards in Australian herbaria.
- Catalyst for the discovery of gaps in the management of Australian plant specimen data.

Lessons learned:

- It is more expensive (both in terms of time and money) to develop software that implements ABCD.
- Supplying data in ABCD format doesn't prevent some types of errors in data consistency because the data can't always be validated fully.

• The difficulty of installing BioCASE on Solaris has proved to be a stumbling block to the implementation of web services to support AVH.

Support is acknowledged from: Council of Heads of Australasian Herbaria, Australian State, Territory and Commonwealth herbaria, Andrew Thyne Reid Charitable Trust, The Ian Potter Foundation, John T. Reid Charitable Trust, The Myer Foundation, The R E Ross Trust

## 8.2. CMKb - integrating Australian customary medicinal plant knowledge with the framework of Atlas of Living Australia

Jitendra Gaikwad<sup>1</sup>, Shoba Ranganathan<sup>2</sup>

<sup>1</sup> Macquarie University, <sup>2</sup> Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW 2103, Australia

Australia is among the 34 global biodiversity hotspots, with unique and endemic biodiversity. For more than 40,000 years biodiversity has been part of Australian Aboriginal traditions and customs. Indigenous people have used natural resources such as medicinal plants for treating ailments, diseases and health related problems. The knowledge of medicinal plants possessed by Australian Aboriginal communities is a significant medicinal resource. The documented information on Australian Aboriginal medicinal plant knowledge is scattered in different literature resources, available in heterogeneous data formats, and the knowledge is distributed among various indigenous communities across Australia, due to varied languages. Unfortunately, this knowledge is in danger of being lost due to loss of biodiversity, cultural impact and demise of the custodians of the ancient knowledge.

To document, conserve and disseminate traditional and contemporary medicinal plant knowledge, we have developed a prototype, Customary Medicinal Knowledgebase (CMKb), an integrated multidisciplinary resource (Gaikwad et al, 2008). Currently, CMKb disseminates information collated from published literature pertaining to taxonomy, phytochemistry, biogeography, and biological activities of customary medicinal plant species as well as images of individual species. Each species in CMKb is linked to online resources such as the Integrated Taxonomic Information System (ITIS), NCBI Taxonomy, and Google images, while the bioactive compounds are linked to the PubChem database using PubChem identifier.

CMKb is a species-centric knowledgebase and all CMKb data are linked using scientific name as the fundamental biological descriptor. This fundamental descriptor and the compliance with Darwin Core (DwC) schema provide a facile link to the new national initiative, the Atlas of Living Australia (ALA). The linkage with ALA will enable CMKb to serve as an important data resource on customary medicinal plants in Australia. Similarly analysis tools from ALA such as ecological niche modelling and species prediction will be accessible for species of medicinal interest. Further, the database schema is designed for integration with different informatics infrastructure initiatives, both nationally and globally.

Overall, CMKb serves as a single knowledgebase for holistic plant-derived therapeutics and can be used as an information resource for biodiversity conservation, to lead discovery and conservation of customary medicinal knowledge. From the experience of developing CMKb, we realise that there is an urgent need for the development of standard schemas and protocols dedicated to customary knowledge on medicinal biota. CMKb can be accessed at <a href="http://biolinfo.org/cmkb">http://biolinfo.org/cmkb</a>.

#### Reference

Gaikwad J, Khanna V, Vemulpad S, Jamie J, Kohen J, Ranganathan S: CMKb: a web based prototype for integrating Australian Aboriginal customary medicinal plant knowledge. BMC Bioinformatics 2008. (Accepted)

Support is acknowledged from: Macquarie University, Sydney, NSW, Australia

## 8.3. TAXAMATCH, a "fuzzy" matching algorithm for taxon names, and potential applications in taxonomic databases

#### Tony Rees

#### CSIRO Marine and Atmospheric Research

Misspelled, or variant spelled taxon names present a common problem in taxonomic data systems, leading to recognition (in some systems) of a role for "near" or "fuzzy" match techniques for detection of similar, but not identically spelled taxon names. This approach can be of value in a number of use cases including:

- user queries against a taxonomic database or similar resource where either the input term, or the available target term may be misspelled;
- de-duplication of existing content, for example in disparate sources prior to merging, or within a single resource post such merging;
- handling distributed queries, where the same name may be present in multiple forms (variants or misspellings) in different resources to be searched, which ideally should all be returned to satisfy a relevant user query; and
- spelling error detection and suggested corrections if a suitably complete and authoritative reference database is available.

The present author's activities in this area have included a series of phonetic "near match" algorithms developed over the period 2001-2007, which have been deployed in a range of taxonomic information systems in Australia, Europe, and U.S.A., and more recently (2007-8) a more comprehensive algorithmic solution that has been termed TAXAMATCH. This algorithm is capable of detecting non-phonetic as well as phonetic spelling errors or mismatches. It currently has an execution time of less than 1-2 seconds against a database of over 1.4 million species names on the author's reference system (the Interim Register of Marine and Nonmarine Genera (IRMNG) database of genus names for plants, animals and bacteria, plus many associated species, maintained at CSIRO Marine and Atmospheric Research, Australia). TAXAMATCH provides good performance for both recall (return of most or all "relevant" near matches) and precision (ratio of "relevant" to "non relevant" matches returned) for specific queries using real-world test data.

With TAXAMATCH development essentially complete, attention can be turned to (1) deployment across a range of suitable systems as desired, and (2) its application to situations such as those described above. Such situations include, for example, user searching of available resources *e.g.*, in the Atlas of Living Australia project, provision of web services using IRMNG or other content, and the potential for general use as a "taxonomic spell checker". This last application would depend on availability of a relevant reference list or lists that are authoritative, correct, and complete. While both are to an extent subjectively defined and also a continuously moving target, the present and emerging availability of extensively scrutinized lists such as Catalogue of Life, WoRMS, ZooBank, Index Fungorum, etc. does present some possibilities in this regard, which will be discussed further during this presentation.

#### Further information on TAXAMATCH can be obtained at

<u>http://www.marinebiodiversity.ca/OBI07/sessions/species-names-management-and-tools/oral-rees/</u>, plus an paper being prepared for the biodiversity informatics literature. While the author's present reference implementation is available for online user exploration via the IRMNG search interface at <u>http://www.cmar.csiro.au/datacentre/irmng/</u>.

### 8.4. Making effective use of the Atlas of Living Australia in conservation assessment through advances in spatial analysis and modelling

Simon Ferrier CSIRO Entomology

An important aim of the Atlas of Living Australia (ALA) is to improve access to primary biological specimen and observation data for use in a wide range of environmental planning and management activities. Biodiversity conservation assessment will probably be one of the most prominent of these applications. This presentation explores the contribution that advances in spatial analysis and modelling are likely to make in enhancing the effectiveness with which ALA data are incorporated into real-world conservation assessment and decision-making. Existing approaches to integrating primary biological data with remotely generated environmental layers through species distribution modelling are increasingly being supplemented by techniques for modelling emergent community-level patterns across highly diverse, less well-known, biological groups. Improved understanding and spatial delineation of such patterns can greatly enhance the rigour with which conservation actions are prioritised at multiple scales – continental, regional and local. Modelling of patterns in the distribution of biodiversity is also providing an essential foundation for the ongoing development of dynamic process-based models aimed at forecasting potential impacts of changing climate, land use and management on biodiversity persistence.

## 8.5. Linking phylogenetic trees, taxonomy & geography to map phylogeography using *Biodiverse*

Dan Rosauer, Shawn Laffan University of New South Wales

#### Overview

We describe techniques to link phylogenetic and spatial information to analyse and map phylogenetic diversity. We use the program *Biodiverse* to explore phylogeographic relationships and present results of a new measure, phylogenetic endemism. One challenge for large scale biodiversity mapping projects, taxon name matching, is discussed.

#### Background

Biological diversity is distributed unequally and great effort has been directed to identifying and studying those places where biodiversity is concentrated, such as Global Biodiversity Hotspots. However, biodiversity is also distributed unequally among taxa with some taxa contributing far more to biodiversity than others. Analyses which include both spatial and evolutionary relationships between taxa can better address questions of biogeography and conservation priorities.

Phylogenetic trees, which estimate the evolutionary or genetic relationships within a group of species, are increasingly available electronically, including online at sites such as <u>http://treebase.org</u>, in the standard nexus format. These trees can be integrated with species location data such as are available from GBIF participants.

#### Biodiverse - a spatial tool to analyse species (and other) diversity

*Biodiverse* is a spatial analysis tool developed by Shawn Laffan, Eugene Lubarsky and Dan Rosauer. We present the phylogeny related functions of *Biodiverse* and demonstrate how *Biodiverse* can simplify phylogeography. It can link phylogenies to species location data, and enable users to visually explore the phylogenetic and spatial relationships between taxa, for example identifying where the taxa within a clade (branch and its sub-branches) are found.

#### Phylogenetic endemism

Phylogenetic diversity (PD) uses the lengths of branches on a phylogeny to estimate the diversity represented by the species of an area. Areas where PD is restricted are of importance both for questions of evolution and conservation. Various studies have identified the amount of phylogenetic

diversity or evolutionary history endemic to particular areas. We present a new, general measure, phylogenetic endemism. Phylogenetic endemism estimates the degree to which diversity is restricted to each part of a study area, letting the data define the areas of significance and working consistently at all levels in a phylogeny. This means that it should be applicable to any taxonomic unit, named or not, so long as its spatial distribution and position in the phylogeny can be defined. We present examples of mapped phylogenetic endemism results generated using *Biodiverse*, for various species groups.

Support is acknowledged from: University of New South Wales, CSIRO Centre for Plant Biodiversity Research, Australian Government Department of the Environment, Heritage & the Arts, Australian National University Centre for Macroevolution & Macroecology

## 9. Atlas of Living Australia - Architecture and deliverables

### 9.1. The Atlas of Living Australia – an introduction

Donald Hobern, Nick dos Remedios, Dave Martin, Wolf Wanjura, Lynette Woodburn CSIRO Entomology

Many users have a great need for easier access to well-managed information on Australia's biodiversity. Uses include identification of native and exotic organisms, pest control, conservation, population management, land-use planning and taxonomic research.

The Atlas of Living Australia (ALA, <u>http://www.ala.org.au/</u>) is funded under the Australian government's National Collaborative Research Infrastructure Strategy (NCRIS) programme and involves contributions from a wide range of Australian natural history collections. The project aims to catalogue data resources relating to Australian biodiversity and to assist users in locating and retrieving the data items most relevant to their interests. The ALA will focus particularly on collection data, ecological and observational data sets, images, online literature, diagnostic tools and molecular data, but aims to provide discovery tools which can accommodate the full range of relevant information resources. A primary goal will be to integrate geospatial data on the occurrence of all Australian organisms, including data from specimens held in herbaria and faunal collections and observational data from ecological fieldwork and recorder networks.

The ALA is intended to provide infrastructure services on behalf of Australian collections and researchers and will therefore work with existing Australian biodiversity informatics initiatives to ensure that all data are made accessible using standard data models and interfaces and to provide tools to integrate data from different projects. Important partner projects include:

- The Australian Biological Resources Study (ABRS)
- o Australia's Virtual Herbarium (AVH)
- The Online Zoological Collections of Australian Museums (OZCAM)
- The Australian Plant Pest Database (APPD)
- The Australian Microbial Resources Information Network (AMRiN)
- o The Taxonomy Research and Information Network (TRIN)

The ALA will work with GBIF to develop tools to harvest taxon occurrence data at the national level and to make these data accessible to GBIF in an efficient form. GBIF and the ALA are also exploring the use of Ecological Metadata Language (EML) as a preferred metadata representation for biodiversity data resources. Other data management links are being formed with the Encylopedia of Life (EOL), IdentifyLife, Earthwatch Australia and the Ocean Biogeographic Information System (OBIS).

The project will as far as possible make use of existing standards, tools and best practices already in place within Australian or international biodiversity informatics projects. The ALA will promote the

use of TDWG standards including: TAPIR; LSIDs; Darwin Core, ABCD or the TDWG TaxonOccurrence vocabulary; TCS or the TDWG TaxonConcept and TaxonName vocabularies; and SDD.

A major focus of the project will be around the development of tools for harvesting and enriching metadata for all information resources of relevance to the study of Australian biodiversity.

Support is acknowledged from: National Collaborative Research Infrastructure Strategy

# 9.2. Taxonomy Research & Information Network (TRIN): seeking biodiversity data standards

**E.** Margaret Cawsey<sup>1</sup>, Jim Croft<sup>2</sup>, Garry Jolley-Rogers<sup>2</sup> <sup>1</sup>CSIRO Sustainable Ecosystems, <sup>2</sup> Centre for Plant Biodiversity Research

The Australian Commonwealth Environmental Research Facilities (CERF) supported Taxonomy Research & Information Network (TRIN) is a four year project using contemporary technology and information management standards to enhance the effectiveness, quality and rate of taxonomic research. The network includes a small informatics team to parallel the taxonomy projects, involving a multidisciplinary systems analysis of the practice and processes of taxonomy using comparative analyses, interviews, and structured workshops with taxonomists to understand the complexity and diversity of modern taxonomy.

A major goal of TRIN is to promote the use and development of robust data and information management frameworks and standards in the taxonomy task chain. A major impediment appears to be that those who know and care about biodiversity data standards are not necessarily those who produce the biodiversity data content *e.g.*, the taxonomists themselves. Bridging this gap is a social as well as a technical issue.

The TRIN informatics group is investigating deployment of contemporary internet tools and protocols that may enhance the process and productivity of taxonomy and collaborative taxonomic research.

Early progress in documenting current and potential future models of informatics frameworks reveals consistent patterns suggesting that the costs and benefits of complying with information management standards are unevenly distributed: the imposed overheads and costs tend to fall upon individual research programmes and the benefits accrue at the institutional level. At all levels, compliance depends on the perception of immediate relevance of the data standards and ease of compliance. Institutional and external policies and protocols have the capacity to change these perceptions of relevance. Targeted application of information technology has the capacity to reduce compliance costs. Policy unsupported by technology and technology in the absence of sound institutional polices are both unlikely to be effective.

Perceptions of limited relevance and lack of technological capability at each stage of the taxonomic process limit the retention of multimedia and other digital data. Data lost this way cannot be harvested for use in other unforeseen applications.

The process of taxonomy is ongoing and many products and benefits are not realised until many decades after the initial research. Biodiversity information management and biodiversity data standards need to be applied in similar time frames.

Support is acknowledged from: Australian Government

#### 9.3. Australian Faunal Directory (AFD) and Australian Plant Census (APC): Content, Architecture and Services

Greg Whitbread<sup>1</sup>, Helen Thompson<sup>2</sup>, Matthew Hand<sup>3</sup> <sup>1</sup> Australian National Botanic Gardens, <sup>2</sup> Australian Biological Resources Study, <sup>3</sup> Answerz Pty Ltd

The Australian Faunal Directory (AFD), the Australian Plant Census (APC) and the Australian Plant Name Index (APNI) are the most complete information resources available online for the taxonomy and nomenclature of Australian plants and animals. They are of high quality, authoritative, and widely accepted as the single point of truth for Australian taxa. Together they contribute the nomenclatural and taxonomic core for the Atlas of Living Australia (ALA).

The databases are housed within the Department of the Environment, Water, Heritage and the Arts (DEWHA) and maintained by the Australian Biological Resources Study (ABRS) and the Australian National Botanic Gardens/Centre for Plant Biodiversity Research (ANBG/CPBR). APC operates under the auspices of CHAH (Council of Heads of Australian Herbaria).

In partnership with the ALA we are integrating AFD and APC/APNI at the datum level and building the services required for distribution using standard protocols and forms. The aim is to deliver vertical slices of the data set to client systems in a way that enables maintenance of data quality and update of extracts in place. In the process, we will improve data management, add Life Science Identifiers (LSIDs), increase usability, implement web services over the Taxon Concept Schema (TCS) and support the TAPIR protocol for interoperability. The work is designed to position the application architecture to provide taxonomic and nomenclatural services over the existing dataset and create the platform for a collaborative infrastructure supporting all aspects of taxon profile development.

The foundation for these services is a generic layer supporting XML-RPC, SOAP and REST that can handle all of the protocols we are asked to support – including TAPIR, OAI-PMH, SRW, LSID and XQUERY – and data models specified using the core ontology. The implementation should provide a path to participation in the Semantic Web.

A preliminary iteration uses an XML Java framework to mimic the operation of existing providers implementing TAPIR to SQL translation over hibernate like mappings between provider objects and the underlying database. However TCS, being inherently more complex than other TDWG schemas, is not well suited to SQL processing and the resulting TAPIR solution demands considerable dedicated code for each service and data model supported. The resulting system is not a generic solution.

Our current implementation takes a very generic approach using eXist-DB - an XML database framework - to provide the required services, query processing and database layers. The strong procedural capability of the XQUERY language and the ease with which TAPIR queries can be taken to XPATH expressions reduces TAPIR implementation to near declarative simplicity. Support for our mandatory protocols is pluggable using easy to write translation modules. Code is reduced to a minimum. All required services are bundled, semantic technologies supported and the integrated XSLT capabilities offer benefits for local content delivery. Importantly, direct publication of the production database to XML further simplifies support for multiple schemas through separation of data and provider deployment.

This solution is simple, elegant and portable.

Support is acknowledged from: Department of the Environment, Water, Heritage and the Arts; National Collaborative Research Infrastructure Strategy

#### 9.4. Providing Annotation Services for the Atlas of Living Australia

Ron Chernich, Stephen Crawley, Jane Hunter University of Queensland

The eResearch Group within the School of Information Technology & Electrical Engineering (ITEE) at the University of Queensland (UQ; <u>http://www.itee.uq.edu.au/~eresearch</u>) is collaborating with the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in the development of the annotation services to be provided as part of the Atlas of Living Australia (ALA).

Annotation services enable users to attach additional information or interpretive data to existing resources, datasets, or individual data records. Over the past few years, collaborative tagging and annotation systems that involve communities of users creating and sharing their own metadata have exploded on the Internet. Sites such as Flickr, Del.icio.us, Connotea and LibraryThing are considered exemplary of the Web 2.0 phenomena because they use the Internet to harness collective intelligence. Such systems provide a community-driven, "organic" approach to classifying information and resources on the Web, so that they can be discovered, browsed, and re-used.

The UQ eResearch group has been developing annotation services for text, images, audio, and video files used by cultural and scientific communities. Research has been directed to determining how annotation/tagging services can be used to harness the collective community intelligence to add value to the datasets. This work will be extended to encompass the unique requirements of the biodiversity community underpinning the ALA.

Our objective is to enhance interoperability with other disciplines and systems using Annotea, a W3C specification for an extensible core data model for annotations (Figure 1). This model provides an XML/RDF schema to which annotations must comply. A key activity is to determine the specific extensions required to the underlying Annotea schema in order to support biodiversity data and the needs of the biodiversity community – for example, the annotation type classification as keywords/tags, descriptive, question, comment, correction, ranking, etc.

One of the first objectives is to determine and agree the functional requirements with respect to the annotation services for the ALA project. The aim of this phase will be to address the key use cases specified by the community. Initial feedback and discussions have identified the following broad user needs:

- The Annotation structure is to be defined by an established, extensible metadata schema tailored for ALA specific use cases
- The schema is to be available from a repository in standard formats (RDF & XML Schema)
- Annotations are to be centrally stored using a Globally Unique Identifier (GUID) with multiple indexes, flexible administration, and access control
- Annotations are to be retrieved through a REST-full interface supporting result sets, specific annotations, or harvesting via OAI-PMH with set support
- o Annotations may entered singly or in bulk through a schema validating REST-full interface

A critical aspect of this project will be to develop simple, easy-to-use Web-based user interfaces to support the broad range of use cases including the easy attachment of annotations to images, maps, and particular data records or database rows or columns. Users will be able to search and browse existing annotations and may wish to be notified automatically via RSS subscriptions of any new annotations. Some communities may want to restrict access to their annotations to authenticated users only.

A further requirement will be the ability to harvest annotations stored on remote community annotation servers and aggregate them with the ALA metadata store to enhance discovery. Figure 2 provides a high level view of a proposed system architecture that uses OAI-PMH to periodically harvest community annotations.

Support is acknowledged from: National Collaborative Research Infrastructure Strategy

## **10.** Atlas of Living Australia - Exploring User Scenarios

#### **10.1.** Uses for biodiversity data – the Atlas of Living Australia user needs analysis

Paul Flemons<sup>1</sup>, John Tann<sup>1</sup>, Lynda Kelly<sup>1</sup>, Donald Hobern<sup>2</sup> <sup>1</sup> Australian Museum, <sup>2</sup> Atlas of Living Australia

The Atlas of Living Australia (ALA, <u>http://www.ala.org.au/</u>) has been funded to deliver tools and services to assist users in discovering and using biodiversity data. Ultimately the ALA is intended to serve all groups with an interest in Australian biodiversity. However it is essential that priorities are established to guide the planning and implementation of the project. Such priorities could relate to taxonomic groups of particular significance, to regions of special interest, to specific classes of data, or to information products likely to benefit critical groups of users.

The ALA has commissioned a team from the Australian Museum

(<u>http://www.australianmuseum.net.au</u>) to survey a wide range of Australian users of biodiversity information to help to develop such priorities. Hundreds of individuals, including officers working for commonwealth and state agencies, researchers from universities and museums, private consultants, and others, have responded to an initial questionnaire including the following questions:

- What biodiversity data do you currently use?
- When looking for biodiversity data where do you tend to go?
- Can you give an example of a task, process or application where you use biodiversity data to achieve an outcome that is core to your work or study?
- If you create biodiversity data, tell us briefly about the data you create. Do you make your data available to others? If so, in what form?

Users have been encouraged to identify all information sources used, including literature and expert knowledge as well as online digital resources.

The survey team has also held workshops to facilitate conversations between groups of users to explore their requirements together. A team carrying out a natural resources assessment for several shires in Western Australia has also been documenting their experiences in gathering data for their work.

The survey team is working to document their findings as an online web resource for use by the ALA and other interested projects (for publication before the end of 2008). This report will seek to identify tasks of importance to numbers of users, the sources of information contributing to these tasks, and the key difficulties facing users in performing these tasks today. The ALA will then use these findings to prioritise data sets and tools which can contribute most effectively to addressing the needs of identified user groups. The approach is also expected to allow the ALA to include representative users to help to measure the benefit of the ALA's activities.

Early results have allowed the team to identify several essential tasks which are important to many users. Sessions 11 and 12 of the TDWG 2008 Annual Conference will be devoted to parallel workshops to explore these tasks:

- Distribution analysis how to make the best possible use of all available specimen, observational and other data to determine the likely range for any given species.
- Site assessment how to make the best possible use of all available specimen, observational and other data to report the list of species known and expected to occur at a particular site or in a particular area.

- Identification how to present users with the fullest and most useful resources for identifying organisms from a given taxonomic group and region (including identification of organisms for which no key is available).
- $\circ$  Maintaining web databases how to assist those who maintain specialised web databases (*e.g.*, databases on toxicology, ethnobotany, herpetology or the wildlife of a local region) with locating new data resources and maintaining taxonomic currency.
- Recording amateur observations how to assist and encourage the capture of observational data from amateur naturalists and others, and to manage associated quality issues.
- Including sensitive data how to integrate and manage sensitive data (of any kind) to meet user needs while maintaining safeguards to the satisfaction of data providers. (This will be discussed in relation to the five other tasks.)

In each case, the following questions will be considered:

- What data need to be mobilised?
- How can they be mobilised?
- What data integration services are required?
- What user interfaces and applications would benefit users?
- Are there opportunities for TDWG member projects to work together in this area?

Support is acknowledged from: National Collaborative Research Infrastructure Strategy

### 11. Plenary - Invited Speakers

#### 11.1. The Value of Data and its Management: Confessions of an Applied Environmental Scientist

#### Neil Burrows

Department of Environment and Conservation, Western Australia

We live in a world that is undergoing transition to knowledge-based economies driven by scientific, engineering and technological advances. Science-based knowledge underpins our future as a thriving, cultured and responsible community and is essential for meeting new challenges, such as global warming. Applied scientists contribute to this knowledge and at the core of their work is the collection of data, or observations of the natural environment that form the basis of knowledge about the environment. Data gathered from field research is extremely valuable at many levels, but do we adequately recognize this and how well do we preserve, protect and manage these data? There is a growing need, and in some jurisdictions, a legal obligation, for institutions to develop systems and infrastructure to preserve, protect and facilitate access to data.

# 11.2. Biodiversity Information Standards: Are We Going Wrong, or Just Not Quite Right?

Jim Croft

Australian National Herbarium

Communication in any field requires mutually agreed and understood terminologies, vocabularies, grammars, etc. and a reasonable degree of compliance. This is especially true of the sciences, including biodiversity information management. A suitable lexicon and compliance mechanism did not exist for biodiversity. It had to be invented and TDWG grew to meet this need. The niche was empty and evolution was diverse, rapid and at times chaotic; sometimes the best have survived, sometimes not.

The early drivers of biodiversity information standards were biologists, forced by necessity to deal with information systems, and consequently TDWG's early standards were understandable by any competent biologist. However, shared biodiversity information required more than simple specifications and list of values. Computing professions were engaged to help TDWG, and they revealed a bewildering array of information complexity and implementation options. Complexity has been accelerated by the explosion of the Internet and related technologies. Have TDWG standards now become too complex and too difficult for ordinary biologists to understand and too costly to implement? Could this split the community into those who know and can afford and those who don't and can't? Should TDWG try to 'pick winners' or be more conservative so as to maximise the return on its investment? How do we move forward?

Support is acknowledged from: Australian National Herbarium

### 12. Biodiversity-Related networks and Tools

#### 12.1. speciesLink & openModeller: Network and Tool Enabling Biodiversity Research

Renato De Giovanni, Alexandre Marino, Sidnei de Souza, Dora Ann Lange Canhos, Marinez Ferreira de Siqueira, Vanderlei Perez Canhos

CRIA

The speciesLink network is currently the main source of online species occurrence data in Brazil, serving ~2.8 million records from natural history museums, herbaria and microbial culture collections. The network was launched at the end of 2002 by the Reference Center on Environmental Information (CRIA).

The architecture being used by speciesLink allows providers to either directly join the network by setting up their own Web Service or to mirror their data in one of the regional cache nodes available. This architecture helps to overcome problems associated with poor Internet connectivity and with limitations of hardware and humanware in some collections. Cache nodes enable the participation of most data providers and help to isolate the central database, balancing the process of updating records. Recent statistics show that nearly 70% of the collections that are connected through regional cache nodes have updated their data in the last 6 months, giving a clear indication that the system being used facilitates the process. This strategy allows data providers to use almost any collection management software in any operating system with any kind of Internet connection. Sensitive data can be filtered wherever necessary. The speciesLink website offers a search interface and a set of online tools, helping collections to improve their data and allowing users to monitor the status of the network. New data cleaning tools are constantly being developed and are among the most sophisticated and well-known services provided by the network.

The openModeller project was derived from speciesLink to demonstrate the importance of making primary biodiversity data publicly available. The main goal of the project is to provide computational tools and practical solutions to facilitate analysis, synthesis and visualization of biodiversity data through ecological niche modeling. Besides CRIA, two other Brazilian institutions are participating: Escola Politécnica da USP (Poli) and Instituto Nacional de Pesquisas Espaciais (INPE). One of the main outcomes of the project was a generic framework to generate species' potential distribution models. It is the only software in this category that has a modular architecture, is multi-platform and multi-algorithmic, supports many different data formats, has many different interfaces, is open source and freely available to the scientific community.

These two initiatives recently achieved unprecedented interoperability levels as all speciesLink data are now accessible through a TAPIR/DarwinCore Web Service. The latest version of openModeller also includes a new driver to read species occurrence data from TAPIR/DarwinCore providers. These circumstances facilitate batch processing of modeling experiments that can be used in conservation

planning, impact prediction of climate changes in biodiversity, invasive species prevention, among others.

The presentation will address the evolution of speciesLink and openModeller over the years and the plans for the future.

Support is acknowledged from: FAPESP

#### 12.2. EDIT Platform for Cybertaxonomy – An Overview

#### Andreas Müller, Andreas Kohlbecker, Pepe Ciardelli, Niels Hoffmann, Patricia Kelbert Botanic Garden Botanical Museum Berlin

The EDIT (European Distributed Institute of Taxonomy) Platform for Cybertaxonomy aims at bringing the full taxonomic work process to the internet. It offers a collection of services and tools, and interfaces and transformation tools to be highly interoperable with existing developments. The services and tools developed for the platform can be divided into three groups: community, central and individual components.

Community components are setup once for each taxonomic community, with a community being a group of taxonomists working *e.g.*, on a common set of taxa. Each of these communities uses a community web store based on the CDM (Common Data Model) to store and publish their data and to communicate with the public and other members. The individually configurable EDIT data portal is the preferable tool to publish these data via the web. Users may also implement their own web interfaces by using the web service Application Programming Interface (API) provided by all CDM data stores.

Other facilities available in the community web store are modern communication tools like blogs, fora, mailing lists and other collaborative tools based on the content management system Drupal. Similar to the CDM community store, a store for ATBI (All Taxa Biodiversity Inventory) data may be set up for each community to centrally store observation data and to make these data available to GBIF via web services.

Central Components of the EDIT Platform offer services to be used by all other EDIT and non-EDIT applications. These components are hosted by one EDIT partner for the entire platform. Among these services are geo services to compute and visualize georeferenced data and bibliographic services for searching and accessing literature data that are not necessarily publicly available via the web.

Efficient and reliable security is needed to ensure that read or write access to some data is limited. To make authentication and authorization as easy and secure as possible an IdP (identity provider) using the security framework Shibboleth has been setup for the EDIT Platform. To fulfill some special needs of the platform a couple of new modules have been developed.

To easily find taxonomically relevant specimen and observation data, a web portal has been developed as a further central platform component. The portal searches on the GBIF (Global Biodiversity Information Facility) index by using checklist based thesauruses to receive more complete results. The outcomes can be imported into the user's data store by an easy to use process based on the TDWG standard ABCD.

To complete the set of platform components, a couple of tools are developed for individual work. The EDIT Taxonomic Editor plays an important role with its modern approach to reflect the daily taxonomic work process by intensively using techniques like drag&drop, on the fly parsing, passive warning instead of active asking.

A basis for many of the core platform components is the CDM Library - a Java library that implements all CDM classes and offers an API for local and remote data access as well as for transformation services for all major taxonomic standards from and to CDM.

### 12.3. Data Integration Services for Biodiversity Informatics: An Example from Plant Collections and Morphbank

Greg Riccardi<sup>1</sup>, Boyce Tankersley<sup>2</sup> <sup>1</sup> Florida State University, <sup>2</sup> Chicago Botanic Garden

An interoperable information system has been created that combines the Morphbank image repository, PlantCollections botanical information portal, and several collections management systems. Putting these systems together poses many challenges that are common in biodiversity informatics. Plans are in place to make the system compatible with the TDWG Access Protocol for Information Retrieval (TAPIR). TAPIR compatibility will allow small repositories to be information providers and to allow consumers to feed back information to providers.

TAPIR specifies an XML-based protocol for accessing structured data that may be stored on distributed databases of varied structure. TAPIR combines and extends features of the BioCASe and DiGIR protocols to create a new means of communication between data providers and client applications using the Internet.

PlantCollections (PC) publishes information about public gardens on the Web using the Google Base database and XML technologies. PC relies on Morphbank for storing its images and uses Morphbank services to populate images on its pages. The PC project is an important educational and scientific activity that is integrating beautiful images and garden information with taxonomic information and other scientific content.

Botanical gardens keep extensive records of their specimens, with taxonomy and locality information related to current location and natural habitat. This information is however not readily available outside of the individual gardens. The gardens typically have little informatics infrastructure and no ability to provide server availability 24/7. The infrastructure is therefore often insufficient to allow the gardens to become direct TAPIR providers.

The major challenges for PC were to provide infrastructure to aggregate and publish the databases and the photographic collections of the gardens, and to provide for information update from garden to portal and back. The PC project began with a survey of the database fields in the garden databases, the development of a common metadata schema, the deployment of that schema using Google Base, and the transfer of many images to the Morphbank image and metadata repository.

The typical information flow to publish garden data is:

- o garden information including image metadata is pushed to Google Base;
- o image files are pushed to Morphbank via ftp;
- o specimen metadata are pushed from Google Base to Morphbank via Web services; and
- the PC Portal provides an interface to access metadata and images with active support from Morphbank.

The interaction between the garden database, Google Base, the PC Portal and Morphbank is accomplished using XML documents and Application Programming Interfaces (APIs) that are consistent with TAPIR. The full implementation of updates is not present and is not fully described in TAPIR.

Two major requirements for interoperability for PC are to extend the Darwin Core, Dublin Core, and TAPIR metadata standards to include support for images and annotations, and to create a suitable update specification for biodiversity metadata.

The schema for data exchange will be created from the Dublin Core and Darwin Core metadata schemas and the TAPIR schema. Minor modifications to the existing Morphbank API (Application

Programming Interface) and schema will be necessary for conformance. The existing Morphbank services were designed to be adaptable to new schemas. The creation of these standards will be in collaboration with TDWG, GBIF and the Atlas of Living Australia.

Support is acknowledged from: U.S. National Science Foundation through DBI 0446224, IMLS through National Leadership Grants for Museums

### 12.4. Community Building and Collaborative Georeferencing using GEOLocate

#### Nelson Rios, Henry L. Bart

Over 2.5 billion biological specimens are archived in the world's natural history museums and herbaria. Major efforts are ongoing to computerize and network the data associated with these collections. Sadly, much of these data lack geographic coordinates, so vital to our utilization of this vast information resource in large-scale studies. Development of automated georeferencing tools have greatly facilitated the task of generating geographic coordinates from textual locality descriptions, yet a bottleneck still exists whereby users must manually verify each record georeferenced. We have developed, GEOLocate, a software tool for computer-assisted georeferencing and verification of natural history collections data. Recent advancements include user-defined locality expressions & multi-lingual georeferencing.

Another recent advancement provides a built-in community-based collaborative georeferencing framework where end users can form communities, combine data and distribute workloads of record verification. The Darwin Core format & DiGIR protocol are used in harvesting end user data. Support for additional protocols and models are under development. Integrating techniques for collaborative georeferencing with automated tools has great potential for significantly decreasing the workload of record verification.

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

# 12.5. Halt the loss of biodiversity data! – Towards a new archiving infrastructure for biodiversity information

#### Anton Güntsch, Walter G. Berendsohn

Freie Universität Berlin, Botanic Garden and Botanical Museum Berlin-Dahlem

Over the last decade, the biodiversity informatics community has set up an impressive worldwide infrastructure for instant access to biodiversity data, including collection and observational data, scientific names and taxon-level information. Within a relatively short time, this infrastructure has made available hundreds of millions of raw data records, which are in turn provided to and processed by a growing number of applications; these include the GBIF portal (<u>http://data.gbif.org</u>), BioCASE (http://search.biocase.org), and the German DNA-Bank Network (<u>http://www.dnabank-network.org/</u>).

Although the access and networking aspect of this infrastructure is rapidly developing, the problems of availability and of long-term maintenance of distributed information resources have received relatively little attention. This applies to primary data used for scientific publications, scientific collection data and data for other large projects representing considerable investment. Consequently, huge amounts of important data are still not available in digital form; worse, an increasing number of important data sets are becoming being lost due to the lack of robust data management and archiving strategies.

We therefore propose complementing the existing biodiversity data infrastructure with the development and implementation of a standardized distributed biodiversity data archive, which should offer at least the following services:

- Depository: transformation and upload of source data into domain-specific standard formats such as Access to Biological Collection Data (ABCD), Taxonomic Concept Transfer Schema (TCS), and Structure of Descriptive Data (SDD).
- Metadata: content description including Intellectual Property Rights (IPR) statements and technical parameters supporting the migration process.
- o Migration: transformation of content into contemporary data formats and versions.
- Application repository: storage of software capable of processing biodiversity data in the repository as well as required software environments such as operating systems. For example, entire server environments could be frozen as virtual machine files.
- Access: searching and browsing the archive.

The proposed archiving infrastructure will stop the waste of resources caused by an unstandardized and sometimes frivolous treatment of valuable primary data that have been generated with considerable amounts of public funding. The benefits of this infrastructure investment should easily outweigh its costs. However, we strongly believe that the biodiversity data archive should be implemented in the context of an already accepted existing or emerging joint infrastructure such as LifeWatch - the European facility for biodiversity research - for which a preparatory project has just started (http://www.lifewatch.eu). The triad of efficient generation, reliable availability, and standardized accessibility of primary data will form a tremendously valuable resource for biodiversity science, thereby increasing scientists' ability to tackle the urgent questions brought about by a rapidly changing environment.

## 13. Species Profile Model

### 13.1. A gentle stroll through the Species Profile Model

Robert A. Morris UMASS-Boston

The Species Profile Model (SPM, <u>http://wiki.tdwg.org/twiki/bin/view/SPM/WebHome</u>) is an OWL/RDF ontology based on the TDWG Ontology

(<u>http://wiki.tdwg.org/twiki/bin/view/TAG/TDWGOntology</u>). It was crafted by Roger Hyam with input from an intensive workshop at GBIF in 2007, documented on the SPM wiki. As is the TDWG Ontology, SPM is expressed in RDF, the Resource Description Framework, of the Worldwide Web Consortium, but we do not presume familiarity with RDF, nor will we subject the audience to any of its syntactic arcana.

An SPM object provides various named types of information about a taxon, or more precisely, about a Taxon Concept expressed in the TDWG Ontology controlled vocabulary. The associated information (SPM "InfoItems") comprise a collection of strongly typed attributes drawn currently from one of 37 classes of information about the taxonomic, ecological, and economic properties of the taxon. These include traditional morphological descriptions, information critical to the management of invasive or endangered species, and attributes important for field biology, for ecological science and for molecular studies.

SPM InfoItems, of which an SPM object may contain several, present a simple information model based on seven properties (with possibly repeated instances). One of these is redundant and ought to be deprecated. The others are of three sorts: (1) the information itself; (2) the context (*e.g.*, temporal or geospatial) in which the information is valid); and (3) other taxa that may be associated with the modeled taxon by biological, ecological, or other relations relevant to the InfoItem. The first two types can be given as plain text, by using a controlled vocabulary, or both.

Following Hyam's easy-to-read HTML rendering we will start at <u>http://rs.tdwg.org/ontology/voc/SpeciesProfileModel</u> for a leisurely tour a through the SPM, stopping

along the way for informal opinions on improvements that might be considered. For example, some of the InfoItem context information could profitably be decoupled from specimens and observations. Time permitting, we will also discuss the (mostly easily remedied) impediments to using machine reasoning on SPM objects, and why one would want to do so. A separate presentation by Catapano discusses experience in generating SPM from digitized legacy systematics literature.

Support is acknowledged from: GBIF and the U.S. National Science Foundation

#### 13.2. Plazi: Implementing SPM

Terry Catapano

Plazi is a Swiss-based non-profit organization dedicated to the digitization of legacy scientific literature. Currently Plazi is engaged in a project with GBIF and The Encyclopedia of Life (EOL) on an experimental implementation the Species Profile Model (SPM). Our aim is to use SPM RDF to describe data in ca. 5,000 ant species drawn from taxonomic treatments encoded in the TaxonX XML Schema and served at <u>http://plazi.org</u>. The SPM data in turn will be accessed and processed by EOL agents for incorporation in their resources. Plazi TaxonX documents often document or refer to valid nomenclatural acts and as such may have rich information available for provision to SPM about the Taxon Concepts they define or cite.

SPM documents have two major components: the Taxon Concept which the model documents, and a series of Information Items ("InfoItems)" further elucidating attributes of the described taxon. In general, the InfoItems fall in one of several named classes covering different aspects of scientific interest, and specified using the TDWG OWL/RDF Ontology mechanisms. These range from descriptions of the biology of the taxon, to its ecological impacts and relations, to management and social impacts. Our focus in mainly to represent taxonomic descriptions and the support for them (*e.g.*, the specimens documenting them). InfoItem attribute values can be described either with controlled vocabularies representing, in the case of descriptions, characters and states, or with textbased descriptive phrases extracted from the TaxonX document. To refine the latter into the former is a machine learning research problem actively pursued by Plazi and others in collaboration with Dr. Hong Cui (University of Arizona) and by other groups. In this presentation we only discuss how we generate the textual descriptions. Because SPM is a developing specification with many issues remaining to be encountered and further investigated, we also will discuss our thoughts on some of these such as:

- Syntax of SPM RDF: We will be producing RDF serialized in XML. The XML itself can express the relationships among the data in a variety of ways. What is the optimal expression for the purposes at hand?
- Adequacy and completeness of data: What data should be included? In what form? To what extent should data be explicit in an SPM instance as opposed to being obtained through dereferencing of URIs?
- Provenance data: How can information about the source of both the data (*i.e.*, the publication containing the treatment) and the SPM instance itself be best expressed?
- Validation: What mechanisms can/should be available for validation of the SPM instance? What degree of validity is necessary?
- Profiling: How can all of the above be formally or semi-formally communicated to enable efficient interoperability?

Support is acknowledged from: GBIF/Encyclopedia of Life

### 13.3. Experiences in consuming SPM marked-up data

Patrick Leary Marine Biological Laboratory

In February 2008, the Encyclopedia of Life (EOL) made its official debut with a website (<u>http://www.eol.org/</u>) that featured 30,000 interactive species pages. This was the first major step in this planned 10 year project, which aims to create web pages for the estimated 1.8 million named species on Earth. EOL is made up of five components, each responsible for a different aspect of this grand project. Together they plan to create the foundation for the delivery of species pages, resurrect information held in collections and libraries across the planet, promote EOL as an educational tool, and make use of an index and collection of content to answer lingering biological questions.

The Marine Biological Laboratory in Woods Hole is leading the Biodiversity Informatics component of EOL. The Biodiversity Informatics component creates software which leverages the latest standards and protocols to index and aggregate content from web-based biodiversity projects. Content is identified by using existing query and harvesting protocols as well as by projects volunteering their content directly to EOL. Once identified, content is indexed within a specialized names-based database infrastructure.

EOL is exploring the Species Profile Model (SPM) to help structure biodiversity content in preparation for sharing via query and harvesting protocols. A pilot project was started between EOL, GBIF, and the Plazi.org repository of taxonomic literature to assess the benefits and challenges of using SPM to transfer descriptions of species. The content shared by Plazi is structured using the Species Profile Model and any other components of the TDWG LSID (Life Science Identifier) vocabularies, such as the TaxonConcept and PublicationCitation ontologies. The results of this pilot project will help EOL form recommendations about how to best structure species descriptions for sharing over the web.

Support is acknowledged from: Encyclopedia of Life

### 13.4. The Species Profile Model from an Avian Perspective.

Jeff Gerbracht, Steve Kelling Cornell Lab of Ornithology

The Cornell Lab of Ornithology has several projects that provide life histories of birds via the web. For example, the Birds of North America Online (BNA) (<u>http://bna.birds.cornell.edu</u>) provides the most comprehensive, up-to-date educational resource concerning the life histories of all birds that live and breed in North America. Each of the 750 species accounts includes detailed information on all aspects of the birds' life history, images that describe many of the plumage and behavioral characteristics, many vocalizations, and often video. The full contents of the BNA are available through subscription and are now linked to more than 10 million individual IP addresses.

The content of the BNA is maintained using three hierarchical levels of detail. For example, in each account the highest level (identified as an article) distinguishes categories such as breeding, systematics, and migration. The second level provides more specific information for each of the articles (*e.g.*, for breeding this includes information on phenology, nesting, eggs). The most detailed level provides further specifics (*i.e.*, for breeding phenology this includes information on topics such as pair formation and nest-building). This level of detail provides sufficient information to categorize all aspects of the life histories of birds.

The Cornell Lab of Ornithology has a great interest in providing portions of the BNA content for repurposing by other applications. For example, the Cornell Lab of Ornithology provides access to partial BNA accounts to the All About Birds project. Additionally, other organizations, such as the Encyclopedia of Life (EOL) initiative, have requested access to certain aspects of the BNA for inclusion within their datasets. One method of providing this access would be using the Species Profile Model (SPM), which is designed to enable an entity to express information about a taxon to a

range of different client applications. SPM does not express categories of information in the detail currently used to organize a BNA life history account. Instead SPM info item classes can be viewed as analogous to the highest level of BNA categories.

To best map BNA content into SPM requires an accurate linking of the content at various levels of BNA to the appropriate SPM info item classes. When we attempt to map BNA content to the SPM we find that few categories can be mapped directly and in many cases, the detailed content of a BNA section such as Breeding - Phenology would be mapped to several different sections of the SPM. For this to be efficiently accomplished via automation, a many to one mapping of BNA content to the SPM would need to be created and maintained within the BNA system. Once this mapping is in place, the current BNA life histories as well as any revisions or new species accounts would be available to SPM consumers such as EOL.

## 14. Multimedia in Biodiversity Informatics

# 14.1. Introduction to the session: The relationship of the GBIF Multimedia Resource Task Group to the presentations and to the aims of the TDWG Image Interest Group

Robert A. Morris<sup>1</sup>, Vishwas Chavan<sup>2</sup> <sup>1</sup> UMASS-Boston, <sup>2</sup> Global Biodiversity Information Facility

The TDWG Image Interest Group (IIG, <u>http://www.tdwg.org/activities/img/</u>) is a focal point for multimedia discussion by TDWG members. The GBIF Multimedia Resource Task Group (MRTG, <u>http://wiki.gbif.org/gbif/wikka.php?wakka=MultimediaResourcesTaskGroup</u>) was convened to make social and technical recommendations to GBIF about multimedia resources towards its goal to increase by an order of magnitude the primary biodiversity records offered by its Participants. MRTG has a limited life with a mandate to identify the obstructions to that goal. MRTG's biggest technical focus is on metadata that documents taxon occurrence and supports metadata discovery and exchange, both as to repositories and media.

Several MRTG members urged the development of metadata for descriptions of the content beyond documenting specimens, *e.g.*, to documenting species, phenology, and ecological and other, relationships. . Members also suggested metadata to support applications on whether and how to acquire and render the media and not just exploit the metadata. These issues are beyond MRTG's current mandate.

GBIF together with Encyclopedia of Life is hosting a MRTG sub-group meeting in September 2008 to draft metadata specifications for comment by the public and possible consideration by TDWG.

This presentation will report on the MRTG efforts, GBIFs plans for considering and acting on MRTGs recommendations, and briefly explore how the MRTGs efforts relate to the aims of the TDWG Image Interest Group. Although not selected expressly for that purpose, each of the remaining presentations speaks in one way or another to the intersection of the goals of IIG and MRTG.

Support is acknowledged from: GBIF and the U.S. National Science Foundation

# 14.2. The NBII's Digital Image Library – bringing together diverse images means quality control and standards

Annette L. Olson

National Biological Information Infrastructure

The National Biological Information Infrastructure (NBII) has been building a Digital Image Library (DIL; <u>http://images.nbii.gov</u>) at the request of its partners in order to store images of projects and to make biological images freely available for nonprofit use. The focus is on images that come with

detailed metadata, such as scientific name, date, location, and context, so that the images become resources for research, natural resource management, decision-making, and education. So far, researchers, photographers, and organizations have donated more than 10,000 photographs of species, habitats, environmental issues, and methodology. Another 40,000 images are already in agreement to be added, and many more have been offered either for direct deposit into the library or through web services (*e.g.*, via NatureServe and Discover Life). Also incoming are audio and video files. The NBII DIL will make all of these media available through its website, but it also will redistribute permitted media to Global Biodiversity Information Facility, the Encyclopedia of Life, and others.

With the diversity of media types and sources, however, there are many issues involved in creating one operable media gallery, much less in creating interoperable services. Issues reviewed here: 1) quality control of content – *e.g.*, how much digital editing is allowable; 2) the likelihood of correct species identification; 3) the need to extract or generate metadata automatically; 4) accounting for name changes and broken links; 5) copyright, privacy, and publicity laws; and, especially, 6) determining which standard(s) will be used. Regarding this last issue, there is no one standard suitable for the range of information that comes with biological media, especially for video: one media resource can be the creation of many people, and can cover multiple subjects, dates, locations, devices used, resolutions for display, formats, and more. Many communities consider the development of a recommended model schema, or standard(s), to suit biological media as critical. The use of media, especially as biological records, has been growing extremely fast. Answers to these issues are being sought now.

Support is acknowledged from: National Biological Information Infrastructure

#### 14.3. Creative Commons - Have Your Copyright and Share It Too

Chuck Miller Missouri Botanical Garden

The Creative Commons Corporation was formed in 2001 for "charitable and educational purposes... including, but not limited to, designing methods and technologies that facilitate sharing of scientific, creative, and other intellectual works with the general public". One of the methods created has been Creative Commons licenses that enable copyrighted works to be released by the owner with "some rights reserved", as opposed to a standard copyright which reserves all rights. These licenses may be employed by anyone without fee or an attorney. They all require attribution back to the copyright owner (the licensor) and enable release of different levels of rights.

Creative Commons licenses are becoming more commonly used by the online media community and could be very useful to the natural science community. Many institutions currently utilize all-rights-reserved copyrights to protect their investment in their online information, which can present a barrier to the progress of science and effective use of the information for conservation, sustainability, biodiversity protection, etc. Creative Commons licensing offers a mechanism by which the interests of both the institutions and the community may both be served.

Creative Commons licenses do have some confusing and challenging aspects. The legal language of the actual Creative Commons license contains more restrictions than are revealed on the Creative Commons "Deed" web page (<u>http://creativecommons.org/licenses/by/3.0/us/</u>) which is intended for a general audience. No standard method(s) for permanently attaching the "manner of attribution" to the licensed object has been defined by Creative Commons and as a result the original license and copyright can be lost when the work is transferred. Creative Commons offers no options to restrict obscene or illegal use of a work. And, the applicability of a United States license to the rest of the world is unclear. However, a recent court decision has upheld the enforceability within the United States of a Creative Commons license via court injunction.

TDWG standards should be augmented to include of metadata for copyrighted or public domain status, copyright owner, copyright licensing terms, and manner of attribution required by the license.

This presentation will discuss the Creative Commons licensing approach, the licenses available, issues in using them, and ideas for how they may be brought within TDWG standards.

#### 14.4. Using Fedora Commons to create a persistent archive for digital objects

Phil Cryer

Missouri Botanical Garden

With the amount of digital data and demand for open access to view and reuse such data continually increasing, the adoption of open source digital repository software is critical for long term storage and management of digital objects. Fedora Commons (http://www.fedora-commons.org) is an open source, integrated repository platform that enables the storage, access and management of any kind of digital content. By utilizing this software, the Missouri Botanical Garden has created a stable, persistent archive for Tropicos digital objects, including specimen images, plant photos, and other digital media. Metadata, organized in standard Dublin Core extracted from Tropicos, are stored alongside the digital objects providing search and sharing of data via open web standards such as Representational state transfer (REST) and the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), opening the door for mash-ups and alternative uses. The presentation will cover initial discovery, required hardware and software, and an overview of our experience implementing Fedora Commons. Lessons learned, pros and cons, and other options will also be covered.

Support is acknowledged from: Moore Foundation

# 14.5. Using the JPEG2000 image format for storage and access in biodiversity collections

Chris Freeland

Missouri Botanical Garden

The JPEG2000 (JP2) image file format offers significant benefits to biodiversity collections over traditional JPEG (JPG) images through its superior compression and decoding capabilities. Lossless compression is possible at reductions comparable to the lossy "highest" JPEG compression. Regions of interest can be decoded without decoding the entire image file, and decoding can take place at less than full resolution. These techniques can reduce transmission time in many applications. Finally, JPEG2000 has good facilities for embedding XML annotations in the image file itself, allowing for the storage and distribution of imagery plus metadata within a single file. However, adoption of the JPEG2000 format has been slow because of the relatively limited software solutions available for encoding, decoding, and serving JPEG2000 images. The presentation will cover examples of JPEG2000 in use by museums and libraries, including a discussion of the various software applications currently available, plus potential future uses of the format.

Support is acknowledged from: IMLS, Moore Foundation, Keck Foundation, Mellon Foundation

### 14.6. OntoMorphBankSter: Image-driven Ontology and/or Ontology-driven Image Annotation

Greg Riccardi<sup>1</sup>, Austin Mast<sup>1</sup>, Daniel Miranker<sup>2</sup>, Ferner Cilloniz<sup>3</sup> <sup>1</sup> Florida State University, <sup>2</sup> University of Texas at Austin, <sup>3</sup> University of Texas at Austin

We describe a method of linking images and ontologies realized as a combination of two systems: the Morphbank image and metadata repository (<u>http://www.morphbank.net/</u>) and the Morphster ontology illustration and management system (<u>http://www.morphster.org/</u>). Our formal treatment of images allowed us to create an image-driven ontology editor to enable domain experts to build ontologies as a part of the research process.

Two concepts are linked in this system: image-driven ontology, which couples images with classes and properties to provide a rich definition of the concepts described, and ontology-based image annotation, which tags images with ontology terms to identify the areas of interest within the image that are described by the terms. The two ideas combined support the notion of standard view descriptions—a combination of ontology terms that describe what can be seen in an image. A group may predefine standard views and associate appropriate views with their images. This aggregation of terms into standard views makes it easy for users to give precise descriptions of their images using terms that are understood within the group.

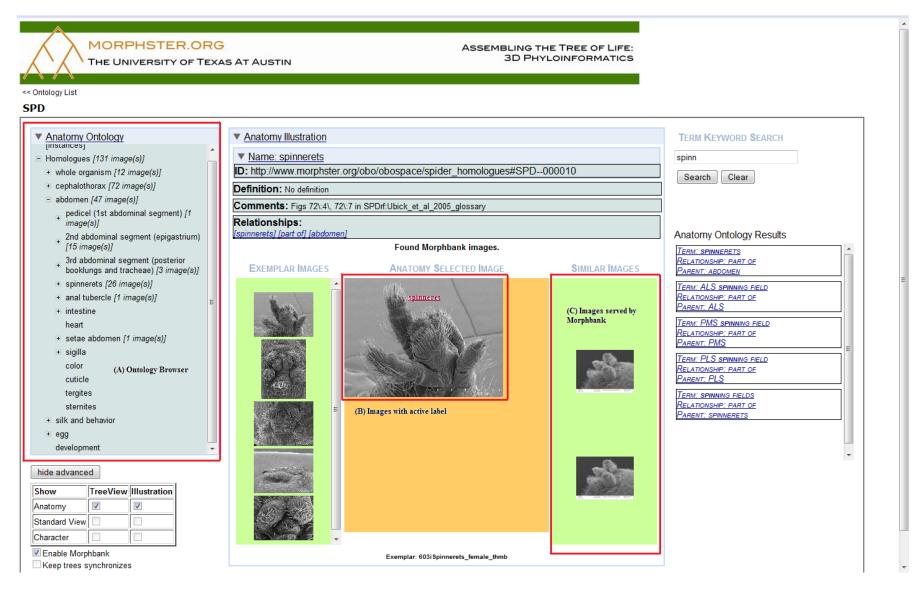
The combined system relies on standard interaction methods. Morphbank has Web services that support query, insert and update. Morphster provides interoperability among a variety of ontology systems. Morphster interacts with Morphbank services to create a dynamic linking between concepts and images. This integration is embodied in OntoBrowser (http://www.morphster.org;8080/OntobrowserV3/Ontobrowser.html).

As images supplant hands-on observation of specimens in systematic research, the use of search methods to retrieve related images is a critical function. An open research question in this environment is how can semantically rich systems (*i.e.*, ontology) exploit those semantics when accessing and retrieving images from legacy databases. In our context, OntoBrowser is a semantically rich environment. Users may choose among ontologies of different organisms. Individual images may be found by browsing by ontology terms or through a search mechanism that leverages the ontology.

Ontology-based image retrieval systems use a basic definition of similarity that involves a graphbased distance measure among the set of ontological terms in the query image and the set of ontological terms associated with each image in the database. This is the same problem as measuring the similarity of the function of two genes per their annotation in the Gene Ontology. A search will be BLAST-like in the sense that the ontology annotations characterize the similarity between images. The results will be ranked based on the information that scientists have associated with the images.

Our ontology activities are informed by our collaboration with the Spider Assembling the Tree of Life (SATOL) project (<u>http://research.amnh.org/atol/files/</u>), Phenoscape (<u>https://www.nescent.org/phenoscape</u>) and the Hymenoptera Ontology (<u>http://ontology.insectmuseum.org</u>).

The attached figure displays information from SATOL. Morphbank contains the standard view manual and the images collected by SATOL. Morphster combines the SATOL ontology and images from the view manual to create an illustrated anatomy atlas. In the figure, the left panel is a browsable tree view of the anatomy ontology that shows the term spinnerets, labeled images documenting the term appear in the center panel, and other images in Morphbank tagged with the ontology term are on the right.



Support is acknowledged from: U.S. National Science Foundation through grants DBI-0446224 and IIS-0531767

## 15. The Global Invasive Species Information Network

# 15.1. Developing a cyberinfrastructure: Experiences from a regional network of plant taxonomists

Zack Murrell, Derick Poindexter, Michael Denslow Appalachian State University

The SouthEast Regional Network of Expertise and Collections (<u>http://www.sernec.org/</u>) is in year three of five years of funding from the U.S. National Science Foundation. The SERNEC community is derived from 215 herbaria distributed in 14 states in the southeastern USA. The curatorial staffs and associated herbarium users come from many career paths, such as systematists, ecologists, land managers, conservation biologists, and educators. This diverse population has agreed to a common goal of making biodiversity information available to various public groups, such as decision-makers, students, teachers, corporations and researchers. The process of mobilizing this community can be compared to herding cats toward a moving target. This cyberinfrastructure includes layers of data infrastructure, computation infrastructure, and communication infrastructure. For this and similar projects to function efficiently and effectively, the most significant component is the human infrastructure.

The SERNEC community is faced with 1) conflicts between existing technologies to develop and enhance databases, 2) disagreements among institutional administrators on the value of providing data to repositories, and 3) changing concepts of how to obtain funding and sustainability of the community effort. Most of the potential data providers find themselves in situations where they are overworked and lack appropriate hardware, information technology support, or administrative recognition of the value of their efforts. Some of these impediments can be overcome through the development of shared resource protocols and community level computational infrastructure. However, major barriers to data availability are 1) the need to expand the community to include computer, social and library scientists, 2) the need to generate a paradigm shift in how productivity is evaluated by administrators, and 3) the need for effective means of communication.

The use of standards and workflow protocols are critical for effective communication within any cyberinfrastructure. To this end, we have worked to inform our community of the ongoing efforts of TDWG and the value of embracing the standards adopted by TDWG. We are encouraging the use of Specify software as our recommended data management tool. Our "Live Plant Imaging Standards Working Group" has developed recommendations for standards of living plant images. We are collaborating with Morphbank to develop workflow protocols for imaging herbarium specimens and making them available via Morphbank and other portals. In addition, the National Biological Information Infrastructure (NBII) and the University of Tennessee's SunSITE are helping us develop portals that utilize information technology tools provided by the Global Biodiversity Information Facility (GBIF) for individual herbaria and thematic and geographical groups of herbaria.

In the 21st century, it is critical that our social networks include the expertise available at the smaller institutions. We must ensure that researchers and data providers at smaller institution with fewer resources are not left behind. We need to help those with greater resources learn how to manage their resources in a sustainable fashion that includes meeting the needs of those at smaller institutions. Only by mobilizing all available expertise can a regional taxon-based cyberinfrastructure generate data of the highest quality.

Support is acknowledged from: National Science Foundation, National Biological Information Infrastructure

#### 15.2. Building a TAPIRLite Toolkit for the Global Invasive Species Information Network (GISIN)

Jim Graham<sup>1</sup>, Catherine Jarnevich<sup>2</sup> <sup>1</sup> Colorado State University, <sup>2</sup> U.S. Geological Survey

The Global Invasive Species Information Network (GISIN) will enable the exchange of data from data managers around the globe. The GISIN Toolkit allows data owners to quickly become a provider to the GISIN. The Toolkit provides a TAPIRLite implementation with Data Models that specifically address the needs of the invasive species data community. To insure the Toolkit is usable by the largest number of users, it was created based on interviews and needs assessment surveys with data providers, data consumers, and data users. Based on provider feedback and limited funding the Toolkit needed to be as small as possible, provided in multiple programming languages, and customizable as needed. The Toolkit has been designed to provide the functionality required by the invasive species community without the overhead of unneeded features. When combined with a transparent, object-oriented approach, this makes the Toolkit much more flexible for providers to adapt to their needs than previous solutions. It also reduces support costs and allows providers to create their own Toolkits if needed. The Toolkit includes an administrative front-end to make configuring a web service quick and easy.

The GISIN Toolkit provides a link between the GISIN TAPIRLite protocol and standard invasive species databases. The Toolkit uses XML-based configuration files to translate the large variety of database schemas from the invasive species community into information compatible with the GISIN data models. The GISIN needs assessment surveys (mentioned above and in previous presentations to TDWG) showed that a TAPIRLite implementation without hierarchical queries and other complex features was preferred by users. The Toolkit is accessed as a URL in a standard web directory structure and should not interfere with any other web services. The skills required to implement the GISIN Toolkit are the ability to copy a folder into an existing web directory structure, the ability to use a browser to edit forms to configure the web services, and knowledge of the design of the database containing the invasive species data to be provided through the GISIN.

The design of the Toolkit is flexible to allow providers to diagnose problems that may arise. The GISIN data models are abstracted into a single file, allowing additions as needed. The Toolkit is installed as a single folder into an existing web-enabled file structure. This is all that is required to install the Toolkit, so an installation program is not currently provided. One may be added in the future if data providers request it. The Toolkit is all OpenSource software and can be obtained at www.niiss.org/gisin.

The first GISIN Toolkit requires any version of PHP 5 and does not require any extensions. The Toolkit comes with a sample database that allows the provider to immediately have a working web service. A web-based administrative interface has been added to make configuring the service to their database quick and easy. The Toolkit includes test tools to ensure the service is operating before registering it with the GISIN network. The Toolkit is currently available in PHP and work is starting on an ASP version.

Support is acknowledged from: National Biological Information Infrastructure, U.S. Geological Survey

### 15.3. Invasive Species Information Management and Exchange in the Americas: IABIN Invasive Information Network (I3N)

#### **Christine Fournier**

National Biological Information Infrastructure

Information on invasive alien species (IAS) from published and unpublished accounts and databases is often scattered in locations and formats not easily accessible. Informatics tools for collecting, organizing and sharing IAS information can help countries to better manage biological invasions. To

overcome this challenge, the Invasives Information Network (I3N) of the Inter-American Biodiversity Information Network (IABIN) created a distributed network of databases for the Americas which includes IAS experts, projects, and datasets. I3N is composed of in-country information providers working to implement common standards for IAS information exchange. Each national node controls its content, though information is documented, validated, and posted in a standard format. The use of a standard taxonomic authority (Integrated Taxonomic Information System) is promoted, a standard vocabulary is available in Spanish, Portuguese, and English, and the Dublin Core Metadata Standard is also used. I3N technical representatives have played an important role in the development and adoption of biodiversity information standards by its parent organization, the Inter-American Biodiversity Information Network, and by the Global Invasive Species Information Network.

Software tools to assist with cataloguing and distributed searching were developed by the National Biological Information Infrastructure (NBII, USA), Instituto Horus (Brazil), and Universidad Nacional del Sur (Argentina). The public can search the records for free from each country's website or from a single Web page. Invasive alien species profiles or fact sheets and occurrence information can be collected using the I3N Database for Invasive Alien Species Template, served on the Internet using the I3N Web Template, and exchanged in the globally recognized Extensible Markup Language (XML). I3N has also developed value added tools including risk assessment and pathways analysis protocols that use the information in the network for preventing new invasions. The I3N facilitates cooperation among countries and provides education and training on the importance of IAS information exchange and the use of I3N tools. Interest in the I3N concept and tools continues has spread beyond the Americas. Representatives in Africa and Asia are now seeking to collaborate with the I3N and learn how to implement similar networks in their own countries.

Support is acknowledged from: Global Environment Facility, Organization of American States, U.S. Geological Survey, U.S. National Biological Information Infrastructure

# 15.4. Ensuring the relevance of invasive species information management tools for policy development and effective management

#### Judy Fisher Fisher Research Pty Ltd

In order to effectively manage invasive species to prevent and/or reduce their impacts on biodiversity, comprehensive up-to-date information, adaptable to wide-ranging conditions is essential. Biological data concerning individual species, their known locations throughout the world, and the species' natural origin and associated climatic conditions are important. However, also critical to ensure the usefulness of databases in the applied management of invasive species is the inclusion of ecological data, as it is often ecological interactions which contribute to the dominance of invasive species. A case study will be described from Perth Western Australia of a transformer invasive species, *Ehrharta calycina*, and its impact within a *Banksia* woodland biodiversity hot spot. The case study demonstrates how the collection of ecological data on the soil seed bank, vegetation composition, fire history and fire responses of the species has led to the successful management and reduction of *Ehrharta calycina*'s dominance. Implications for the incorporation of ecological and management data into invasive species information management tools will be discussed. The links between enhanced invasive species information management tools and informed policy development will be demonstrated.

Support is acknowledged from: School of Plant Biology, University of Western Australia, Botanic Gardens and Parks Authority, Perth

#### 15.5. Data Models to share Invasive Sspecies Information through the Global Invasive **Species Information Network**

Annie Simpson<sup>1</sup>, Jim Graham<sup>2</sup>, Elizabeth Sellers<sup>3</sup>, Michael Browne<sup>4</sup> <sup>1</sup> U.S. National Biological Information Infrastructure, U.S. Geological Survey, <sup>2</sup> Colorado State University, <sup>3</sup> National Biological Information Infrastructure, U.S. Geological Survey, <sup>4</sup> Invasive Species Specialist Group of the Species Survival Commission of the IUCN

The sharing of standardized invasive species information is vital to improve our understanding of biological invasions and to enable the coordination of effective control efforts. The Global Invasive Species Information Network (GISIN) is the working body of the Biodiversity Information Standards - TDWG Invasive Species Interest Group (ISIG), and is a collaboration of like-minded invasive species information managers interested in sharing data globally. In 2008, GISIN held two workshops to establish Data Models for information sharing, with funding from the Group on Earth Observations (GEO), the U.S. National Biological Information Infrastructure (NBII), and the U.S. National Aeronautics and Space Administration (NASA), and with logistical support from The Polistes Foundation (Discover Life).

Although smaller working groups (with less than 20 participants) were more effective at holding standards discussions, workshop participation did include strong representation from both global and regional data providers. Three GISIN Data Models were accepted: Occurrences, SpeciesStatus (including concepts such as Abundance, Distribution, Harmful, RateOfSpread, Persistence, and RegulatoryListing), and SpeciesResourceURLs (for URL lists of one of the following types: species profile, image, video, experts, or references). The GISIN Models use existing concepts of the Darwin Core draft standard, the Dublin Core standard, and several ISO standards. The GISIN Protocol that utilizes these data models is simplified from, and compatible with, TAPIR (TDWG Access Protocol for Information Retrieval). Because the GISIN Protocol is non-relational and contains fewer elements than the full TAPIR protocol, it has been dubbed a TAPIR lite application.

Funding is being sought for a third workshop (GISIN3) to be held in January 2009, to discuss the four remaining Models of the GISIN Protocol: ManagementStatus, ImpactStatus, DispersalStatus, and Citations.

To facilitate communication among its members and with the interested public, the GISIN has several information tools:

- Non-technical online interface, email listserv, and information repository hosted by the U.S. 0 National Biological Information Infrastructure at http://www.gisinetwork.org.
- Technical documents repository and GISIN information system (including directory) hosted 0 by the U.S. National Institute of Invasive Species Science at http://www.niiss.org/gisin.
- Wiki for collaborative discussion and standards development hosted by TDWG at 0 http://wiki.tdwg.org/twiki/bin/view/InvasiveSpecies/WebHome.
- Technical email listserv hosted by the Hawaiian Ecosystems at Risk (HEAR) Project of the 0 U.S. National Biological Information Infrastructure http://www.hear.org/hearlists/gisintech.htm.

Support is acknowledged from: U.S. National Biological Information Infrastructure (U.S. Geological Survey), Group on Earth Observations, U.S. National Aeronautics and Space Administration (NASA), and The Polistes Foundation (Discover Life)

## **16. Bringing Organism Observations into Bioinformatics** Networks

### 16.1. The Significance of Observations to Biodiversity Studies.

Steve Kelling

Cornell Lab of Ornithology

As the types of data included in biodiversity clearinghouses expands outside of the traditional realm of natural history collections opportunities and challenges arise. More data provides a greater opportunity for synthetic analysis across broad spatial and temporal landscapes, but since these data are collected in different ways more care is required in how data are repurposed.

The bird monitoring community classifies species occurrence data by protocol, project design, and data analysis opportunities and this approach may be more broadly applicable. Protocols are an essential part of species occurrence data because they define the context in which the data were collected, and facilitate the combination of observations made by multiple participants in many locations. Protocols directly influence project design and analysis, which can be classified into 3 general categories:

- Directed surveys used when *a priori* knowledge of a given system or biological mechanism already exists. The design attempts to control for known sources of variation, while sampling one or a few well defined variables. As such, directed surveys are the form of observational data collection that closest resembles experimental studies.
- Broad-scale surveys generate probabilistic estimates of species occurrence. They do not provide direct evidence, but allow inferences for the causes of species occurrence. Broadscale surveys gather tens of millions of observations annually and provide the bulk of nonspecimen observational data available.
- Biological collections are zoological, botanical, and paleontological specimens in museums, living collections in botanical or zoological gardens, or microbial strain and tissue collections. They are the foundation for taxonomic and historic occurrence of species. While most use of specimen collections has been for taxon-oriented research, they have been used for predictive modeling of species occurrence.

Each category of species occurrence data has its own issues. Directed surveys are expensive and conducted on small spatial and temporal scales. Broad scale surveys often gather data opportunistically, rely on volunteer participation, and use protocols that are less stringent. Biological collections often provide presence-only data with significant and often undocumented sampling biases.

While important inroads have been made in organizing broad scale surveys and biological collection data via standardized schemas, few directed surveys have been added to data clearinghouses. They represent a huge untapped biodiversity data resource, but integrating them with broad scale surveys or biological collections will require increased information provided at the individual record level. For example, the explanatory factors needed to understand the processes (*i.e.*, protocol design and sampling plan) used to collect the data must be included. Consequently, individual data records will need to include both biologically-relevant factors that affect organisms' distribution and abundance as well as information on the factors that impact the data collection process. With sufficient additional record-level data, potential sources of bias may be investigated and be accounted, thus increasing the significance of the analysis or visualization process.

# 16.2. Towards Global Observation Networks: GBIF's plans for mobilizing primary biodiversity data

Vishwas Chavan, Éamonn Ó Tuama Global Biodiversity Information Facility

Increased digitization, mobilization and access to "fit for use" primary biodiversity data through the GBIF network will enable its participants and others to effectively manage their biodiversity resources. During the past 6 years the GBIF network has been able to mobile over 145 million primary biodiversity records. What is now required in addition to this core on-going mobilization effort, is an increase in discovery, digitization and access to various types of primary biodiversity data. Therefore it is critical to also focus on supporting the development of network capacity to discover and mobilise a greatly expanded array of primary biodiversity data types complementing the existing specimen based records. To explore the feasibility of mobilizing non-specimen based data records, GBIF convened two tasks, namely, the Observational Data Task Group (ODTG) and Multimedia Resources Task Group (MRTG).

The ODTG was tasked with addressing issues around the mobilization of observational data for the GBIF network, as inclusion of observation data types will contribute significantly to achieving the target of data mobilisation as outlined in GBIF's 2009-2010 Work Program. During the last 8 months, the ODTG has brainstormed on several aspects including (a) what are biodiversity observations, (b) how should GBIF work to integrate observational data in the short and long term, (c) how should GBIF expand descriptive metadata for observational data sets, and (d) how can GBIF increase participation of observational data providers.

Our presentation will describe the outcomes of the ODTG that will help build the global network of observational data providers by catalyzing both technical and social mechanisms for mobilizing increased volume, depth and density of observational data.

#### 16.3. Directions in observational data organization: from schemas to ontologies

Matthew B. Jones<sup>1</sup>, Chad Berkley<sup>2</sup>, Shawn Bowers<sup>3</sup>, Josh Madin<sup>4</sup>, Mark Schildhauer<sup>2</sup> <sup>1</sup> NCEAS, UC Santa Barbara, <sup>2</sup> NCEAS, U. California Santa Barbara, <sup>3</sup> U. California Davis, <sup>4</sup> Macquarie University

Ecological studies generate data on the distribution and abundance of organisms and the processes and mechanisms that lead to the maintenance of biodiversity. Studies range from uniform monitoring of the spatiotemporal distribution of taxa to small and large scale experimental manipulations of the environment in order to elucidate ecological processes. These studies use methods and sampling designs that vary in both significant and subtle ways, making synthetic studies that utilize these data labor intensive. Mechanisms for capturing these data have traditionally employed structured metadata standards such as the Ecological Metadata Language (EML) and the Biological Data Profile (BDP). These metadata systems describe the data set and methodologies using natural language in order that researchers can properly interpret and use the data. EML also provides mechanisms to encode key aspects of the metadata, such as the physical and logical structure of the data set and the units of measurement, in a structured format that can be computer processed without human interpretation. This allows semi-automated loading and parsing of EML-described data sets, which can substantially streamline data synthesis.

Nevertheless, significant methodological information about the experimental design, sampling methods, and contextual and semantic relationships among data observations are only represented in natural language in these traditional metadata standards. Multiple projects have been developing ontologies and semantic annotation systems to represent scientific observations so that these important semantic subtleties can be available to computing systems for automated data processing. Several of these ontologies, such as the Extensible Observation Ontology (OBOE), explicitly model observations and the context in which these observations occur. They allow for accurate models of complex experimental designs, and provide a semantically robust means of describing ecological data so that observations in different studies conducted using varying methods can be integrated in cross-

cutting analyses. While the promise of these semantic observation ontologies is great, there is significant work remaining to standardize the ontological frameworks and to develop software tools to make annotating data and using ontologies approachable for research scientists.

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

# 16.4. Dealing with Sensitive Data: managing access, issues of data quality, and data ownership

#### Arthur David Chapman

A survey conducted for the Global Biodiversity Information Facility identified that as many as 64% of data providers surveyed were generalising sensitive biological data in some way. How those data were generalized varied greatly and there was little documentation to inform users that the data had been generalised and if so, how. Following on from this an attempt has been made to develop some standard methodologies for data providers to generalise their data, and to standardise on documentation. Special emphasis has been applied to spatial attributes, but non-spatial attributes have also been covered. The results of the study have led to a Guide to Best Practices for Generalising Sensitive Species-Occurrence Data co-authored by myself and Oliver Grafton. Iwill discuss this Guide; examine potential scenarios as to how biodiversity data repositories can adequately maintain accurate information on sensitive species and data while at the same time regulating the level of generalisation for direct access. The paper will also examine methods for proper attribution of data owners and custodians, look at 'where to form here', and seek to engender discussion on how best to encourage date providers to adopt the Guide's recommendations.

#### 16.5. Domain-centric observation networks: Experiences gained from the Avian Knowledge Network

Denis Lepage Bird Studies Canada

The Avian Knowledge Network (AKN) is a network of North American institutions dedicated to the ecological study of bird populations. AKN nodes have contributed so far nearly 50 million observation records, mainly generated through broad-scale surveys. The goals of the AKN include: facilitating access to bird data in a standardized format, create a long-term archive to reduce data loss and develop visualizations and analysis tools that will facilitate understanding the patterns and dynamics of bird populations. The data federation standard used by AKN is an extension of the DarwinCore schema called the Bird Monitoring Data Exchange (BMDE). In addition to the DarwinCore fields that describe the taxonomy, the location, the time, etc. and allow describing presence only, the BMDE also includes fields to describe characteristics of the sampling events during which the observations were made, such as unique sampling identifiers and information on protocol and effort. Those additional fields allow making inferences about the data, such as the absence of organisms or the detection probabilities, and allow for example examining spatial and temporal changes in relative abundance. Additional extensions of BMDE are also created for more specialized types of datasets and specialized types of questions (*e.g.*, banding recovery data based on bird banding or productivity based on nest records).

The AKN is being developed as a collaborative environment where nodes play a central role. When acting as data providers, either as primary data custodian or on behalf of other partners, nodes have a primary role in ensuring data quality and compliance to standards. Nodes must also ensure that datasets are properly documented through metadata and that data access issues are properly addressed. Nodes are also acting as data consumers, and are engaged in developing interoperable visualization and analysis tools that can address the needs of various user communities, including the general public, researchers and students, conservation biologists and decision makers.

In my presentation, I will provide a brief overview of the architecture of the AKN and the roles played by the nodes; I will describe how schema extensions are built to allow addressing specific questions related to bird population dynamics; I will explain some of the ways by which the AKN is addressing various sensitivities related to data sharing; finally, I will provide some examples of how the AKN concept facilitates the development of interoperable tools by the nodes, such as web services for data analyses and visualizations (maps and summaries), as well as toolkits such as an R package for data analysts that is based on AKN standards.

## 17. Wild Ideas!!

### 17.1. A Hadoop-based Prototype for the Filtered-Push project

Zhimin Wang<sup>1</sup>, Hui Dong<sup>2</sup>, Maureen Kelly<sup>3</sup>, James A. Macklin<sup>3</sup>, Paul J. Morris<sup>4</sup>, Robert A. Morris<sup>2</sup>

<sup>1</sup> University of Massachusetts Boston, <sup>2</sup> University of Massachusetts, Boston, <sup>3</sup> Harvard University Herbaria, <sup>4</sup> Harvard University Herbaria and Museum of Comparative Zoology

The Filtered-Push project aims to establish a cross-institutional infrastructure to help taxonomists share and improve digitized collection data via the exchange and management of record annotations. The project addresses three major challenges: The identification and annotation of specimen records in multiple collections that arose from a single collection event; the quality control of new annotations; and more generally the dissemination of annotations specimen records, whether or not representing duplicate specimens.

To address those concerns, first we decompose this system into five modules: The client Application Programming Interface (API); network communication module; schema translation module; network node adapter; and storage module. All these modules are glued together through well defined interfaces, which also give us the flexibility to change the underlying technique for each module. In a prototype, we are adopting the Apache Hadoop map-reduce framework (<u>http://hadoop.apache.org</u>) for the communication and storage modules in our system for the following reasons: first the network discovery of duplicates fits the map-reduce model well, in that one can understand the process as combining (reduce) results from local searches (map). Hadoop makes it easy for us to distribute programs and computation tasks across the net. The Hadoop HBase distributed database provides high availability by its robust, transparent file replication architecture. The column-oriented database structure is particularly suited to a global annotation store, from which local participants can accept or reject annotations based on local policies. This global repository also helps to retain knowledge that is independent of local nodes such as pending annotations.

The current underlying data model is a duplicates-oriented global view of specimen data of participants. Through this view, local changes can be provided globally and new global annotations can be applied to local copies. To facilitate the data exchange and data sharing, we use the Darwin Core as the common vocabulary, which is also extended to address some application specific problems, such as managing Globally Unique IDs of duplicates. The architecture imposes a strict separation between the message passing network and the computation models required to filter and respond to messages about annotations.

We have begun building a Hadoop-based prototype annotation sharing network. It holds promise as a platform for more complex research-oriented computations related to collection data, such as clustering of potential duplicates, identification of outliers for quality control. A separate poster by P.J. Morris et al. describes the principal use cases and the messages required to describe them in the network.

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

#### 17.2. Data Exchange Standards – The Case for Being Stupidly Simple

Chuck Miller, Missouri Botanical Garden

Why are Biodiversity Information Systems (TDWG) data exchange standards so complex? The schema based standards like Structured Descriptive Data (SDD), Access to Biological Collections Data (ABCD), and the Taxonomic Concept transfer Schema (TCS) can involve hundreds of nodes, making mapping to data sources a difficult task. Extensions to deeply layered ontologies and Resource Description Framework (RDF) triples are adding even more complexity. Although there are sound reasons for full analysis and detailed, complete accounting of all possible data values and structures, there are counterbalancing reasons to also consider simpler structures. The majority of institutions with biodiversity data to share do not have the skilled informatics staff needed to implement complex data interfacing techniques. Furthermore, they may lack the knowledge to understand even the descriptions of the interfacing techniques. TDWG runs the risk of producing products that are so complex they further isolate rather than capture a large segment of the biodiversity data available.

Can we relearn the lesson of Darwin Core (DwC)? Despite its limitations, multiple versions in use, difficulties as an XML Schema, and, some say, lack of scalability, it is ostensibly the single most used biodiversity data exchange standard in the world, exchanging over 140 million records from 3,000 datasets within the GBIF network alone. Why? It's simple and linear. Although its simplicity necessarily means information is missing, it still provides the essential facts about specimens and occurrences. And being simple and linear it can be implemented with basic informatics knowledge. The lesson from Darwin Core is that we need simple and linear exchange standards for all of the categories of data: concepts, names, descriptions, and more. Not instead of TCS, SDD or ABCD but in addition to and integrated with them. This presentation will suggest some approaches to a simpler side of TDWG.

#### 17.3. 3D and beyond. What graphics cards can do for you

Bart Meganck, Patricia Mergen Royal Museum for Central Africa

Biological mapping and visualisation tools like the ones developed within Biodiversity Information Standards (TDWG) and GBIF (Global Biodiversity Information Facility) context are presented to a public accustomed to gimmicks and eye candy. Compare any map viewer with the rich fly-through world of Google Earth for example, and you'll notice we're missing a dimension. There's something unnatural about our current 2-D presentation of biological datasets, about not being able to grab it and look at it from all possible directions - be it a bunch of specimen gathering spots, or a river trajectory, or measuring points on an ape skull. This is more than just eye candy: this is crucial for a good interpretation. And why are we always registering heights for our specimens, if not for using them in 3-D? A bold Star Trek style prediction (this is the "wild ideas" section after all): "We need to conquer the 3rd dimension, or we condemn ourselves to oblivion".

3D- in visualisation and 3-D in data space as well: Spatial dataset calculations are an order of magnitude more complex than our current "distances" and "nearest neighbour" questions.

How to get the 3-D power for this brave new world?

A tantalising new approach is GPGPU, or General Processing GPU (Graphical Processing Unit) processing. In normal language: Look for the processing power where it can be found. No other line of computer hardware is under such rapid development as the GPU (Graphical Processing Unit) on the graphics card, with throngs of gamers crying out for ever faster, more spectacular games. GPUs come relatively cheap as well, compared to your normal processor (CPU or Central Processing Unit).

And you can string together more than one graphics card nowadays, to get one 3D super-number cruncher. An "idiot savant" supercruncher that is, because GPU's traditionally are built for a single

task: Rendering as many frames per seconds as possible on your screen. They excel at shading, rendering 3D scenes and ray tracing (all things 3D), but don't ask them to calculate a square root. Clever software engineers have been working at just that: Tricking 3D graphics cards into "multi-purpose" processing, thus extracting an enormous processing power "on the cheap".

Different frameworks exist for programming general science problems into the single-minded graphics card. But rather than losing ourselves in the technical tidbits, we'd like to present some examples of what people are doing with this technology:

- o 3-D scenery rendering for movie special effects,
- o 3-D physics animations (effects of gravity and wind on vegetation and constructions ),
- o 3-D particle behaviour and collision detection (fluids, clouds, sand grains),
- Realistic texture rendering (*e.g.*, skin texture) for movies,
- o Software virus signature matching,
- Encryption and decryption,
- Random-number generation, and
- Earth subsurface imaging from seismic data.

We'll also address how GPGPUs could possibly serve in a biological/biodiversity context - as "wild idea" and food for thought:

- 3-D river trajectory calculations for extracting biological parameters (*e.g.*, speed of flow, total length of rivers),
- o Random-number generating for Monte Carlo algorithms for biological statistics,
- o 3-D calculations of (nearest neighbour) patterns within a specimen gathering dataset,
- o 3-D correlation search between parameter layers (e.g., height) and specimen data points,
- o Biological virus signature matching / DNA pattern matching (e.g., in biogeography),
- o Fast 3-D rendering for online GIS (Geographical Information Systems) applications, and
- o 3-D skull and skeleton reconstructions and calculations.

Many of the examples are drawn from GPU Gems 3 / edited by Hubert Nguyen Addison Wesley/Pearson education Copyright 2008 NVIDIA Corporation

### 17.4. SpeciesIndex: A Practical Alternative to Fantasy Mashups?

#### Roger Hyam

Royal Botanic Garden Edinburgh

Generating consensus descriptions of species by mashing together data gleaned from different sources has been proposed by various projects but will never be practical. Different accounts not only come in different formats they are also authored for different audiences from different perspectives. An afternoon in the library with a photocopier and a pair of scissors will demonstrate the impossibility of creating non-trivial descriptions by cutting and pasting from original sources even for a human.

This wild idea is that we shouldn't bother doing mashups but instead enhance the indexing and federated search services such as iSpecies by answering the simple question "Where are all the species pages?".

The proposed way to do this is to exploit the well established SiteMaps format to allow authors of taxonomic accounts to provide an index file to their description pages.

SiteMaps is a protocol that allows webmasters to inform search engines of the URLs they would like to have indexed. Practically the protocol consists of an XML or plain text file that contains a list of URLs and metadata for a particular website. It is a very simple protocol, easy to implement and supported by a wide range of search engines including Google, MSN, Yahoo! and Ask.com. There

are only three metadata elements in the protocol. They consist of the last modification date, change frequency and priority. Importantly SiteMaps is extensible by addition of XML elements in other namespaces.

As a first level of implementation data suppliers could be asked to generate a SiteMap file as per the protocol that only includes their species pages. There are tutorials and validation tools already available from Google to support this process. They should then submit the URL of the file to a Species Index Registry(SIR) along with a simple description of the taxonomic and geographic scope of their data. SIR consists of a human with an email account and a manually edited web page. It should be relatively simple to write an indexer that uses the SiteMaps listed in SIR to index just species page of interest to a particular project. The cost and risk of implementation of this strategy is very small yet it would enable a great deal of innovation going forward.

A second level of implementation would involve extending the SiteMaps file format to include metadata about what each species page included. This could involve reusing the TDWG ontology namespaces. Including more metadata in the SiteMaps file would enable the generation of a more complex registry and more intelligent indexers. It is highly likely however that discussions about metadata extensions will bog down the first level implementation and so prevent anything practical happening at all. Perhaps that is why this is a wild idea?

## **18.** Computer Demonstration

### 18.1. Introducing the EDIT Desktop Taxonomic Editor

#### Pepe Ciardelli, Andreas Müller, Anton Guentsch, Walter Berendsohn Botanic Garden Botanical Museum Berlin-Dahlem

Taxonomy is a fiendishly complex problem domain. We attempt to write software addressing every conceivable aspect of this domain. In doing so, we often overlook the importance of making software useable. The EDIT Desktop Taxonomic Editor attempts to make use of hard-won lessons to strike a harmonious balance between painstaking attention to taxonomic detail and keeping the application intuitive, error-tolerant, and fun to use. This presentation will go into the philosophy behind the design of the Editor, and walk the audience through some of its highlights.

The core of the European Distributed Institute of Taxonomy (EDIT) Platform for Cybertaxonomy is the Common Data Model (CDM) for modeling taxonomic information. The EDIT Desktop Taxonomic Editor is the main tool for editing CDM data, from fine-tuning an imported checklist to compiling a complete dataset from scratch.

We start with the proposition that no taxonomist wants to see yet another form-based web application. They would prefer tools more attuned to their traditional workflow: Taxon-centered, list-based, and heavily oriented towards cutting and pasting blocks of text. If our design is successful, the application should correspond to this workflow.

Users should be kept aware that the goal of the Editor (and of EDIT) is to get their data into a structured format which can then be used to share results with other taxonomists worldwide. Our solution is to present the data side by side in two different layouts. To the left is a free text area, where the taxonomist enters data by pasting it, by dragging it from other application windows, or by entering it manually. To the right is a so-called "property sheet" view of the data element which currently has a focus that reflects the structure with which it is saved to the CDM. For instance, if a scientific name is currently being edited, the property sheet will show its rank, genus, species epithet, the components of its nomenclatural reference and its nomenclatural status.

The interface between the free text area and the property sheet is an on-the-fly parser. This parser atomizes the data as it is entered in the free text area, and displays the results in the property sheet. Our goal is for the parser to correctly atomize the data in at least 95% of all cases; fine-tuning will

then be done by the taxonomist in individual property sheet fields. The other main function of the parser is to point out data elements which it recognizes as syntactically incorrect. Offending passages are underlined in red, and the application will offer suggestions for correcting the mistake. The taxonomist can at any time produce a report with all syntactical violations in the current treatment.

Using autocomplete menus, the user is gently prodded toward reusing data which already exist in the CDM. For instance, if the parser recognizes that the user is currently entering author team data, a dropdown list appears consisting of author teams in the CDM which match the characters the user has already typed.

In general, we will replace enforced integrity ("you can't do this") and confirmation messages ("are you sure you want to do this?") with a less intrusive model of warning messages. We will show the user the implications of what has been done, and then make it easy to undo the step just taken. Warnings will often come with suggestions - we often have a fairly good idea of what the taxonomist is trying to do, but at the same time the Editor should give the taxonomist the benefit of the doubt and not force him to do anything against his will.

We have decided that as a matter of principle, design should focus on identifying and simplifying default cases; we must avoid making workflows complex to accommodate cases that occur infrequently. In practice, this means automating as much as possible and enforcing data integrity as little as possible - in short, trusting the taxonomist to do what is right.

http://dev.e-taxonomy.eu/trac/wiki/TaxonomicEditor

Support is acknowledged from: European Distributed Institute of Taxonomy (EDIT)

### 18.2. SilverCollection: A Standardized Web Portal for Biological Collections

Michael Giddens SilverBiology

SilverCollection is an interactive web portal for public viewing of biological collections. In contrast with other general web portal creations where a form is submitted for searching a collection, or hyperlinks are followed to browse taxa, this software combines the TDWG standards for data interoperability with real-time designs. In these designs, dynamically-generated web pages that have the behavior and the look-and-feel of a common desktop application replace traditional forms and hyperlinks. The main objective of this software is to provide web access to functions commonly found in collection management software, such as browsing, searching, and the generation of distribution maps and checklists from collection data. Additional features include the ability to download and send by email any set of data.

Independent of the collection data are images tagged with searchable morphological character metadata. These images would be provided by the institution housing the collection. Once metadata are bound to the image, these fields can be used by any web user to identify taxa, similar to using an interactive key.

SilverCollection does not replace collection management software; instead, it serves its own Darwin Core v1.4 compliant data from a MySQL database. A harvester is used to synchronize the authoritative database to the SilverCollection's database.

The goal of this software is to form a network of collections that use the same common web portal interface with common tools that work easily with all the biological collections.

The software is setup to work with a TapirLink provider, TapirLink's LSID resolver, and other third party tools, *e.g.*, Google maps, BioGeomancer, ICUN Red List of threatened species, and the International Water Management Institute's online climate summary service.

A video demonstration is available at <u>http://collection.silverbiology.com/videos/flash/1</u>. Further information about this project and its progress, updated periodically, may be found at <u>http://collection.silverbiology.com</u>.

### 18.3. Mandala 8: A Database System that Can Start Out Small and End Up Big

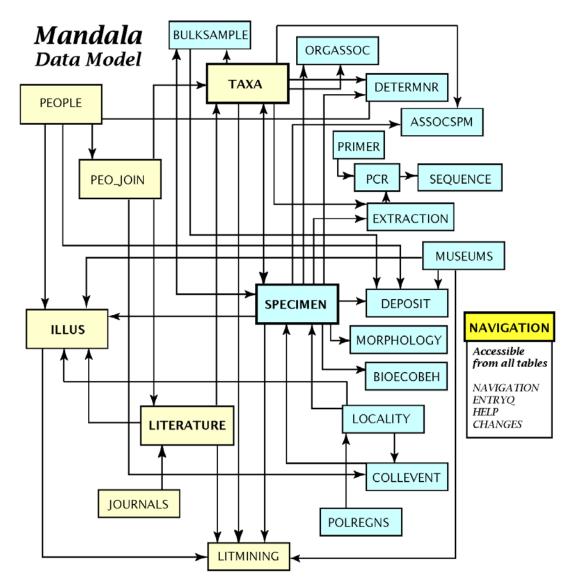
Gail E. Kampmeier Illinois Natural History Survey

Mandala<sup>™</sup> is a database system for tracking detailed information about specimens, nomenclatural history of taxonomic names, scientific literature, and illustrations. Written using the cross-platform (Macintosh and Windows operating systems) database engine, Filemaker® Pro, Mandala has been optimized for specimen-based systematics studies of a target group of organisms; for tracking samples in bioinventories, which are collected in bulk, sorted, labeled, and shipped as loans to specialists for identification; and it incorporates features for conducting a biodiversity blitz.

Mandala consists of 3 files (TAXA, SPECIMEN, NAVIGATION) with 27 interrelated tables.

Development on Mandala was begun in 1995 with support from the National Science Foundation and Schlinger Foundation. Mandala 8 was recently released and it is compatible with FileMaker 8 and above. Among the changes in this upgrade was the introduction of detailed tracking of the progress of molecular studies for specimens, following extractions, PCR products, and sequences all the way to links to GenBank.

Mandala is most suited for use by small- to medium-sized work groups. Clients using FileMaker can operate Mandala from their desktops or authorized clients can remotely access Mandala from a server hosting the files with FileMaker Server 8 or above. Project data of DarwinCore-compatible fields may be exported from the production database to a stand-alone file optimized for use in PHP-based web queries (we use FX.php, FMStudio®, FileMaker Server, and Adobe® Dreamweaver®). Interactive mapping is provided in collaboration with DiscoverLife.org. http://www.inhs.illinois.edu/research/mandala/



#### Mandala Data Model

Support is acknowledged from: U.S. National Science Foundation's program on Partnerships for Enhancing Expertise in Taxonomy (PEET) Grants DEB 95-21925, 99-77958, and 07-31528; the NSF Biodiversity Surveys & Inventories, DEB-0425790 "Terrestrial Arthropod Survey - Fiji;" and the NSF Diptera Tree of Life project, EF-03-34948.; as well as the Schlinger Foundation, the University of Illinois Agricultural Experiment Station Hatch Project (ILLU 875-380), and the Illinois Natural History Survey. A special thanks is due to John Pickering and DiscoverLife.org for encouragement and mapping integration between the online PHP databases and DiscoverLife.

### 18.4. The new EDIT Specimen and Observation Explorer for Taxonomists

Patricia Kelbert, Niels Hoffmann, Jörg Holetschek, Anton Güntsch, Walter G. Berendsohn BGBM Berlin

Since 2001, the GBIF (Global Biodiversity Information Facility) has been networking worldwide biodiversity data and making them freely available on the Internet. Currently, more than 150 million records are stored in the GBIF database and accessible through several web portals. In parallel, European projects such as SYNTHESYS and BioCASE (Biological Collection Access Service for Europe) focus on developing specialised web services and Internet portals to access collection and observation data in the GBIF infrastructure.

The generic XML-based web-service Thesaurus Optimized Query Expander (TOQE) was developed at the BGBM to allow access to thesaurus databases of any kind. It uses a fixed set of methods, thereby hiding the complexity of the underlying thesaurus. A taxonomic checklist is used as the source for a thesaurus's synonyms and related concepts. Results for a scientific name are delivered as well-formed XML documents.

The BioCASE portal (<u>http://search.biocase.org/europe</u>) is a web portal that allows users to search European biodiversity data. The BioCASE portal uses TOQE to incorporate the Euro+Med Plantbase checklist for European flora and the Fauna Europaea checklist for European fauna, expanding user queries to include synonyms and related concepts for a given name. Using thesaurus expansion in the BioCASE portal improves search efficiency by discovering records that may (or in some cases may not) be identified by a related name, but it is an on/off mechanism: it does not allow users to control the query expansion process.

To address this shortcoming, the TOTO prototype for checklist-driven access to collection and observation data (<u>http://search.biocase.org/toto</u>) was developed by SYNTHESYS. It improves the BioCASE portal's query mechanism by giving users full control of the query expansion process. Users can:

- o choose which thesaurus to use
- choose to include or exclude types of relationships (synonyms, related taxa in the taxonomic hierarchy, related taxonomic concepts such as misapplied names)
- o individually mark or unmark discovered "related" names for inclusion in the search.

The BioCASE portal and TOTO have been developed further under the name "EDIT Specimen and Observation Explorer for Taxonomists" (<u>http://search.biocase.org/edit</u>), and made part of the EDIT Platform for Cybertaxonomy. Available in 11 languages, this new tool provides users with fast and easy-to-use access to worldwide biodiversity data, and offers full control over query expansion. The portal accepts one or more Latin names, suggests related query terms for both zoological and botanical data, expands the query accordingly and offers complete BioCASE portal functionality for resulting specimen and observation data.

As of June 2008, this new explorer for taxonomists uses the following sources for query expansion:

- "Euro+Med Plantbase" an information resource for Euro-Mediterranean plant diversity
- "European Register of Marine Species" an authoritative taxonomic list of species occurring in the European marine environment
- "Fauna Europaea" a database of scientific names and distribution of all living multicellular European land and fresh-water animals
- o "German standard checklist" the standard list of the ferns and flowering plants of Germany
- The Reference list for the German Bryophytes (Koperski & al.)

# **18.5.** EDIT Data Portal – A flexible publishing tool for biodiversity information networks

Andreas Kohlbecker<sup>1</sup>, Andreas Müller<sup>1</sup>, Niels Hoffmann<sup>2</sup>

<sup>1</sup> Botanic Garden & Botanical Museum Berlin-Dahlem, <sup>2</sup> Department of Biodiversity Informatics and Laboratories, Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin

The EDIT (European Distributed Institute of Taxonomy) Data Portal, one of the most important components of the EDIT Platform for Cybertaxonomy, is a powerful tool for building websites for scientific communities sharing a common taxonomic interest. It is a user-friendly but detail-rich application for the online publication of valuable scientific work in progress.

The EDIT Platform for Cybertaxonomy is an extensible set of distinct tools supporting the work of taxonomic communities. A community stores its data in a so-called CDM Community Store, which is

accessed by applications such as the EDIT Data Portal via simple web services. CDM stands for Common Data Model, the domain model for core EDIT components. The CDM is primarily based on the TDWG ontology, although it also builds upon the work of other models and standards.

The EDIT Data Portal publishes taxonomic data either to a closed community or to the public at large. It has been implemented as a module for the content management system Drupal. The term "content management system" is somewhat misleading here: since the actual content is stored and managed by the CDM, ensuring scientific accuracy and reusability of data. Drupal offers a great deal of flexibility and can be extended almost indefinitely, making it possible to adapt the EDIT Data Portal to the individual needs of a community. Custom web pages, forums, and blogs are only a few of Drupal's capabilities available out of the box, and there is practically no limit to the number of modules that can be added to enhance a site with for example image galleries. The layout and look and feel of the application can be easily adapted to fit any taste.

The EDIT Data Portal supports zoological and botanical names (support for viral and bacterial names should be available soon), as well as multiple concurrent taxonomies. Rules as specified by the nomenclatural codes ICBN and ICZN are built into the CDM, ensuring that taxon names conform to these standards.

User interface design was guided by the following design principles: provide easy access to all taxa and names; display all relevant information in one place where possible; and reduce complexity without compromising scientific accuracy. A challenge particular to taxonomy is keeping the user from losing his or her bearings while navigating through vast taxonomic trees; to this end, a sophisticated taxonomic tree browser has been developed to reduce confusion and make daily work with such trees more productive. A search function provides an additional way to access taxa.

Each taxon is represented by an individual page, which displays the list of synonyms, related taxa, occurrences, descriptions, media, and more. The list of homotypic and heterotypic synonyms is ordered historically and grouped by types; its layout resembles that found in printed publications. Images of protologues and of type specimens are linked directly from taxon names and synonym lists. In addition, taxon pages will allow users to contribute annotations as text or images.

Currently, EDIT Data Portals and CDM Community Stores are used to store and publish the work of the EDIT exemplar groups focusing on the taxonomic groups Cichorieae, Diptera and Palmae.

### 18.6. NatureGate: An Online Service to Promote Education for Sustainable Development and Biodiversity

Eija Lehmuskallio<sup>1</sup>, Arja Kaasinen<sup>2</sup>, Mauri Åhlberg<sup>2</sup> <sup>1</sup> NatureGate Ltd., <sup>2</sup> University of Helsinki

NatureGate is a new, open access tool for promoting both local and global education on biodiversity, sustainable development, and sensitivity to environmental issues for a better quality of life. This free online service offered by the company NatureGate Ltd, and supported by a worldwide patent, has been developed in cooperation with the Research Group of Biology and Sustainability Education at the University of Helsinki.

The NatureGate Online Service (<u>http://www.naturegate.fi/</u>) offers rapid, interactive, and easy-to-use identification software for species (or any type of object). The software supports both identification and subsequent finding of the biological and ecological details about a species through a species home page. NatureGate online service can be used for planning a trip, or out in the wild through a mobile phone, if there is a fast connection to Internet. After the field trip, NatureGate online service can be used to verify the observations. If users have access to mobile phones or cameras with GPS (Global Positioning System), NatureGate can accept the georeferenced data as people upload their digital photos.

Species may be linked to information about ecosystem services (benefits accruing to humans from an ecosystem) and ecosystem models, which may help advise managers on sustainable use of biodiversity, sustainable development, environmental education, etc. There will be special NatureGate online publications connected to the core species identification service, with a magazine expected later in 2008. Languages supported on the website now include Finnish, Swedish, English, and German; French and Spanish are anticipated later in 2008.

The software is easy to use. Many wild flowers, trees, and bushes, and certain invasive species are easily recognizable by observable characteristics. However, even for characters that can be easily described, there are often too many matching species to consider, if one is not working with a certain group of organisms every day. Options to quickly and reliably reduce the number of characters needed to make an accurate identification are therefore paramount. This is just what NatureGate software and online service offers for plants. Identification software will also be developed for birds, butterflies, rocks, and minerals during 2009.

The number of plant groups and species covered by the service will continue to grow. Scandinavian wild flowers are featured in the currently offered pilot service. Later on, North American wild flowers will be added. NatureGate already boasts nearly 400,000 digital images of nature objects (*e.g.*, plants, butterflies, birds, fungi, and their habitats) but collaboration with nature photographers and other interested partners will be welcomed to expand the services.

Support is acknowledged from: University of Helsinki Funds, Nokia Corporation.

### 18.7. NBII Clearinghouse (Mercury)-Distributed Metadata Management, Data Discovery and Access System

#### Giri Palanisamy, Vivian Hutchison, Mike Frame ORNL

Metadata records that describe scientific datasets in a standardized format are a valuable way for scientists and researchers to discover completed or on-going research projects in a particular area of study. Metadata records support data management requirements by preserving the data creation history in order that data can be re-used or adapted to new science. Clearinghouses provide a critical link to data sources in multiple agencies, and as a result, can prevent duplication of scientific studies. This promotes collaboration and data sharing, benefiting both science and the financial bottom line.

Mercury is a federated metadata harvesting, search and retrieval tool (<u>http://mercury.ornl.gov</u>) based on both open source tools and Oak Ridge National Laboratory (ORNL)-developed software. It was originally developed for National Air and Space Administration (NASA), and the Mercury development consortium now includes funding from NASA, U.S. Geological Survey (USGS), and the Department of Energy (DOE). Mercury supports widely used metadata standards such as Federal Geographic Data Committee (FGDC), Dublin Core, Darwin Core, Ecological Metadata Language (EML), and ISO-19115, and protocols and specifications such as XML and Z39.50.

The National Biological Information Infrastructure (NBII) Clearinghouse tool, developed using Mercury, serves as a portal for record and data discovery. It represents a single gateway to metadata records and data on disparate data management systems. Contributors to the Clearinghouse are able to distribute their metadata while maintaining complete control and ownership. This leverages investment in existing information systems and research. Users of the NBII Clearinghouse range from biologists in Federal and State Agencies, land managers, data managers, research scientists, and the public.

In April 2008, a newly enhanced version of the NBII Clearinghouse (<u>http://mercury.ornl.gov/nbii</u>) with powerful search capabilities and updated features debuted. The NBII Clearinghouse makes available over 40,000 metadata records contributed by 41 partners.

The newly redesigned user interface organizes metadata records by presenting relevant records through a ranking and filtering system controlled by the user. The new NBII Clearinghouse system is based on a Service Oriented Architecture and supports enhanced search features, including an ability to search geographically or by specific data provider, then bookmark or email record results. Users can view search results in a combined view from all providers, or use filters to dynamically sort search results. A Really Simple Syndication (RSS) feed can be set up to inform a user about new records in the Clearinghouse reflecting a particular query. RSS feeds are currently being broadcast to several NBII partners including the Global Forestry Initiative, World Data Center for Biodiversity and Ecology and the primary NBII web-site (www.nbii.gov).

Support is acknowledged from: NASA, USGS, DOE

#### 18.8. IRMNG – The Interim Register of Marine and Nonmarine Genera

Tony Rees

CSIRO Marine and Atmospheric Research

IRMNG, the Interim Register of Marine and Nonmarine Genera (<u>www.obis.org.au/irmng/</u>), is an initiative of OBIS Australia implemented at CSIRO (Commonwealth Scientific and Industrial Research Organisation) Marine and Atmospheric Research, Australia (CMAR), which aims to provide a machine-addressable system that will discriminate marine from non-marine, and extant from fossil taxa in species lists, either on the basis of their genus alone, or (if required) by the full species binomen. To enable this, a database of as many genus names as possible (with associated species names as readily available) has been compiled at CMAR, with all names embedded in a "provisional" or working taxonomic hierarchy, with the facility to mark up ("flag") families, genera, or individual species with their relevant characteristics – either marine, non-marine, or both (where "marine" also includes brackish water species), and either extant, fossil, or both.

This project presents interesting challenges in two areas: first, assembly of a sufficiently comprehensive reference database (which could be viewed as an expanded "Catalogue of Life" with particular emphasis at genus level and above, also including extinct organisms), and second, application of the relevant habitat and extant flags in as efficient a manner as possible. The present (version 1) IRMNG database has been assembled as a merger of information from over 40 printed and electronic sources, and currently includes some 220,000 genus names and more than 1.4 million species (with hybrids and infraspecies currently excluded). With the addition of data not yet loaded, IRMNG version 1.1 is likely to include significantly more names, with the aim of over 90% complete coverage at genus level.

Application of the relevant habitat and extant flags is implemented as a hierarchical approach, in that flags are applied at as high a level as possible in the taxonomic hierarchy, with all child taxa inheriting flags from their parent taxon wherever the latter are flagged unambiguously: marine or non-marine, extant or fossil, *e.g.*, all echinoderms are marine, all dinosaurs are fossil, most insects are non-marine, etc. By this means, the task of applying flags at an individual genus or species level is considerably reduced. In addition, reasonably comprehensive (although in some aspects dated) family-level treatments, for example Parker's "Synopsis and Classification of Living Organisms" (1982) and Benton's "The Fossil Record 2" (1993), are followed as far as available. Flagging then needs to proceed to the genus level and below only where a particular family has an ambiguous habitat or extant flag (*i.e.*, both permitted states present).

Mechanisms for the further ongoing development and population of IRMNG, potentially in collaboration with similar or overlapping projects elsewhere, are appropriate topics for discussion.

Support is acknowledged from: OBIS Australia (part); CSIRO Marine and Atmospheric Research. Creation of the initial genus-level coverage of IRMNG was assisted considerably by the re-use of data generously made available by original custodians Sheila Brands (Netherlands) and Ellen Farr (U.S.A.), from the Systema Naturae 2000 and Index Nominum Genericorum compilations, respectively. OBIS Australia contributed financially to the compilation of habitat and extant flags for several key groups. I also thank numerous other contributors of individual taxonomic data compilations that have been used as input to IRMNG, as listed on the IRMNG website.

# 18.9. Using AquaMaps for biodiversity assessment, including a prototype MPA (Marine Protected Area) network design tool

Tony Rees

CSIRO Marine and Atmospheric Research

AquaMaps (<u>www.aquamaps.org</u>/) is a biogeographic and biodiversity information tool for marine species developed by R. Froese, K. Kaschner and others. It uses a niche modelling approach to generate predicted distributions for marine species based on presently available distribution data and/or described ecological tolerances. In essence, provided that a minimum "acceptable" level of information is available, the tool will generate a modelled (unreviewed) predicted distribution for any species, as calculated probability of occurrence in all 0.5 x 0.5 degree cells in the ocean. This model can then be further refined if necessary by appropriate experts with relevant detailed knowledge of the particular species concerned and its distribution. Most species information systems display either individual data points or approximate, hand-drawn maps to depict known areas of species' presence. The aim of AquaMaps is to produce an improved, transparent and reproducible version based on the available point data and other information. The system attempts to "intelligently" fill in gaps between disparate data points where suitable environmental conditions are deemed to exist. AquaMaps produces modelled quantitative (probability) distribution maps that are independent of sampling effort or the geographic patchiness of available point data.

Once predicted (modelled) distributions are available for a sufficiently large number of taxa, and their quality is considered acceptable, it then becomes possible to use AquaMaps data in an integrated fashion for biodiversity assessment purposes. An example would be to determine the predicted diversity of all, or any desired subset of taxa at any scale from global down to the limit of resolution of the system (0.5 degree squares, measuring approximately 55 x 55 km at their maximum size, adjacent to the equator). As one of a range of currently available options, a prototype MPA (Marine Protected Area) network design tool has been developed that permits a user to specify a subset of species according to their particular characteristics for conservation purposes. These currently include endemism; IUCN listing; resilience (capacity to withstand exploitation); and overfishing status, in a region of choice (currently large marine ecosystem, country's exclusive economic zone, or any predefined FAO area). The system will then plot a map in real time, colour-coded as to the number of the selected species occurring in every 0.5 degree cell; clicking on any cell will generate the respective list of species with probability of occurrence, plus additional information on the locality. Areas of highest computed biodiversity may be considered as possible MPA sites.

In practice, the tool is limited by the availability of suitable base data (currently some 8,000 species maps have been generated covering fishes, marine mammals and selected invertebrates) and the status of the maps themselves (auto-generated, *i.e.*, unreviewed, vs. expert reviewed). The potential of the system is apparent from its utility, content and associated tools that have been developed or envisaged.

Support is acknowledged from: Pew Foundation; European Commission/IncoFish. AquaMaps has been developed by K. Kaschner, J. Ready, E. Agbayani, P. Eastwood, T. Rees, K. Reyes, J. Rius and R. Froese. The AquaMaps MPA Planning Tool was designed by Rainer Froese and Will Le Quesne and implemented by Eli Agbayani in the context of the European Commission INCOFISH Project

# 18.10. Open Access Data Policy and Tools for its Implementation at the Finnish Museum of Natural History

Hannu Saarenmaa<sup>1</sup>, Miruna Badescu<sup>2</sup>, Alex Morega<sup>2</sup> <sup>1</sup> Finnish Museum of Natural History, <sup>2</sup> Eau De Web S.l.r.

The Finnish Museum of Natural History (FMNH) is a national institution that keeps in its collections about 15 million specimens and has an equal number of observational records in its databases. In early 2008, the FMNH adopted a new strategy for its biodiversity informatics activities, including a new open access data policy. The policy, which is available at

<u>http://www.fmnh.helsinki.fi/english/about/data-policy.htm</u>, requires that all data be made publicly available through the Internet, but sometimes with delay, generalisation, and with certain necessary restrictions for sensitive data. The aims of the policy are to comply with the requirements of the national legislation, be ready for implementation of the EU INSPIRE directive, motivate the data collators and allow them respite to first publish results of their data, and in general increase the productivity of the FMNH through a more efficient and wider use of data holdings.

Implementation of the strategy and data policy requires efficient tooling. The GBIF Data Repository Tool, which is being used by a number of GBIF Nodes worldwide for data sharing, has been customised to support the new requirements. This tool can be used to archive and share smaller datasets in CSV spreadsheet format. It enables the owners to directly manage their data files and control their release. The Repository Tool supports Dublin Core and DiGIR metadata for each dataset. It also allows setting up the parameters for fuzzying or hiding geographic coordinates, exact locality, time, and collector. In compliance with the policy rules, data from the archived spreadsheets is parsed and copied into a cache which has openly accessible DiGIR and TAPIR interfaces. There also is a simple TAPIR search tool to present the data locally. A demo site is available through http://www.gbif.fi/repo/.

A central data warehouse is also under construction for FMNH. This warehouse will be used as the main vehicle for enabling public access to large institutional databases which are constantly updated and for which spreadsheet format is not suitable. Data from the operational source databases are regularly copied into the data warehouse through a consolidation process, where taxonomic and geographic details are harmonised and metadata are augmented using EML. The data warehouse itself is not publicly available, but data from it are copied to a publicly available cache which has a TAPIR interface. During the copying, the data are filtered in a similar fashion as in the Data Repository.

With these two tools, public access to data do not need to be set up for each and every dataset separately, but can be managed centrally. The idea is to remove the obstacles of data sharing to a minimum and cope with the cultural change of sharing data freely and openly. The tools keep the data custodians in control of their own data, and help in complying with the institutional policy.

# **18.11.** Feature recognition for herbarium specimens (Herbar-Digital)

Karl-Heinz Steinke, Robert Dzido, Martin Gehrke, Klaus Prätel University of Applied Sciences & Arts Hannover

Our research project, "Rationalizing the virtualization of botanical document material and their usage by process optimization and automation (Herbar-Digital)" started on July 1, 2007 and will last until 2012. Its long-term aim is the digitization of the more than 3,5 million specimens in the Berlin Herbarium. The University of Applied Sciences and Arts in Hannover collaborates with the department of Biodiversity Informatics at the BGBM (Botanic Garden and Botanical Museum Berlin-Dahlem) headed by Walter Berendsohn. The part of Herbar-Digital here presented deals with the analysis of the generated high-resolution images (10,400 lines x 7,500 pixel).

The image of a herbarium specimen may contain different kinds of objects, such as: mounted plant material; envelopes with additional parts of the plant; printed or handwritten labels with printed

headlines (or without); annotation labels; a metric chart; a colour chart; stamps indicating ownership, specimen donation, accession numbers, or other events in the history of the specimen (*e.g.*, digitization); specific markers (such as the red "Typus" label in the Berlin herbarium); barcode; and handwritten annotations directly on the sheet. Known, recurring stable objects can be found automatically by template matching methods (strong classifiers), which compute the similarity between the image and the template at each position in the whole image. The similarity is transformed to a brightness value. By choosing the maximum brightness, the location of the template in the image is resolved, and the object can be cut out of the image. More variable objects require more sophisticated methods, based on classifiers, *i.e.*, entities that serve to classify an object into categories. One approach uses boosting-algorithms from face recognition applications. Boosting-algorithms utilize an iterative approach that combines the results of several weak classifiers to arrive at a decision to a problem instead of matching a single strong classifier. The idea suggests that it is easier to find some rules of thumb than relying on a general rule to resolve the problem.

More than 30% of the images contain handwriting. Because commercial optical character recognition (OCR) software is not able to interpret handwriting in old documents, we developed an algorithm to distinguish handwriting in complex surroundings such as those on a herbarium sheet, where it may be mixed with leaves and roots. With the help of an interactive method, single characters of handwriting may be extracted from the text. The characters can be described mathematically *e.g.*, by Legendre polynomials and the coefficients are then stored for comparison in a database. The aim is to identify the writer, for example "Alexander von Humboldt".

The computer demonstration will show several approaches for image analysis. First, stable objects like stamps are identified and eliminated. Afterwards, areas containing writing are localized and analysed by self-developed methods or by Optical Character Recognition. An adaptive boost classifier can find unstable objects. Identification of coloured objects is possible by means of a colour space transformation. Eventually, the separation of the plant from the background is done. All the approaches have to be combined in order to extract the largest amount of information from the image.

Support is acknowledged from: the state of Lower Saxonia and the Volkswagen Foundation.

# **18.12. INOTAXA implemented - a digital future for taxonomic literature**

Anna Weitzman<sup>1</sup>, Chris Lyal<sup>2</sup> <sup>1</sup> Smithsonian Institution, <sup>2</sup> The Natural History Museum

INOTAXA (INtegrated Open TAXonomic Access) is a means of unlocking and accessing taxonomic information from published literature. It provides a means of locating and accessing relevant information within the content of taxonomic publications, including taxonomic descriptions, identification keys, catalogues, names, specimen data and images. The project has been set up to provide access to users around the world, complementing other digitization initiatives by allowing much more detailed search and data-mining than is otherwise possible. INOTAXA is designed as a tool primarily for taxonomists, but will serve other users; responses to this prototype will assist in developing new tools and refining / modifying data content and display in responses to queries. All of the content accessible through INOTAXA can be searched or browsed simultaneously with a single command, and results displayed together for selection of a more detailed view. The INOTAXA interface, which was designed based on a user needs assessment and refined during development following tester feedback, allows access to taxonomic literature marked up in a purpose-written XML schema, taXMLit, which itself provides interoperability with other TDWG standards. The schema is highly atomized, providing the means to isolate data elements (*e.g.*, taxon names, citations, specimen data) for subsequent repurposing. Although INOTAXA is mainly focused on digitized taxonomic publications, it will expand in future to allow simultaneous search of specimen data, gazetteers and information elsewhere on the web, using interoperability of the standard schemas.

In the INOTAXA prototype, searches may be simple and unrestricted, or restricted by factors we have found valuable for taxonomists: geography, taxon, publication and context, or refined still further by

use of Boolean logic across some 50 indexed fields. Alternatively, the content can be browsed by taxon, geography, or people (as authors, collectors, or editors) to provide data and information.

The functions of the prototype are to:

- (1) Provide proof of concept to support further development;
- (2) Identify user needs through feedback;
- (3) Demonstrate the value of the approach to the user community;
- (4) Provide a real resource for taxonomists working on the groups covered.

Support is acknowledged from: Atherton Seidell Fund and the National Museum of Natural History of the Smithsonian Institution; the Natural History Museum, UK and Biodiversity Informatics Standards (TDWG). Site developed by Information International Associates.

# **19.** Poster

### **19.1. ETI BioPortals: customizable common access system to distributed heterogeneous** data sources

#### Ir Wouter Addink, Dr Peter Schalk ETI BioInformatics

Today, much knowledge on biological diversity (*i.e.*, on taxonomy, ecology, genetics, conservation, legislation) is available in digital form. The information is often compiled for a specific purpose and stored in custom-made, geographically distributed databases using different formats. Mining this information can therefore be time consuming, and recombination of data can be cumbersome because of incompatibilities. International initiatives, such as the Global Biodiversity Information Facility (GBIF), the Encyclopedia of Life (EoL), and the Consortium for the Barcode of Life (CBOL), are paving the way to make data in a specific domain globally accessible.

However, much of the demand for information is on a national or local level, driven by defined user groups with specific questions or problems. These call for a custom-made answer to their information needs. ETI developed a 'Google-like' web portal solution that provides a single access point to a large array of heterogeneous information sources. The so-called BioPortal can be customized to specific information needs and user levels: from governments to schools.

The BioPortal toolkit has a modular design that can be easily tuned to user needs, including customizing the web-interface and providing support for multiple languages. The toolkit combines content management system (CMS) functionality, like news items and static pages, event tracking, and forum modules, with functions to access the scientific core data. These are biological collections, observations, ecological relations, molecular data, and image libraries compiled on a national or local level, combined with external sources such as GBIF or EMBnet (European Molecular Biology Network). Sophisticated technology allows for simultaneous server-side asynchronous searches in several distributed data sources on the web.

The toolkit includes the Linnaeus II Species Bank compiler that allows the creation of targeted (*e.g.*, national) species information systems and identification tools for the Internet (*e.g.*, <u>www.soortenbank.nl</u>) or mobile phone platforms. The systems can also be dynamically linked to external sources (*e.g.*, EoL) and can provide information in TDWG SDD format (Standard for Descriptive Data). The BioPortal also integrates the TDWG NCD (Standard for Natural Collections Descriptions) toolkit that can be used to collect and share information on national capacities (collections, experts, museums, research institutes, national parks). Another module gathers biodiversity news from a range of national and international (RSS feed) sources. The Geographical Information System (GIS) module uses Google Maps combined with local GIS data to provide advanced search options by location. The BioPortal can deploy TapirLink (used by GBIF) with

TAPIR (TDWG Access Protocol for Information Retrieval), using TDWG DarwinCore standard data and custom web services. The BioPortal uses the Catalogue of Life as a validated taxonomic index to match searches to connected data sources.

ETI's BioPortal toolkit was used to build the Netherlands' national GBIF portal (NLBIF), <u>http://www.nlbif.nl</u>. More recently it provides a foundation for the Tanzanian national biodiversity portal (TanBIF: <u>http://www.tanbif.or.tz</u>). This portal is optimized for their national information needs with emphasis on the use of the information in education and outreach (tourism). Colombia and Bulgaria are considering similar BioPortal projects.

Support is acknowledged from: NLBIF, GBIF

# 19.2. Sampling biodiversity sampling

Arturo H. Ariño, Javier Otegui University of Navarra

The world's vast biodiversity data repository, the Global Biodiversity Information Facility (GBIF), federates taxonomical databases that, presumably, encompass most of the taxonomically- and geographically-oriented, openly accessible occurrence data. Therefore, GBIF can be considered the best sampling of biodiversity information available in the world. However, to conclude that this very vastness (more than 145 million records) should account a priori for high significance and confidence, may be misleading.

For a sample to represent the universe being sampled, a number of statistical assumptions must be made. Most importantly, the samples should ideally be random samples from across any potential pattern of interest. If, on the other hand, samples do have a hidden intrinsic pattern, any resulting observed pattern might not reflect the universe being sampled but, rather, the sampling bias. Thus, use of this mass of available data is subject to knowing or estimating what inferences can be drawn now, and what others cannot (yet) be ascertained.

Some of the GBIF data biases are well known or easy to discover. For instance, the famous geographical concentration of available data from the Western hemisphere, incommensurate with the biodiversity hotspots distribution, can be readily observed in the map of all geolocated records already available at GBIF portal. Other patterns, however, may be suspect but less obviously assessed or even detected. For instance, taxonomical patterns in the data, such as taxon concentration in geographical areas, may require cross-examination of each data provider's own geographical and taxonomical bias.

An obvious practical problem for such studies is the difficulty of putting into play all relevant records. As many data providers and databases are thematic, patterns that transect or combine space, time, and taxa might require the examination of very large tracts of available data. Queries performed on indexes, such as the ones that GBIF supplies, may not answer certain questions being asked. In the most extreme case, one wishing to analyze certain patterns in the data should, ideally, have immediate access to the full dataset. This could amount to downloading all datasets from all data providers, which is impractical for most.

In this work, GBIF data were queried using sampling techniques adapted in part from ecological studies, and inferences drawn from these smaller samples were compared to inferences drawn from the full dataset. Also, samples were used to search for specific patterns in GBIF data that may, among other things, (a) influence the use of these data as samples for ecological studies, and (b) very significantly, allow for the discovery of knowledge gaps, identifying data regions worth exploring such as underrepresented taxa, neglected areas, etc.

Some of these patterns were already discovered by newly developed exploratory procedures in data from our own database, Zootron 4, during a previous series of analyses presented at TDWG-2006 and TDWG-2007. Other patterns were conceptually new, as they pertain to the multi-dataset provenance of GBIF data.

Performing these analyses to detect patterns in GBIF data did not prove especially taxing to the load on GBIF servers, as only standard web queries were performed, requiring no complex table interaction. This is particularly relevant in view of the performance of the database engines under large size asymmetry (see Holetscheck, this issue).

# **19.3.** A Global Information System for the Conservation and Sustainable Use of Plant Genetic Resources for Food and Agriculture (PGRFA)

Elizabeth Arnaud<sup>1</sup>, Michael Mackay<sup>1</sup>, Dag Terje Filip Endresen<sup>1</sup>, Sonia Dias<sup>1</sup>, Rajesh Sood<sup>1</sup>, Kiran Viparthi<sup>1</sup>, Milko Skofic<sup>1</sup>, Adriana Alercia<sup>1</sup>, Tito Franco<sup>1</sup>, Frederick Atieno<sup>1</sup>, Xavier Scheldeman<sup>1</sup>, Anzar Shamsie<sup>2</sup>, Selim Louafi<sup>3</sup>, Pete Cyr<sup>4</sup>

<sup>1</sup> Bioversity International, <sup>2</sup> International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA) Secretariat, <sup>3</sup> International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA, <sup>4</sup> United states Department of Agriculture-Agricultural Research Service

The overall goal of the project "Global Information on Germplasm Accessions" is to improve access by breeders and other users to the germplasm they need in gene banks around the world. The project covers all major food crops, with a focus on 22 crops: banana, barley, beans, breadfruit, cassava, chickpea, coconut, cowpea, faba bean, finger millet, grass pea, maize, major aroids, lentil, pearl millet, pigeon pea, potato, rice, sorghum, sweet potato, wheat, and yam, and a limited number of other crops of interest to the project collaborators.

The project specifically addresses the challenge of making readily available information about germplasm collections, which has been repeatedly identified as a key to increasing their use. It is being carried out in three components: 1) Developing information standards to describe the characteristics of genetic resources of most interest to users, by mobilizing germplasm users to agree on core subsets of characterization and evaluation standards; 2) Deploying the state-of-the-art global Germplasm Resources Information Network (GRIN-Global) gene bank data management system being developed by the United States Department of Agriculture (USDA). GRIN-Global will be made available to national programmes and other gene banks; and 3) Building a global system for accessing and managing accession-level germplasm data to link national, regional and international gene bank databases in support of the conservation and use of the Plant Genetic Resources for Food and Agriculture (PGRFA) and in the framework of the International Treaty.

This global system will be based on two major databases:

- SINGER (System-wide Information Network for Genetic Resources), which gathers the records of 11 international CGIAR (Consultative Group on International Agricultural Research) gene banks with records from the World Vegetable Center (AVRDC) together, the members of SINGER hold more than half a million samples of crop, forage, and tree diversity in their germplasm collections (<u>http://singer.cgiar.org</u>); and
- EURISCO (European Internet Search Catalogue), which is a web-based catalogue that provides information about ex situ (maintained or found outside their natural habitat) plant collections maintained in Europe. EURISCO is based on a European network of ex situ National Inventories (NIs) and contains 1.1 million samples of crop diversity representing 1,450 genera and 8,665 species from 38 countries (<u>http://eurisco.ecpgr.org/</u>).

The project "Focused Identification of Germplasm Strategy (FIGS)" (<u>http://www.figstraitmine.org/</u>) is a final component of this network. This website allows users to efficiently integrate the data associated with this collection and facilitates identification of custom subsets of accessions with traits that may be of importance to breeding programmes.

Support is acknowledged from: the Global Crop Diversity Trust ; ITPGRFA Secretariat

# 19.4. LifeWatch – e-Science and Technology Infrastructure for Biodiversity Research

Walter G. Berendsohn, Marie Gebhardt

LifeWatch is an initiative for a distributed European research infrastructure, which will link the domains of ecological research and monitoring data collected from marine and terrestrial observatories with the vast amount of data from biological collections. LifeWatch was initiated by a collaboration of existing European networks active in the domain of biodiversity research (see footnote).

LifeWatch plans to construct and bring into operation the facilities, hardware and software as well as governance structures to create a massive biodiversity research infrastructure.

The new infrastructure would revolutionize European biodiversity research and biodiversity informatics by bringing it to a novel scale of cooperation and interoperability. It would open up new areas of research and new services by creating and providing access to large data sets from different (genetic, population, species and ecosystem) levels of biodiversity together with analytical and modelling tools.

The sheer size of the infrastructure with respect to costs (first estimate at €1.5 billion over 25 years), functionalities, and user communities will require large-scale international cooperation for its implementation. The European Strategy Forum on Research Infrastructures (ESFRI) identified LifeWatch as a European Research Infrastructure to be supported by European countries. Currently 19 European countries have expressed interest in the initiative. Earlier this year, the European Commission granted a 5 million Euro financial support for a three-year preparatory phase, with the aim of rallying national support and developing a construction plan for the LifeWatch infrastructure.

Workpackage 5 of LifeWatch (Construction Plan Strategy) is jointly coordinated by the BGBM Berlin-Dahlem, the Fraunhofer Institutt für Intelligente Analyse- and Informationssysteme and Cardiff University. It is working on construction logistics and technical plans for a construction phase of the LifeWatch infrastructure. Aims include the preparation and assessment of the construction specifications, such as enabling data capture, data flows, analytical and modelling tools, web portals, Grid infrastructure, and system management. Case Studies contributed by country members are to test and demonstrate the potential of the infrastructure.

If the LifeWatch construction phase is realised, it will strongly support and integrate TDWG's efforts, *e.g.*, in the definition and interoperability of ontologies and the cooperation with further standards organisations.

For further information see <u>www.lifewatch.eu</u> and the websites of the initiating networks: ALTER-Net (A Long-Term Biodiversity, Ecosystem and Awareness Research Network - <u>www.alter-net.info</u>), BioCASE (Biological Collection Access Services – <u>www.biocase.org</u>), EDIT (European Distributed Institute of Taxonomy - <u>www.e-taxonomy.eu</u>), ENBI (European Network for Biodiversity Information – <u>www.enbi.info</u>), EurOCEANS (European Network for Ocean Ecosystems Analysis -<u>www.eur-oceans.eu</u>), MarBEF (Marine Biodiversity and Ecosystem Functioning – <u>www.marbef.org</u>), MGE (Marine Genomics Europe - <u>www.marine-genomics-europe.org</u>), and SYNTHESYS (Synthesis of Systematic Resources - <u>www.synthesys.info</u>).

Support is acknowledged from: European Commission

# **19.5.** Proof of concept study of the Socio-Ecological Research and Observation oNTOlogy (SERONTO) for integrating multiple ecological databases.

Nicolas Bertrand<sup>1</sup>, Herbert Schentz<sup>2</sup>, Bert Van der Werf<sup>3</sup>, Barbara Magagna<sup>2</sup>, Johannes Peterseil<sup>2</sup>, Sue Rennie<sup>1</sup>

<sup>1</sup> Centre for Ecology and Hydrology, <sup>2</sup> Umweltbundesamt, <sup>3</sup> ALTERRA

ALTER-Net is a network of excellence for Long-Term Biodiversity, Ecosystem and Awareness Research spanning 24 institutions in 7 European countries. The aim is to develop an integrative research framework in biodiversity research and monitoring to address biodiversity issues at a European scale.

A key objective is the development of a framework for distributed data, information and knowledge management. The major challenge in achieving this objective is the provision of consistent data access and querying across multiple institutions and diverse data types.

Semantic approaches to data integration are seen as an enabling mechanism to carry out integrated socio-ecological science at a global scale. The Socio-Ecological Research and Observation oNTOlogy (SERONTO) has been developed building upon Umweltbundesambt's (Federal Environment Agency – Austria) experiences in developing a semantic database system for managing environmental data.

To validate the development of SERONTO and its uses for future data integration, a proof of concept study was conducted. The scope of the proof of concept was to test:

- The feasibility of mapping relational databases to SERONTO and
- The querying of the connected database(s) from the ontological view of SERONTO.

The requirements for accepting the proof of concept were:

- The databases must have different structures and must have been developed independently of SERONTO;
- The databases must feature reference lists (*e.g.*, species lists);
- The database structures must not be altered as a result of the integration work;
- o New concepts may be imported into SERONTO as and when required and
- The databases must contain data relevant to Long Term Ecological Research (*e.g.*, vegetation surveys, records of species occurrences, measurement of biotic and abiotic components).

For the proof of concept, we used OntoStudio<sup>TM</sup> (1), the reasoning language F-Logic(2), and a platform for managing connectivity to the databases, semantic mappings and queries.

Results have been encouraging and demonstrate that SERONTO can be combined with F-Logic to query ecological data in disparate databases.

References:

- 1. http://www.ontoprise.de/de/en/home/products/ontostudio.html
- 2. Kiefer, Michael, Georg Lausen, and James Wu. 1995. Logical Foundations of Object-Oriented and Frame-Based Languages. Journal of the Association for Computing Machinery, 42:741-843.

Support is acknowledged from: We acknowledge financial support by ALTER-Net, a Network of Excellence funded by the EU within its 6th Framework Programme. We would like to thank Dirk Wenke (OntoPrise GmbH) for his support during the integration testing.

# **19.6.** The Register of Antarctic Marine Species: towards a comprehensive inventory of the Southern Ocean biodiversity

Bruno Danis SCAR-MarBIN

In the framework of the Census of Antarctic Marine Life (CAML), the Antarctic Marine Biodiversity Information Network SCAR-MarBIN has established an openly accessible online Register of Antarctic Marine Species (RAMS), to provide accurate and complete taxonomic and distribution data openly accessible online. RAMS aims at compiling and managing an authoritative taxonomic list of species occurring in the Antarctic marine environment, to establish a standard reference for marine biodiversity research, conservation and sustainable management. RAMS in turn contributes to some larger taxonomy initiatives such as the World Register of Marine Species or the Catalogue of Life. Species-level information is of fundamental importance and these scientific names are a crucial link to various kinds of data from different databases. Relying on literature sources and museum collections as well as on new records from CAML cruises, world-renowned specialists contribute their expertise to establish up-to-date and dynamic classifications of the various taxonomic groups. This permits the compilation of accurate species lists and occurrence data. The Register of Antarctic Marine Species constitutes the first attempt to complete the inventory of the Southern Ocean biodiversity from microorganisms to whales, contributing to the establishment of a sound benchmark to assess future biodiversity changes.

Support is acknowledged from: Belgium Science Policy (BELSPO) and SCAR

# 19.7. SCAR-MarBIN: free and open access to Antarctic Marine Biodiversity Data

Bruno Danis SCAR-MarBIN

Because of the unique ecological, biogeographical and political context of the Southern Ocean, and given the rate and potential impact of global changes on the Antarctic ecosystems, and their potential effect on biodiversity, an efficient mechanism for the exchange of scientific information is needed to enable sound, science-based adaptive management, in a realistic time frame, of Antarctic marine biodiversity. For research, conservation and management purposes, Antarctic marine biodiversity information has to be widely published, instantly accessible, and thoroughly checked. To answer this need, the Antarctic Marine Biodiversity Information Network (SCAR-MarBIN) has managed since May 2005 a dedicated data portal, <u>www.scarmarbin.be</u>, which provides free and open access to Antarctic marine biodiversity information. SCAR-MarBIN is supported by the Belgian Science Policy (BELSPO), as a long term and significant support for Antarctic research activities.

To date, the two main achievements of SCAR-MarBIN are the development of the Register of Antarctic Marine Species (RAMS) and of a distributed network of interoperable databases holding information on Antarctic marine biodiversity.

RAMS is a fully functional, browsable/searchable online species list of Antarctic marine species, dynamically maintained by a dedicated board of expert taxonomic editors. Using WebGIS (Chameleon, chameleon.maptools.org/), SCAR-MarBIN also allows users to visualize and download raw baseline data on the occurrence and abundance of marine organisms.

Science activities within the International Polar Year (IPY) require mandatory ongoing management of data and information. SCAR-MarBIN is designed to act as the tool that effectively safeguards the legacy of these efforts and spreads this information and knowledge within and beyond the Antarctic community.

Support is acknowledged from: Belgium Science Policy (BELSPO) and SCAR

# **19.8.** Development of the Integrated Publishing Toolkit by the Global Biodiversity Information Facility (GBIF)

### Markus Döring, Tim Robertson GBIFS

The new Java-based GBIF Integrated Publishing Toolkit (IPT) is under development and expected to be available for early release and testing by January 2009.

This tool has features that allow for efficient but easy sharing and hosting of organism occurrence data, taxonomic and nomenclatural information, and general dataset metadata.

By focusing on specific biodiversity data types instead of providing a generic wrapper solution, the software can provide a richer environment, *e.g.*, specialized interfaces or data validation routines. Similarly, the IPT addresses the needs of data holders to serve custom, extended data schemas that go beyond simple one-to-one extensions as provided by the Darwin Core (DwC). For example, it allows multiple taxon identifications per occurrence record. Additionally, the toolkit provides a simple web portal for data holders, which permits easy exploration of the data and offers access to simple and easy-to-use tools for data reporting and cleansing.

The IPT will continue to support the protocols and standards defined by Biodiversity Information Standards (TDWG), such as the TDWG Access Protocol for Information Retrieval (TAPIR), Darwin Core, and the Taxon Concept Schema (TCS). The IPT will offer additional interfaces such as an Open Geospatial Consortium (OGC) Web Feature Service (WFS). A simple dataset harvesting protocol will be available to allow frequent and efficient indexing by global aggregators like the GBIF Data Portal or the Ocean Biogeographic Information System (OBIS).

The poster will provide a simple-to-understand graphical illustration of the internal components of the IPT, along with clear explanations of the interfaces available, and will highlight where the tool would be positioned in the GBIF network.

Support is acknowledged from: GBIF Secretariat

# **19.9.** The DNA Bank Network – How GBIF Technology Enables a New Generation of Communicating DNA Repositories

Gabriele Droege, Holger Zetzsche, Anton Guentsch, Birgit Gemeinholzer Botanic Garden and Botanical Museum Berlin-Dahlem

The DNA Bank Network is a service to facilitate research on biological diversity. Since spring 2007, the network links separate DNA banks of four major biological research collections in Germany (Bavarian State Collection of Zoology (Munich), German Collection of Microorganisms and Cell Culture (Braunschweig), Zoologisches Forschungsmuseum Alexander Koenig (Bonn) and the Botanic Garden and Botanical Museum Berlin-Dahlem). The main target is to offer an online platform that contains information about voucher specimens, DNA samples and all associated data available to improve transparency of scientific analyses and allow for verification and further improvement of data.

The core of the network consists of two basic software systems: the DNA module and the central webportal. The DNA module was developed at the BGBM during the last year and includes a MySQL database and PHP forms for data input and administration. Three of our currently four project partners are already using installations of the module in their insitutes; an integration of older DNA information records was possible by importing Excel data files. Since June 2008, new DNA data can be entered and administered by using the module software. It is also possible to add relevant publications for sequenced data and to administer customer requests (orders, shipping, stock administration). To find the related specimen data, the module sends a query to the respective specimen database via BioCASe Provider Software (wrapper) and stores a copy of a few attributes in

the DNA cache. It is possible to connect any GBIF-compliant specimen database worldwide, using the BioCASe and DiGIR protocol.

For presenting the DNA and related specimen data on the central webportal, it is necessary to install another BioCASe wrapper on the DNA databases. The preferred ABCD 2.06 schema with the existing part for 'Sequences' does not include all required fields. ABCD offers two options to add supplementary contents: 'MeasurmentsOrFacts' and 'UnitExtensions'. Since the hierarchical structure of the DNA specific fields is too complex for 'MeasurementsOrFacts' we decided to use the 'UnitExtensions' to integrate an xml schema definition for DNA data, similar to the ABCDEFG extension for geosciences. This DNA schema definition covers more than 35 elements, *e.g.*, 'ExtractionDate', 'ExtractionMethod' or 'Concentration' and can easily be implemented into the BioCASe provider software. The DNA Sample means the ABCD-Unit and the identifier (Triple-ID) for the related specimen is defined by 'UnitAssociation'.

Presumably at the beginning of 2009 it will be possible to search for DNA data on our webportal <u>http://www.dnabank-network.org</u>.

Support is acknowledged from: DFG (German Science Foundation)

# 19.10. Invasive Species Tools for Risk Assessment in the Americas: IABIN Invasive Information Network (I3N)

Christine Fournier National Biological Information Infrastructure

Information on invasive alien species (IAS) from published and unpublished accounts and databases is often scattered in locations and formats not easily accessible. Informatics tools for collecting and organizing IAS information can help countries to better manage biological invasions. To overcome this problem, the Invasives Information Network (I3N) of the Inter-American Biodiversity Information Network (IABIN) created a distributed network of databases for the Western Hemisphere, which includes IAS experts, projects, and datasets.

The I3N Database for Invasive Alien Species Template is a database shell that both standardizes the kind of invasive species information that is collected and enables the exchange of that information in the globally recognized Extensible Markup Language (XML).

New I3N tools have been made available to assess the pathways and risks of invasion by individual species. These tools are designed to interact with, and build on, the existing I3N database information. They are freely available for download from the Web in Spanish and English: <a href="http://i3n.iabin.net/tools/web\_tools.html">http://i3n.iabin.net/tools/web\_tools.html</a>.

Support is acknowledged from: Global Environment Facility, Organization for American States, U.S. National Biological Information Infrastructure, U.S. Geological Survey.

# 19.11. Taxonomic data exchange facilities for descriptive tools

Hélène Fradin, Elise Kuntzelmann, Régine Vignes Lebbe

Identification still remains a crucial activity whatever the taxonomic study concerned: inventory, biodiversity, monitoring, revision, to name a few. For this, modern technologies can bring great help, notably through computer-aided data management and presentation.

Currently, taxonomists can use separate tools to manage their data, for example to describe observations or collections of specimens in the field, to manage personal or institutional collections, to create taxonomic descriptions, or to analyse phylogenetic data.

Creating collaborative tools available to taxonomists would help strengthen the sharing of knowledge in the taxonomic community. Encouraging this interaction among taxonomists, along with projects demanding large-scale analyses and modelling, require improving the compatibility among data resources (*e.g.*, allowing interconnection of specimen databases, as in the Global Biodiversity Information Facility, GBIF), as well as the links and exchanges among different types of knowledge management systems, *e.g.*, specimen, literature, or species data.

For this, taxonomic tools must be developed to include standard exchange formats. For descriptive data, one already usable data exchange schema is the TDWG-SDD (Structured Descriptive Data) model as the exchange format between existing knowledge management systems.

Today, several descriptive tools (DiversityDescription, Lucid, Xper<sup>2</sup>, etc.) already support the SDD format. We focus here on these programs and especially on Xper<sup>2</sup>. It is a taxonomic management system for the storage, building, modifying, and online distribution of taxonomic knowledge bases. To do this, Xper<sup>2</sup> uses its own data format, but an export/import procedure from Xper<sup>2</sup> to SDD format is available. Xper<sup>2</sup> also manages the import/export of comma-separated values (csv), text, and html files. Some preliminary tests have been done to validate these new procedures and assess the potential loss of data due to this reformatting. We will present results of our first tests of a migration between Xper<sup>2</sup> (http://lis.snv.jussieu.fr/apps/xper2/) and Lucid (http://www.lucidcentral.com/) through SDD. We will especially comment on the gap between SDD and the specific format of the two descriptive tools, and on the potentially different interpretations of the standard. We will present a tool that allows translation of numerical data into qualitative data, in particular, allowing one to import more data from SDD to Xper<sup>2</sup>.

# 19.12. A blueprint of the GBIF decentralisation strategy

Samy Gaiji<sup>1</sup>, Éamonn O'Tuama<sup>2</sup>, Vishwas Chavan<sup>2</sup>, David Remsen<sup>2</sup> <sup>1</sup>Global Biodiversity Information Facility, <sup>2</sup>Global Biodiversity Information Facility Secretariat

In advancing "from prototype to full operation", the GBIF community recognises the need to move to a more distributed and decentralised model based on the active engagement of more self-sufficient participants nodes.

Such logical evolution of the GBIF network architecture is aimed first at increasing its capacity to rapidly mobilize and share larger amounts of biodiversity-related information, covering not only the existing primary biodiversity data records but also associated information such as metadata, observational, multimedia, names, and other data types. A particular focus will be on further simplifying the process of contributing data by new providers, and on more frequent indexing updates from existing providers.

To achieve this, the GBIF Secretariat has developed a preliminary blueprint of its decentralization strategy, which is aimed at presenting the main components required for such an ambitious evolution of its network architecture. It presents how participant nodes will be progressively empowered to fully participate with the appropriate standards, tools, and training. For example, the new Integrated Provider Toolkit is being developed to enable users to more easily provide occurrence data, taxonomic information, and metadata to GBIF's distributed indexing nodes.

In summary, the ambitions are to significantly expand the scope of content mobilisation and discovery and to couple this with a decentralised model based on customisable tools to better meet the needs of participants. The poster will provide a simple-to-understand, graphically illustrated blueprint of this decentralization strategy.

Support is acknowledged from: Global Biodiversity Information Facility

# 19.13. BioGeobif: Automated Georeferencing Using BioGeomancer

Andrew William Hill<sup>1</sup>, Paul Flemons<sup>2</sup>, Reed Beaman<sup>3</sup>, Ajay Ranipeta<sup>2</sup>, Philip Goldstein<sup>1</sup>,

Robert P Guralnick<sup>1</sup>

<sup>1</sup> University of Colorado, Boulder, <sup>2</sup> Australian Museum, <sup>3</sup> University of Florida

Although a majority of the 141 million occurrence records that are cached by the GBIF portal are georeferenced to allow mapping of occurrence data, the quality and methodology of georeferencing varies widely. Only a small fraction of the occurrence records have georeferenced data produced according to best practices. We are currently working on a GBIF-funded project called BioGeoBIF (combining Biogeomancer and GBIF), which aims to rectify this problem by providing a system to georeference many of those 141 million occurrences with high quality locality data and then providing the georeferenced data back to the original data providers. The system we are building is very simple and when complete will do the following: 1) data providers will be given the opportunity to opt-in to our service; 2) a data harvester will collect provider records through the provider's public access point; 3) unique sets of geographic descriptions will be sent to the BioGeomancer web service (http://www.biogeomancer.org/); 4) the response for each record will be rechecked to determine if the location result is unambiguous; and 5) the final georeferences as well as metadata (e.g., date collected from provider cache, date georeferenced at BioGeomancer, final gazetteer used) will then be made available to the original data providers for download and possible reincorporation into their databases. We are developing several methods for both notifying providers that new georeferencing data are available and helping them assess the quality of their georeferenced data. These assessments will be based on information such as coordinate uncertainty, or more provider-specific preferences, such as the gazetteer from which the georeferences came. A future goal is to work with data providers to help them more fully incorporate georeferences provided through the BioGeobif service back into their databases.

Support is acknowledged from: GBIF

# **19.14.** How to Reduce Your Coffee Consumption – Performance Investigations on Large Occurrence Databases

Jörg Holetschek Botanic Garden & Botanical Museum Berlin-Dahlem

International networks and initiatives such GBIF, BioCASE, and SYNTHESYS share the vision of free and open access to the world's primary biodiversity data, drawn from 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 147 million occurrence records, and the SYNTHESYS cache for European data holds 61 million records (2008-09-02).

This information can be accessed in two ways. The first uses one of the provided web services in order to retrieve particular records or specific taxonomic or regional subsets of all data, and is intended for use by machines, that is, by other projects utilising the data for certain analyses or setting up their own applications. The second method consists of a number of different data portals intended for a wide range of potential human users (see Kelbert, this volume). Apart from usability (especially a user-specific search functionality), short response times are a key requirement for such data portals. Most users are not willing to wait more than a few seconds for search results.

In recent years, the database management system (DBMS) MySQL has been widely used, since it is open source and, due to its slim architecture, fairly fast. However, as the number of occurrence records continues to grow each month, MySQL has been pushed to its limits: its restricted functionality and unsophisticated algorithms increasingly cause problems, regardless of how powerful the server it runs on. With data growth expected to accelerate, this problem must be addressed to avoid serious restrictions on future development.

As is typical with relational databases, GBIF-type occurrence information is split into several tables. However, these datasets are distinguished by the huge discrepancy in the number of records stored by each table – up to several orders of magnitude – and the consequent data/index file sizes. MySQL uses only a single, naively simple algorithm to join together data from different tables, which does not perform satisfactorily for those disproportionate tables, even when custom-tailored query execution plans are provided. Other DBMSs choose from a variety of join algorithms, depending on record numbers, index availability and selectivity of search criteria. They can also process the affected tables in parallel, whereas MySQL does this sequentially, using only one processor per query at a time.

This study compares MySQL to another open source DBMS (Postgres) and to a commercial product (Microsoft SQL Server) in order to find out how they perform for the special characteristics of GBIF-type databases. It looks at the query execution plans chosen by the DBMS and compares actual run times for several typical queries, using a SYNTHESYS occurrence database with 49.7 million records.

Initial results confirm that MySQL's performance is adversely affected by its simple algorithms. Postgres performs better, but requires fine-tuning of its configuration and statistics settings. MS SQL Server shows by far the best results, even without any special tuning, due to its sophisticated queryexecution optimizer and its exhaustive use of the server's memory resources. In contrast, MySQL and Postgres limit memory usage to the values defined in the configuration files.

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

# 19.15. A Pan-European Species-directories Infrastructure (PESI)

Yde de Jong<sup>1</sup>, Charles Hussey<sup>2</sup>, Mark Costello<sup>3</sup>, Thierry Bourgin<sup>4</sup>, Anton Guentsch<sup>5</sup>, Ward Appeltans<sup>6</sup>, Walter Berendsohn<sup>5</sup>

<sup>1</sup> Zoological Museum Amsterdam, <sup>2</sup> Natural History Museum London, <sup>3</sup> Leigh Marine Laboratory, University of Auckland, <sup>4</sup> Muséum national d'histoire naturelle, Paris, <sup>5</sup> Botanic Garden and Botanical Museum Berlin-Dahlem, <sup>6</sup> Vlaams Instituut voor de Zee, Oostende

PESI (<u>http://www.eu-nomen.eu/pesi</u>) is a new initiative, funded by the European Union under the Framework 7 Capacities Work Programme: Research Infrastructures. This three-year project started in May 2008. Led by the University of Amsterdam, it will involve 40 partner organisations from 26 countries.

Progressing from earlier European programs on taxonomic indexing, PESI provides standardised and authoritative taxonomic information by integrating and securing Europe's taxonomically authoritative species name registers and nomenclators (names databases) that underpin the management of biodiversity in Europe.

PESI defines and coordinates strategies to enhance the quality and reliability of European biodiversity information by integrating the infrastructural components of four major community networks on taxonomic indexing into a joint work programme. This will result in functional knowledge networks of taxonomic experts and regional focal points, which will collaborate on the establishment of standardised and authoritative taxonomic (meta-) data. In addition, PESI will coordinate the integration and synchronisation of the European taxonomic information systems into a joint e-infrastructure and the set up of a common user-interface disseminating the pan-European checklists and associated user-services results.

The organisation of national and regional focal point networks as projected, not only assures the efficient access to local expertise, but is also important for the synergistic promotion of taxonomic standards throughout Europe. PESI will be positioned to liaise with national governmental bodies on the implementation of European biodiversity legislation. In addition, PESI will promote the geographic expansion of the European expertise networks to eventually cover the entire Palaearctic biogeographic region.

PESI supports international efforts on the development of a 'Global Names Architecture' by building a common intelligent name-matching device in consultation with the principal initiatives GBIF (Global Biodiversity Information Facility), TDWG, EoL (Encyclopedia of Life), LifeWatch, and EDIT (European Distributed Institute of Taxonomy)). PESI contributes to the development of a unified cross-reference system and promotes the use of high quality taxonomic standards. PESI will further involve the Europe-based nomenclatural services and link the planned joint European taxonomic e-infrastructures to the global e-gateway.

Support is acknowledged from: European Commission

# 19.16. CABIN: Implementing a Biodiversity Information Network in Central Africa

Charles Kahindo<sup>1</sup>, Franck Theeten<sup>2</sup>, Patricia Mergen<sup>2</sup>, Bart Meganck<sup>2</sup>, Kim Jacobsen<sup>2</sup>, Michel Louette<sup>2</sup>

<sup>1</sup> University of Bukavu, RD Congo, <sup>2</sup> Royal Museum for Central Africa

In 2008, the Royal Museum for Central Africa started the CABIN project (Central African Biodiversity Information Network) in the framework of the wider initiative SABIN (Sub-Saharan Biodiversity Information Network). CABIN is supported by the Belgian General Direction of Development Cooperation for a period of 5 years.

Its purpose is to involve scientists from Central Africa in the Information Technology protocols that are currently in force within the scientific community of biodiversity researchers. It contains a capacity building aspect with the installation of databases and web portals that are able to exchange data with the network.

The activities encompassed by this project are:

- 1 The assessment of needs in Central Africa, in terms of access to information about taxonomy and biodiversity on Internet. This task will run during the first two years of the project.
- 2 The identification of specimen collections that may be digitized and later published on the Internet. This task will be done by the means of a call for projects, which will be issued after the second year of the project. A jury will evaluate submitted projects and will decide which are eligible for implementation.
- 3 Capacity Building: The installation of databases and web portals which are able to exchange data with the Global Biodiversity Information Facility (GBIF) network. There is also the need for training staff for the maintenance of this technical infrastructure.

We plan to publish data from partner institutions on our own website and on the GBIF data portal. GBIF already allows access to the biodiversity data from several international institutions. For the transmission of data between the institution providing the data and this thematic portal, we will use software and protocols promoted by TDWG and GBIF.

The project will require the installation of one or two local mirrors of the GBIF portal in Central Africa. The regional indexing of data providers will be undertaken in line with the decentralized architecture of GBIF for contribution to the GBIF network.

CABIN/SABIN will cooperate with the CEDESURK (a documentation centre based at the University of Kinshasa and co-funded by the Walloon and Flemish offices for inter-university co-operation) and Eb@lé (a telecom project promoting education and research in the Democratic Republic of Congo (DRC)) which seek to connect several universities from the Western and Eastern part of the DRC into a common Internet network by using relay-stations for cell-phones as carriers.

More information may be found at:

http://www.africamuseum.be/sitemap/research/zoology/vertebrates/SABIN-CABIN/index\_html.

Support is acknowledged from: Belgian Office for Development Cooperation

# 19.17. Building and Management of the Korean Biodiversity Information System

Sangyong Kim, Dong-Gwang Jo Korea National Arboretum

Since 1998, the Korea National Arboretum (KNA) has been making continuous efforts to build databases and develop applications for the purpose of accumulating and sharing national biodiversity information. In addition, it is attempting to standardize this information by establishing and managing the Korean Plant & Insect Names Index and Standard Classification System. These developments enabled the networking of national biodiversity information, providing research institutes with access to these resources in a systematic way. The outcome of these integration efforts has been the creation of the Korean Biodiversity Information System, integrating and managing information scattered all over the country.

The KNA allocates part of its national budget each year to work with the Korean Biodiversity Information System to develop high quality content for users of Korea's national biodiversity information. Issues of when and how the content are kept up-to-date are managed by allowing the responsible persons at the original information-holding institutes to manage the content themselves in real time.

Activities for database quality control include both scientific content and database structural integrity. Quality control of database content is the responsibility of experts on each taxonomic group, which secure data accuracy by collaborating with pertinent academic societies for inspecting and supervising scientific content. For quality control of the database structural integrity, a yearly quality diagnosis is carried out by authorized external database analysis experts. The results of these exercises are incorporated back into the database in order to improve its content and performance.

As of August 2008, the Korean Biodiversity Information System contained 16 databases with 3.14 million records, including species profiles, specimen records, and living collections data for plants, insects, fungi, and wild animals (birds and mammals).

Support is acknowledged from: Korea National Arboretum

# **19.18.** Information Integration in Documentation of Biodiversity

### Karl H. Lampe

Zoologisches Forschungsmuseum Koenig

Introduction: The practice of documenting biodiversity in natural history museums shares with analogous practices in other museums, archives, and libraries – the so called memory institutions – a traditional focus on entities such as locality, collector, or author. A trend is now underway across cultural and scientific domains in which the focus of documentation is expanding to include processes and events. Thus, the entities what, who, where, and when are now being related to each other through processes and events.

Methods: Biodiversity information is modelled with an ontology based on the International Committee for Documentation [in museums] - Conceptual Reference Model (CIDOC-CRM; ISO 21127). In informatics, ontology is a formal specification of semantic concepts. The CIDOC-CRM is a compact object-oriented model consisting of 84 named classes or entities (E1, E2...), which are interlinked by 137 named properties (P1, P2, etc.). The properties can be understood as verbalized links pointing to other defined entities of the CIDOC-CRM, thus allowing the creation of simple "Subject Predicate Object" relations. Therefore these classes and properties provide a formalized language – a lingua franca – to describe common high-level semantics.

Results: Some realistic examples taken from the field of biodiversity (collecting-, observation and image capture events as well as determination-, type creation- and expedition events) demonstrate that the formal specification of semantic concepts makes scientific activities understandable to a wider audience. Through a formal ontology, information handling and dissemination can be improved by

multiple views, mutual verification and semantic enrichment. In scientific and scholarly disciplines, a formal ontology can give simple object documentation its full scientific depth beyond the limited purposes of collection management.

Documenting processes and events seems to be a prerequisite for interdisciplinary information integration. This documentation is needed for developing knowledge networks and knowledge representation tools on the Internet.

Support is acknowledged from: German Federal Ministry of Education and Research (BMBF)

# **19.19.** Using LSIDs to Link Taxonomic Concepts to Scientific Names for Efficient Data Integration

Nina M. Laurenne, Markus Penttilä, Hannu Saarenmaa Finnish Museum of Natural History

Scientific names of organisms are not always sufficient to unequivocally specify biological objects. In many cases it is necessary to define and identify the taxonomic concept better, and at least denote in what sense or according to whom the name has been used. Resolution of synonyms, misspellings, history of splitting and lumping of taxa, etc., is not yet widely available in the existing taxonomic names servers (TNS). Nearly all TNS show one preferred taxonomy and nomenclature, and are ill-equipped to display alternative ones. These problems are aggravated when integrating multiple occurrence databases that use conflicting taxonomic concepts. There is a growing need for a database that can manage several species lists with differing nomenclatures.

To overcome this problem, we are developing a database that links alternative taxonomies and nomenclatures between checklists using taxonomic concepts. The names and concepts are kept as separate objects. The database, a beta version of which is available at <a href="http://www.luomus.fi/taxondev/shows">http://www.luomus.fi/taxondev/shows</a> not only valid names with their synonyms and taxonomic rank, but also the taxonomic concept and its relationship to other concepts, and the names used for these. Also, the species content of the higher taxa can be viewed and compared with each other. New checklists can be uploaded into the database in bulk operations, and the names therein are provisionally connected to existing concepts in cases when the name and author strings fully match. Otherwise a new concept is created. Authorised expert users can edit the relationships, such as whether a particular taxon in one publication or checklist is a part of part of another taxon in another publication or checklist, includes another taxon and/or overlaps with yet another taxon.

Life Science Identifiers (LSID) are used to uniquely identify the taxonomic concepts. LSID are obtained from the Catalogue of Life's Annual Checklist when available, and for the remaining taxa the LSIDs are generated using the LSID service of the Finnish Museum of Natural History.

So far, five checklists of North-European Lepidoptera compiled from 1962 to 2008 have been uploaded to the database. Linking of the taxa across these lists has been done for the superfamilies Hesperioidea and Papilionoidea. The database will be expanded to cover other organism groups. The lists can be downloaded in various formats, and a TAPIR interface for Taxonomic Concept Transfer Schema (TCS, <u>http://www.tdwg.org/standards/117/</u>) data will be added. Common names are added to meet the needs of other clients using taxonomic information.

The next step, which is part of our project in the Nordic region, is to introduce the LSIDs to biodiversity databases to denote the taxonomic concepts of specimens and observations. For instance, all specimens and observations of the butterfly Leptidea sinapis (Linnaeus, 1758) sec. Linnaeus, 1758, collected before the year 1990, when the species was split, will get a different LSID than Leptidea sinapis (Linnaeus, 1758) sec. Reissinger, 1990, which are being observed today. When the LSID markup of taxa is widely implemented in occurrence databases, data can be better integrated than simply using scientific names alone.

Support is acknowledged from: NordForsk

# 19.20. The Collaborative Ontology Creation Process Within The ALTER-Net Community For the Socio-Ecological Research and Observation oNTOlogy (SERONTO)

Barbara Magagna, Johannes Peterseil, Herbert Schentz Umweltbundesamt GmbH

This poster describes the ongoing development process of SERONTO within the ALTER-Net community. ALTER-Net is a network of excellence for Long-Term Biodiversity, Ecosystem and Awareness Research spanning twenty-four institutions in seven European countries.

According to Gruber (1993), an ontology is a formal specification of a shared conceptualization of a domain of interest. "Shared" means a common understanding of the knowledge to be formalised; a foundation to which domain experts can map their own specific views. "People can't share knowledge if they don't speak a common language" (Davenport, 1997). The real challenge for developing ontologies lies in finding this common language and understanding. This seems to be the essential requirement for acceptance and reuse of the ontology for sharing data and knowledge within a community.

The efficiency in the different development phases must consider the social dynamics within the development group. This is a well founded and proven methodology for the creation process. Five potentially distinct phases were identified:

- 1 Identify the scope and distribute the tasks to working groups with different focuses (core and domain ontologies)
- 2 Structure the information space ordered by relevance and resulting in loose concepts
- 3 Create a conceptual model with a derivation hierarchy and a relation structure
- 4 Create a formal model including restrictions and rules
- 5 Create examples, sometimes by importing instances from existing databases

For each of these phases, different working environments, tools, methods and group sizes seem to be appropriate. Face to face meetings of a week appear to be the most efficient way to advance the ontology work because the whole group can focus its attention solely on this task. Nevertheless, not all problems can be solved by such a group. Issues arising there must be described in ways that make it easy for the group to reach decisions and fed back into a resolution process. Usually this is done by one person who is responsible for presenting outstanding issues back to the group. A Wiki is a collection of web pages that can be edited by all members of a group. A Wiki an ideal platform for exchanging viewpoints and discussing issues around a developing ontology. In our decision process, working periods with deadlines are defined to address each issue, and every member of the entire group is given the opportunity to vote for or against each proposed solution.

The size of the group required depends on the scope of the problem. Our experience shows that small working groups (up to 4 persons) are most efficient after phase one. The later phases demand more advanced skills in ontology engineering, for example in the formalization of concepts within ontology editors. This implies a differentiation of roles within the group. The process facilitator takes on the tasks of maintaining an overview of the process and extracting group opinions while keeping a neutral position during the meetings. The facilitator also has the task of supervising the ontology work in the WIKI.

Support is acknowledged from: ALTER-Net, a Network of Excellence funded by the EU within its 6th Framework Programme

# **19.21.** Kansas State University's Biodiversity Information System: Combining institutional expertise to enhance access to collections information

#### Mark Mayfield, Carolyn Ferguson, Gregory Zolnerowich, David Allen, Harish Maringanti, Mike Haddock, Jenny Oleen Kansas State University

The Biodiversity Information System (BiodIS) of Kansas State University is a collaborative project between the institution's natural history collections (herbarium and entomological) and libraries. It was conceived to provide tools to expose biodiversity data housed within the University. Established tools and standards for data acquisition and sharing are fundamental components of the system, and are critical for the success of BiodIS. We utilize Specify 5.2.3 (www.specifysoftware.org) to manage and enter data. Currently, the core web-based collections data are presented as a web portal (http://biodis.ksu.edu/collections) developed by the commercially available SilverCollection software (www.silverbiology.com). The collections data are mapped to a Darwin Core framework and served via TAPIR (TDWG Access Protocol for Information Retrieval) using TapirLink 1.0. The BiodIS website access will eventually combine both basic and advanced user interfaces, each tailored to the needs of these different communities. It will also encourage the engagement of these user communities in website development through collaborative tools such as wikis for sharing scientific and historical information relating to the specimens as well as for improving website functionality. Core resources focus on organisms of particular importance in the Great Plains region of the central United States. Target taxa include economically and ecologically important groups such as grasses and crop weeds and pests. Digital print and image resources are being developed to make unique local publications widely available. In addition to presenting the BiodIS resource in its current state, we briefly discuss its development, noting the advantages of partnering with diverse groups within our institution.

Support is acknowledged from: National Science Foundation, Kansas State University Targeted Excellence, Kansas NSF-EPSCoR

# 19.22. The roots and branches of the GeoPlatform tree.

Bart Meganck<sup>1</sup>, Patricia Mergen<sup>1</sup>, Franck Theeten<sup>1</sup>, Pere Roca Ristol<sup>2</sup> <sup>1</sup> Royal Museum for Central Africa, <sup>2</sup> CSIC

The EDIT WP5.4 GeoPlatform is being developed as an online, user-friendly tool for taxonomists: A website to fire up first thing in the morning, ready for use when you need it - unobtrusive and reliable. However, GeoPlatform suffers from a lack of "splendid isolation" - linked as it is to a host of contributing ideas and derivative works. We will try to untangle the roots winding up to, and the branches leading out of this tool: Both the solid, for-certain, near-future results and the slender, tentative offshoots.

As roots, we include the work done at the TDWG BioGeoSDI workshop in Campinas, 2007, where crucial insights were gained in stringing together online biodiversity tools into a full "à la carte" workflow. From SYNTHESYS' interoperability work (Itinerary Tool, Generic Query Tool) come the technologies for importing Global Biodiversity Information Facility (GBIF) data and a broad range of file formats. The Worldwide Web Consortium (W3C) technologies like Web Map Server (WMS) and Web Feature Server (WFS) provide the bread and butter for easy interoperability. Finally, the EDIT ATBI inventories provide a grand source of real-life data for testing and fine-tuning.

As branches, there is the firm trunk of the GeoPlatform itself, with its many functions that are still growing. Furthermore, there is the molecular and phylogenetic branch, adding all things DNA to the visualisation - showing that molecular technologies are an addition, not a threat to taxonomy. We see a fruit-laden branch of dissemination through the EDIT summer courses. Then there is the timid, airy branch of 3D promises.

We hope that through this presentation the importance of standards adherence and close collaboration will become apparent, as well as the role of transverse thinking and having a broad view on biodiversity-related projects. The talk should also provide other people with hooks for their own projects and ideas through demonstrating the slow but steady process from fragile brain seed to complexly branched technology tree.

Support is acknowledged from: SYNTHESYS, EDIT, TDWG and GBIF

# 19.23. The roots and branches of the EDIT GeoPlatform tree

Bart Meganck<sup>1</sup>, Pere Roca Ristol<sup>2</sup>, Franck Theeten<sup>1</sup>, Patricia Mergen<sup>1</sup>, Garin Cael<sup>1</sup>, Kim Jacobsen<sup>1</sup> <sup>1</sup> Royal Museum for Central Africa, <sup>2</sup> CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS

In the framework of the EDIT (European Distributed Institute for Taxonomy) Work Package 5 (WP5), a collaborative online working environment called Platform for Cybertaxonomy is being developed, which includes among others, geospatial components referred to as the "Geoplatform".

The development and implementation of the Geoplatform and its modules, has not been in isolation, but is a result of synergies with other GIS-related initiatives. This is seen to guarantee interoperability, conformity to internationally recommended standards (like those of the W3C and Open Geospatial Consortium) as well as making use of already existing components rather than developing new modules from scratch. All this keeps in mind the goal of efficiency, user-friendliness, and matching end-users needs.

The work accomplished at the TDWG Geointeroperability Workshop in Campinas, Brazil in 2007, represented a major step toward stringing together various online biodiversity tools in a full "à la carte" workflow. Other related developments to be considered as add-ons to the EDIT geoplatform and part of this presentation are the Itinerary and Generic Query Tools implemented in the framework of the European Union's SYNTHESYS project. These query tools aim to mobilize the huge amount of information in commonly used formats, which are available in networks such as GBIF.

Recent information collected during the EDIT All Taxa Biodiversity Inventories provides a marvelous source for testing and fine-tuning the Geospatial platform. Tests have also been conducted using 3D technologies, to relate phylogenic trees generated from DNA analysis to the sampling location of the specimens.

The aim of this presentation is to show how a strong foundation of roots based on the adherence to standards and close collaboration among related biodiversity and geospatial information projects can promote the branches that allow this project to reach the common goals of interoperability and meeting end-users needs.

Support is acknowledged from: the EU projects SYNTHESYS and EDIT, TDWG, GBIF

# **19.24.** RDF based Reference Network Architecture for a distributed Digital Library system, The Royal Museum for Central Africa's use case on African Bird information (STERNA project)

Patricia Mergen, Hein Vanhee, Agnes Lacaille, Garin Cael, Michel Louette Royal Museum for Central Africa

The STERNA European project funded by the eContentplus program in the topic of Digital libraries stands for Semantic Thematic European Reference Network Architecture. The system is based on a previous project called RNA for Reference Network Architecture which objectives were to set up a distributed knowledge management system based on RDF (Resources Description Framework) and SKOS (Simple Knowledge Management Structure), both Worldwide Web Consortium (W3C) standards (http://www.rnaproject.org/).

The STERNA project builds on the results of the RNA projects, and is meant to achieve a proof of concept by using the broad thematic of bird-related information. The elements range from specimen and occurrence data on birds with distribution maps and habitat information, to sound recordings, photographs, ethnographic objects and paintings.

The Royal Museum for Central Africa's (RMCA) play multiple roles in the project. As work package leader of "Technology improvement, target user validation, evaluation of the STERNA approach", we will try to identify the target end users, perform a user need survey, do user validation tests of the system, and test interoperability with internationally recognized standards (among which TDWG standards) and then suggest technological and functional enhancement of the existing system.

The RMCA will also provide its unique African bird specimen information as a Use Case Scenario for the project. The collection has approximately 145 000 specimens of which 78% are digitized and 73% georeferenced, including 987 type specimens of 226 nominal species and about 500 high quality digital images of these type specimens. Through collaboration between the departments of Zoology and Cultural Anthropology, the RMCA will also provide information on selected cultural objects either representing birds or composed partially of feathers or other bird parts. In doing so, our intention will be to test the anthropology extension to the TDWG standard ABCD, presented by Charles Copp at the 2007 TDWG annual meeting.

The members of the Sterna Consortium are Salzburg Research Forschungsgesellschaft m.b.H (Project Leader), Archipelagos, DOPPS Birdlife Slovenia, Heritage Malta, Hungarian Natural History Museum, Icelandic Institute of Natural History, Natural History Museum / Municipality of Amaroussion, Natural History Museum of Luxembourg, Naturalis, Netherlands Institute of Sound and Vision, Royal Museum for Central Africa, Teylers Museum, Wildscreen/ARKive. The (OpenSource) software is being developed by the company Trezorix (<u>http://www.trezorix.nl</u>). The Sterna Project is funded by the EU eContentplus Program of the DG Information Society & Media in the topic Digital Libraries (<u>http://ec.europa.eu/information\_society/activities/econtentplus/index\_en.htm</u>). The digital images of the type specimens are financed by a project of the Federal Belgian Science Policy Office.

Support is acknowledged from: The Sterna Project is funded by the EU eContentPlus Program of the DG Information Society & Media in the topic Digital Libraries (<u>http://ec.europa.eu/information\_society/activities/econtentplus/index\_en.htm</u>).

# 19.25. tropicos.org - A New Look

#### Chuck Miller, Chris Freeland, Jay Page Missouri Botanical Garden

Tropicos has been the Missouri Botanical Garden's primary supporting data tool for botanical taxonomic research since 1982. Over 1000 person-years of plant taxonomic research have been recorded in Tropicos, resulting in a repository of one million plant names with synonymy, type protologues, distributions, references and almost four million specimen records cross-linked to the taxa records.

The Internet face of Tropicos is tropicos.org, which provides open worldwide read-only access to the data contained within the Tropicos systems, including integrated links to the botanicus.org repository of digitized botanical reference literature. A completely revised graphical user interface was launched in April 2008 that is much richer and simpler to use than the veteran W3Tropicos. The new graphical user interface utilizes tabs to categorize and provide easy access to the information. High-resolution digitized specimen images are now displayed via enhanced JPG2000 real-time decompression techniques.

tropicos.org has been developed as a web application employing an n-tier object-oriented architecture that separates the database, application, and user interface layers, making the application easier to maintain and enhance. Asynchronous Javascript and XML (AJAX) techniques have been used to create a responsive and intuitive browser-based user interface.

The primary software development components of tropicos.org are:

- o Microsoft Visual Studio 2008 development environment,
- Microsoft SQL Server 2005 database management system,
- o Microsoft ASP.Net web application layer,
- o Microsoft .NET Framework 2.0 foundation layer, and
- AJAX user interface layer.

Support is acknowledged from: The Andrew W. Mellon Foundation

# 19.26. Remote Annotation of Specimens: Use cases and messages of a Filtered Push network

Paul J Morris<sup>1</sup>, Hui Dong<sup>2</sup>, Maureen Kelly<sup>3</sup>, James A Macklin<sup>3</sup>, Robert A Morris<sup>2</sup>, Zhimin Wang<sup>2</sup>

<sup>1</sup> Harvard University Herbaria/Museum of Comparative Zoology, <sup>2</sup> Department of Computer Science, University of Massachusetts, Boston, <sup>3</sup> Harvard University Herbaria

Global collection database networks bring specimen data to the specialists best able to correct and clean those data, but lack easy means to return annotations back to data sources; a radical departure from annotation of labels by visitors to collections. We are designing and implementing a prototype network, which we term "Filtered Push". This network transports structured annotations from remote sites to the authoritative databases of the collections holding the vouchers to which those annotations apply. The name reflects function. 'Push' means thatannotations can be pushed from remote corners of the network to authoritative data sets. 'Filtered' means that data curators can accept or reject annotations. For these functions, TDWG standards and an open architecture are essential.

From expressed community needs, we have developed two use cases for a Filtered Push network, 'Find Duplicates' and 'Make Annotation.' In both cases, people interact with their own software systems. These systems in turn interact with the Filtered Push network through an open client API (Application Programming Interface) This API defines messages injected into the network and hides internal details of implementation.

The 'Find Duplicates' use case addresses the redundant data capture of distributed duplicate herbarium specimens. In this use case, a collection manager begins capture of a herbarium sheet by entering the collector name and collector number. If duplicates of this herbarium sheet have already been captured at one or more institutions, those data are presented to the collection manager who can import it rather than continuing the entry. Knowledge of duplication is returned to the network and to the other herbaria holding members of the set. Requirements from this use case include the return of data from the network faster than typing by hand and retrieval of structured data mapped onto local concepts.

In 'Make Annotation', a taxonomist using a remote data portal examines data on the vouchered occurrence of some organism (images and label data of a museum specimen) and adds information such as a new determination. The portal injects this new information in structured form as an annotation message through the client API to the Filtered Push network. The annotation is delivered to the appropriate collection for review, to be either discarded or accepted into the authoritative collection database. Other network users can discover the specimen data and the annotation, even if the annotation is pending and awaiting review. Requirements include mapping structured data at network boundaries onto local schemas (facilitated by standards), filtering annotations, network knowledge of pending annotations, and an open architecture. 'Make Annotation' implies typed client API messages for the injection of different annotation of different kinds. A new determination needs semantic elements for taxon name, determiner, and the referenced specimen. A correction corrects an arbitrary concept of a referenced specimen. Annotation messages induce response messages informing the network of the acceptance, agreement with, or rejection of an annotation.

We will discuss the open architecture we are developing from these use cases, including the client API, visibility of pending annotations, and filtering. We will separately discuss a prototype network architecture based on Apache Hadoop MapReduce. The project is being openly documented at (http://mantis.cs.umb.edu/wiki/index.php/Filtered\_Push).

Support is acknowledged from: U.S. National Science Foundation DBI:0646266

# **19.27.** Standardizing Mansfeld's World Database of Agricultural and Horticultural Crops by implementing a concept-based data model

Ram Narang, Helmut Knuepffer Leibniz Inst. of Plant Genetics & Crop Plant Res.

Mansfeld's World Database of Agricultural and Horticultural Crops (<u>http://mansfeld.ipk-gatersleben.de/</u>) was developed at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) in Gatersleben, Germany, concurrently with the preparation of "Mansfeld's Encyclopedia" (Hanelt & IPK 2001). It is an online accessible taxonomic database of agricultural and horticultural crop plant species with worldwide coverage. For each of the ca. 6,100 plant species cultivated for purposes other than ornamental or forestry, the database provides information on taxonomy and nomenclature with numerous synonyms, as well as common names in different languages. The factual information provided for each cultivated species includes information items such as the regions of cultivation, distribution in the wild, plant uses, notes on the phylogeny, ancestral species, domestication, variation, history, images, and references.

Originally developed in Microsoft Visual FoxPro, the Mansfeld Database was recently migrated to the Oracle 10g database management system. In a second step, the structure of the database was changed to become compatible with the "Berlin Model" (Berendsohn et al. 2003), a concept-based taxonomic data model. The new implementation of the Mansfeld Database uses the Berlin/IOPI (International Organization for Plant Information) taxonomy core module with a few modifications, to manage and organize information about taxon names and the concepts they represent. Various information items (*e.g.*, scientific names, botanical references) were atomised and transformed into the concept-based database model. The transfer to such a widely-used concept-based data model led to better standardisation and quality improvement of the taxonomic information in the Mansfeld database by increasing its accuracy, resolution, and interpretability. The factual information, *e.g.*, textual information on geography, uses, domestication and history, is linked to the potential taxon section of the Mansfeld Database. The adoption of the "Berlin Model" will help in integrating the Mansfeld Database into global biodiversity information networks and sharing the information with other systems.

Using a suitable web editor designed for concept-oriented taxonomic databases that use the Berlin Model, it will be possible to update and revise the contents of the Mansfeld Database, with the aim to incorporate new knowledge about cultivated plant species and to eventually produce a new edition of the published book.

References:

- Berendsohn, W.G. et al. (2003) The Berlin Model: a concept-based taxonomic information model. In: W.G. Berendsohn (Ed.), MoReTax. Handling Factual Information Linked to Taxonomic Concepts in Biology. Schriftenreihe für Vegetationskunde 39, 15-42.
- Hanelt, P. & Institute of Plant Genetics and Crop Plant Research (Eds.) (2001) Mansfeld's Encyclopedia of Agricultural and Horticultural Crops. Vols. 1-6: 3716 pp. Springer, Berlin, Heidelberg, New York, etc.

# **19.28.** Unlocking Biodiversity Data: Semantic Cyberinfrastructure for Investigation and Discovery (SCID)

Federico Ocampo, Mary Liz Jameson

Scientific discovery plays an increasingly essential role in addressing societal challenges related to the environment, human health, and the economy. Advances in biology are particularly key to developing solutions to these challenges. Three broad research fields account for the majority of our new biological knowledge that will allow us to develop solutions: evolutionary biodiversity science, genomics and proteomics, and bioinformatics. Of these, bioinformatics is the glue that serves to bind these research efforts and will create entirely new fields of research as it integrates genomics-related sciences and evolutionary biodiversity sciences into earth sciences, chemistry, and other disciplines. Without question, scientific inquiry has made unprecedented progress due largely to advances in computer engineering and overwhelming abundance of high-quality biological data. Today's investigators are able to capitalize on advances in physical and computer-based data acquisition methods and networked, electronic data storage – advances that yield exponential increases in the amount of data available for investigation and discovery. Life scientists have benefited from access to data on whole genomes of many organisms, global gene expression sets, cellular metabolic profiles, and global biodiversity data.

Unfortunately, in many disciplines and sub-areas, investigators remain hard-pressed to fully exploit available data due to heterogeneity, large scale, and decentralization. Our project will unify the fields of genomics and proteomics, evolutionary biodiversity science, and bioinformatics by federating and unifying data sources in biodiversity and genomics, unlocking heterogeneous, large-scale databases, and empowering investigation and discovery. This provides the added benefits of unifying diverse data and allowing for broader scientific investigations (*e.g.*, biodiversity studies could leverage data from climate, geology, and even human history).

The long term vision of our research is to bridge the data of several diverse biological fields in order to facilitate scientific investigation in the biological sciences. In accordance with our vision, we will pursue three overarching research goals: 1) unite several separate biological areas by federating their heterogeneous databases; 2) support scientists in the efficient and effective use of ontologies to enhance investigation and discovery; and 3) streamline the use of bioinformatics tools via common user interfaces for querying data, enhancing data analyses, and interpreting scientific results. We provide an overview of the project, our mapping prototype, and our long-term vision.

Support is acknowledged from: U.S. National Science Foundation, CONICET (Argentina)

# 19.29. The GBIF Work Programme 2009-2010: From Prototype towards Full Operation

Éamonn Ó Tuama<sup>1</sup>, Vishwas Chavan<sup>1</sup>, Samy Gaiji<sup>2</sup>, Alberto Gonzalez-Talaván<sup>1</sup>, Nicholas King<sup>1</sup>, Tim Robertson<sup>1</sup>

<sup>1</sup> Global Biodiversity Information Facility Secretariat, <sup>2</sup> Global Biodiversity Information Facility

In advancing "from prototype to full operation", GBIF recognises the need to move to a more distributed and decentralised model based on the active engagement of more self-sufficient participants, thereby enabling scalability and diversification beyond the relatively centralised infrastructure achievable through the limited capacity of the Secretariat and a low number of fully functional contributing members. In a significant change in emphasis, the GBIF Work Programme for 2009-2010 thus focuses on enabling an increased level of activity and involvement by participants to help make operational a greatly expanded and successful network. This will be achieved by building participant capacity for uptake and implementation, *e.g.*, through expanded mentoring and training, and through developing enhanced technologies for participant access and connectivity. In addition, the Work Programme provides for the inclusion of new and wider biodiversity data and

information types (*e.g.*, observational and multimedia) and analyses, and for increases in data volumes.

The 2009-2010 Work Programme centres around the two thematic areas of Informatics and Participation. Both are closely aligned, with Participation undertaking the key role of conveying to Informatics the needs of participants for specific products and services, and, once developed, channeling them back to participants, in a continual process of improvement. GBIF can thus play a core central role in Biodiversity Information Standards (TDWG), by conveying the informatics requirements and experiences of the broader biodiversity community, thereby helping to set priorities around standards and interoperability issues for "real world" practitioners.

The main Informatics and Participation activities of the Work Programme, outlined here, are designed to allow GBIF to realise its role as a global mega-science initiative. Acting as a catalyst for change, GBIF will draw on the experience of its more advanced members and will seek to build capacity in all participants, to develop customisable outcomes based on local needs, but also to commit to sharing biodiversity data as a "common good" that knows no boundaries, recognising that working together to create a successful network is for the benefit of all.

# 19.30. GBIF's Global Biodiversity Resources Discovery System

Éamonn Ó Tuama<sup>1</sup>, Vishwas Chavan<sup>1</sup>, Samy Gaiji<sup>2</sup>, Tim Robertson<sup>1</sup> <sup>1</sup>Global Biodiversity Information Facility Secretariat, <sup>2</sup>Global Biodiversity Information Facility

The core informatics challenge for the Global Biodiversity Information Facility (GBIF) as it moves beyond its prototype phase to full realisation of a successful, global biodiversity network is to design and implement a scalable, decentralised system that adheres to international standards for data exchange formats and protocols, thereby facilitating the maximum degree of interoperability across heterogeneous, distributed data holdings and applications. One fundamental requirement that follows from the need for a decentralised infrastructure is to provide users with ways that substantially increase their ability to discover and access relevant biodiversity datasets no matter where they occur across the network. This, in turn, requires that data providers document their datasets as fully as possible, conforming with international metadata standards.

Users of the GBIF system include both providers and consumers of data (often the same individuals assuming different roles). Such users undertake varying activities relating to inventory, discovery, and access to data. Data providers are key actors in that they must prepare an inventory (or catalogue) of their biodiversity resources in the form of a metadata catalogue (with one metadata document to describe each resource, such as occurrence datasets, taxonomic data, etc.) and make their data available, ideally, via standards-based web services. The metadata catalogue can then be shared through indexing in a central metadata repository, thus allowing unified searching across multiple catalogues and networks.

The related functions of inventory, discovery, and access are being brought together by GBIF through its Global Biodiversity Resources Discovery System (GBRDS), at the heart of which lies a metadata cataloguing system. The GBRDS will form a key component of GBIF's decentralised architecture by providing a unified entry point, at a global level, for discovery of all kinds of biodiversity resources, both digitised and undigitised, including observational datasets, collections, names data, standards and services, and for integrating the GBIF network in other systems such as the Global Earth Observation System of Systems (GEOSS). This poster provides an overview of the internal components and interfaces of the GBRDS and explains its central importance for a global biodiversity network.

# **19.31.** The Encyclopedia of Life: Status report on species pages, contributions, and curators

### Cynthia Parr

Smithsonian Institution

The Encyclopedia of Life (<u>http://www.eol.org</u>) aims to provide high-quality data about species to scientists and the general public worldwide from a single portal. Information about the 1.8 million known, extant species will be assembled in ten years via several routes: existing data providers can agree to share some or all of their information with us, scientists will work to bring new information online, and the general public will be invited to contribute their best information, which is then subject to curatorial review. Below we discuss our progress on each of these routes.

EOL spent the first year and half building infrastructure to serve content from existing data providers. Our February 2008 launch included content from FishBase, AmphibiaWeb, Tree of Life, and Solanaceae Source in addition to a small number of exemplar pages and more than a million stub pages for names in the Catalog of Life. In early September, content came online from Animal Diversity Web, AntWeb, ARKives, and others to reach about 38,000 taxa with text and another 15,000 with no text but images from several sources. At least 40,000 (perhaps up to 150,000) additional text pages will be released in December. Original literature from the Biodiversity Heritage Library is linked to several hundred thousand species pages. To accelerate connections with existing resources, a registration process now allows prospective data partners to establish their own affiliation with EOL. Providers map their schemas to the EOL Transfer Schema, which uses TDWG standards such as the Species Profile Model.

To support the participatory phase of EOL development, we provide ways for the scientific community and the general public to bring new content online. EOL's Biodiversity Synthesis Center and the Species Pages Group have sponsored workshops for taxonomic communities to coordinate content-building – *e.g.*, the Decapods, ants, fungi, Diptera, and others. The EOL Fellows program, piloting this year at the Smithsonian Institution, supports scientists to develop content and mobilize their communities. Several regions have stepped forward to coordinate geographically-relevant pages.

A variety of tools and features that enable EOL participation are coming online. LifeDesk: Expert is a Drupal-based content management environment, modeled after and compatible with EDIT Scratchpads, which scientists can use to assemble and manage information useful to their own communities and to EOL. Later LifeDesk versions will support educational and citizen science audiences. In December 2008, we will launch commenting and tagging features. Flickr (http://flickr.com/) has been chosen as one way for users to contribute images to EOL.

Once the public begins contributing content, a robust curatorial network is necessary to decide which contributions are suitable for authoritative pages. In December 2008, we will begin testing limited curation functionality, phasing in the full system in 2009. An expert can register as a curator and offer publicly viewable credentials. Because multiple curators may volunteer for any given area of the EOL, we will appoint master curators, with the help of scientific societies, to help resolve conflicts. We are designing the process and our policies to be as open and inclusive as possible, while still ensuring quality. We seek input on these policies, and assistance in recruiting curators and in refining ways to credit curators and contributors for their effort.

Support is acknowledged from: MacArthur Foundation, Sloan Foundation

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

#### David P Remsen GBIF

Taxon names are a component of almost all information pertaining to species and, as a consequence, are present as text annotations within most digitized content. Biodiversity content spans an enormous range of sub-domains within the life sciences and is present at all organizational levels of biological information from molecular and cellular data through organism and ecosystem information. The ubiquity and importance of taxon names implies a role as facilitators for linking these data to our understanding of biodiversity. Enabling such a linkage, however, is hampered by a number of significant impediments inherent to taxon names. These impediments fall within at least three major conceptual classes (or properties), which must be collectively addressed to facilitate discovery, access, and ultimately integration, of biodiversity information.

- 1. At its most fundamental, a taxon name is a string of characters within a computerized system and one aspect of a cyberinfrastructure for names must address impediments inherent to the string properties of names.
- 2. Taxon names are nomenclatural units linked to an originating nomenclatural act. Names have properties governing their structure that provide the nomenclatural basis for relationships among different taxon names. Such relationships provide one rationale for how different names can be related, or conversely, how the same name can be distinct.
- 3. Taxon names are labels for taxon concepts. Taxon concepts are defined by a competing number of subjective criteria that can result in varying circumscriptions (or definitions) of a taxon labelled with the same name. Taxon concepts complicate the use of taxon names as the basis for a cyberinfrastructure that facilitates integration of content. Taxon concepts, however, also contribute useful components (*e.g.*, classification, synonyms) to a cyberinfrastructure that facilitates discovery and access to content.

The Global Names Architecture is a simple idea based on the use of indexing and cross-referencing of information resources, which may contribute information that addresses these impediments. Such an infrastructure, in concert with the use of TDWG-governed exchange standards and protocols, provides the foundation for comprehensively facilitating discovery and access to biodiversity information. Such a foundation, when combined with appropriately defined globally unique identifiers, will enable this infrastructure to contribute to the integration of globally distributed species information in a taxonomically meaningful way.

Support is acknowledged from: Global Biodiversity Information Facility

# **19.33.** Ten Years of FloraBase

Ben Richardson<sup>1</sup>, Alex Chapman<sup>2</sup>

<sup>1</sup> WA Herbarium - Department of Environment and Conservation, <sup>2</sup> WA Herbarium - Department of Conservation and Land Management

FloraBase 1 was published online in November 1998 as a set of database-enabled CGI scripts and supplementary HTML pages. It was revised substantially as version 2 in July 2003 as a PHP web application, and again as version 2.5 in September 2007 with further enhancements.

We started with a small web site containing a few critical features. This was driven partly by a lack of resources, which we now think was useful. In the first phase of the site's life, when we were learning what our users found most valuable and what we were able to build with the available resources, we avoided implementing features that were of little interest. Only gradually have we added features as our needs and those of our users matured.

We also made decisions that in hindsight we should have done differently. We chose early on to create single photographs as a collage. The collage contained up to 5 images with a title and the name

of each photographer. This made updating the collage more difficult because changes often required a complete overhaul to the collage, particularly when one photograph needed to be replaced with another with markedly different dimensions. Also, our users became interested in using the photographs for their own projects and the collage made it more difficult to determine the ownership for each photograph.

In future versions we plan to continue to make careful changes to support our users. An area of interest to TDWG members will be data downloads. Supplying data formatted according to the applicable TDWG standards make obvious sense to us, but our users have not yet requested data in these formats.

Support is acknowledged from: Department of Environment and Conservation

# 19.34. EDIT mapViewer: a Gegraphic Web Application for Taxonomists

Pere Roca Ristol<sup>1</sup>, Pablo Sastre<sup>1</sup>, Jorge Lobo<sup>1</sup>, Franck Theeten<sup>2</sup>, Bart Meganck<sup>2</sup>, Patricia Mergen<sup>2</sup> <sup>1</sup>CSIC, <sup>2</sup> Royal Museum for Central Africa

Geographic information is almost always an important part of biological data. An easy-to-use webapplication able to manage georeferenced data is currently being developed within the EDIT project (European Distributed Institute of Taxonomy), providing taxonomists with a tool for their daily work: uploading data, querying, choosing symbols to represent data points, and creating maps by adding remote or local geographic data.

Although 72dpi resolution, suitable for viewing on a computer screen is now available, providing a higher resolution (printing quality) image from the maps is a priority in the project. We also expect to be able to harvest and display GBIF data and provide spatial analysis tools that allow users to assess whether or not the inventory is complete enough for use: Do I have enough data? Which spatial units may not yet be thoroughly surveyed?

EDIT mapViewer: http://edit.csic.es/edit\_geo/prototype/edit.html

Support is acknowledged from: EDIT

# 19.35. e-Biosphere 09: International Conference on Biodiversity Informatics

David E. Schindel, Cynthia Parr Smithsonian Institution

Biodiversity Informatics (BI) has experienced extraordinary growth as an international enterprise over the past decade. Data standards, databases, and networks of interoperating people, data, and software tools have appeared, grown, and become integrated to varying degrees. Information resources have grown faster than their user communities, but important applications of and appreciation for these resources are growing rapidly.

The Encyclopedia of Life (EOL; <u>www.eol.org</u>), the Consortium for the Barcode of Life (CBOL; <u>www.barcoding.si.edu</u>), and approximately ten other international initiatives in biodiversity informatics will be sponsoring "e-Biosphere 09: The International Conference on Biodiversity Informatics" in London on 1-5 June 2009 (see Conference website, <u>www.e-biosphere09.org</u>). The goal of the conference will be to highlight the progress made by BI over the past decade, and to begin the process of developing a more comprehensive, integrative roadmap for BI's future directions. The first three days will be an open International Conference at the Queen Elizabeth II Conference Centre, during which the accomplishments, applications, challenges, and future directions of BI will be discussed. Two sessions during the International Conference have been set aside as free time, during which conference participants can visit demonstrations, poster presentations, and commercial exhibits

in the e-Biosphere "Street Fair". The final two days will be a planning meeting at the Natural History Museum (London) involving invited representatives of diverse initiatives in BI.

Members of the TDWG community are invited to attend e-Biosphere 09, and to participate in the following special ways (see <u>http://www.e-biosphere09.org</u> for application forms and details):

- As software/database demonstrators: Demonstration booths on the Street Fair will be available for hire, in which developers can demonstrate not-for-profit software tools and databases;
- As commercial exhibitors: Commercial database companies, publishers, and other organizations can hire exhibit space on the Street Fair to present their products and services to conference attendees;
- As poster presenters: Researchers and students can submit poster abstracts for consideration by the Programme Committee. Accepted abstracts will be included in the Conference Proceedings and can be displayed as posters throughout the conference; and
- As organizers of side-events: During the final afternoon of the Conference, space is available for special sessions that can be organized by BI initiatives, including TDWG participants.

Support is acknowledged from: Encyclopedia of Life (EOL); Consortium for the Barcode of Life (CBOL); Global Biodiversity Information Facility (GBIF); JRS Biodiversity Foundation; Andrew W. Mellon Foundation; Ocean Biogeographic Information System (OBIS); European Distributed Institute of Taxonomy (EDIT); Darwin200; LifeWatch; BioNET-INTERNATIONAL; Inter-American Biodiversity Information Network (IABIN); Atlas of Living Australia (ALA); and the Pan-European Species-directories Infrastructure (PESI).

# 19.36. Development of the Generation Change Program Domain Models and Ontology Management Embedded in the GCP Platform

Rosemary Shrestha<sup>1</sup>, Jeffrey Detraas<sup>2</sup>, Thomas Metz<sup>2</sup>, Martin Senger<sup>2</sup>, Graham McLaren<sup>2</sup>, Elizabeth Arnaud<sup>3</sup>, Tom Hazekamp<sup>3</sup>, Adriana Alercia<sup>3</sup>, Guy Davenport<sup>1</sup>, Reinhard Simon<sup>4</sup>, Jayashree Balaji<sup>5</sup>, Richard Bruskiewich<sup>2</sup>

<sup>1</sup> International Maize and Wheat Improvement Center - CIMMYT, <sup>2</sup> International Rice Research Institute, <sup>3</sup> Bioversity International, <sup>4</sup> International Potato Center - CIP, <sup>5</sup> International Crops Research Institute for the Semi-Arid Tropics

The volume and complexity of biological data is increasing day by day creating more and more problems for researchers to manage their data. The historical data that could be used in comparative studies are often scattered in various species-specific databases. Each database uses slightly different terminology, creating further difficulty for researchers in understanding the exact meaning of cross-specific terms that are related to phenotypes.

To support and encourage researchers and breeders to share and reuse information among agricultural databases, the Generation Challenge Programme (GCP) has emphasized building an "integrated platform" of molecular biology and bioinformatics tools. Part of this platform includes the development of Domain Models (DM) and GCP ontology. The DM is a generic metadata model upon which other model components are specified. The DM defines the general concept of a system "Entity" with "Identifier" and "Feature" components and relies on extensible ontology to define the semantics of the model. An ontology is a standardized (controlled) vocabulary of common, consistent and internationally recognized descriptive terminology arranged in a structured order, or a network based on biological concepts that can be shared and used uniformly among the species and to which genetic and phenotypic data can be associated. Each vocabulary term has a fixed identification term, a definition, and is structured as a hierarchical Directed Acyclic Graph (DAG). Terms can have one or more parents and zero, one, or more children. The terms are always linked by relationships such as "is-a", "part-of" etc. Currently, the GCP crop-specific ontologies are being developed for chickpea, maize, musa, potato, rice, maize and wheat traits. Initial sources of ontology terms for crop-specific ontologies are identified, for example, the ICIS for rice, wheat and maize and the Bioversity germplasm descriptors. Draft releases of the Open Biomedical Ontologies (OBO) formatted ontology

files for rice, wheat and maize traits are currently available at

<u>http://cropforge.org/projects/gcpontology/</u>. The crop-specific work plans and ontology related other materials are published at <u>http://pantheon.generationcp.org</u>.

The development and curation of general purpose ontologies will be continued and will be made available at the above mentioned Pantheon website and CropForge. In future, the GCP ontology will allow researchers and end users to query the keywords that are related to traits, plant structure, growth stages, molecular functions. Researchers can then link keywords to the associated phenotyping and genotyping data sets of GCP such as germplasm, crop physiology, geographic information, genes and QTL. The ontologies will facilitate comparisons between and within the crop species. The GCP ontology will be integrated with the Plant Ontology (PO) developed by the Plant Ontology Consortium (<u>http://www.plantontology.org/</u>) and Gramene (Trait Ontology, TO; Environment Ontology, EO) to develop a common, internationally-shared crop trait and anatomy ontology.

Support is acknowledged from: Generation Challenge Programme SP4 Project G4005.22: Development of Generation CP Domain Models and Ontology

# 19.37. Data Models To Share Invasive Species Information Through The Global Invasive Species Information Network

Annie Simpson<sup>1</sup>, Elizabeth Sellers<sup>2</sup>

<sup>1</sup> National Biological Information Infrastructure, <sup>2</sup> National Biological Information Infrastructure, U.S. Geological Survey

The sharing of standardized invasive species information is vital to improve our understanding of biological invasions and to enable the coordination of effective control efforts. The Global Invasive Species Information Network (GISIN) is the working body of the Biodiversity Information Standards - TDWG Invasive Species Interest Group (ISIG), and is a collaboration of like-minded invasive species information managers interested in sharing data globally. In 2008, GISIN held two workshops to establish Data Models for information sharing, and three models were accepted: Occurrences, SpeciesStatus (including concepts such as Abundance, Distribution, Harmful, RateOfSpread, Persistence, and RegulatoryListing), and SpeciesResourceURLs (for URL lists of one of the following types: species profile, image, video, experts, or references).

The GISIN Models use existing concepts of the Darwin Core draft standard, the Dublin Core standard, and several ISO standards. Additional concepts that are unique to invasive species science have been added to the GISIN Data Models, which we propose be considered as extensions to the Darwin Core. The GISIN Protocol that utilizes these data models is simplified from, and compatible with, TAPIR (TDWG Access Protocol for Information Retrieval). Because the GISIN Protocol is non-relational and contains fewer elements than the full TAPIR protocol, it has been dubbed TAPIRlite.

This poster will explain the history of the GISIN and describe the Data Models of the GISIN Protocol.

To facilitate communication among its members and with the interested public, the GISIN has several information tools:

- Non-technical online interface, email listserv, and information repository hosted by the U.S. National Biological Information Infrastructure at <u>http://www.gisinetwork.org</u>
- Technical documents repository and GISIN information system (including the data provider directory) hosted by the U.S. National Institute of Invasive Species Science at <u>http://www.niiss.org/gisin</u>
- Wiki for collaborative discussion and standards development hosted by TDWG at <u>http://wiki.tdwg.org/twiki/bin/view/InvasiveSpecies/WebHome</u>
- Technical email listserv hosted by the Hawaiian Ecosystems at Risk (HEAR) Project of the U.S. National Biological Information Infrastructure <u>http://www.hear.org/hearlists/gisintech.htm</u>

Support is acknowledged from: U.S. National Biological Information Infrastructure, U.S. Geological Survey

### 19.38. Xper<sup>2</sup>: an open tool to manage descriptive data

Visotheary Ung<sup>1</sup>, Régine Vignes-Lebbe<sup>2</sup>

Taxon names are the main access key to biological information (Polaszek, 2005) and the link between the real biological world and the conceptual world proposed by systematists through the description and characterization of taxa. The next revolution in taxonomy may be knowledge base management systems that support the generalization and the integration of current taxonomic work.

Nowadays, in the context of a distributed taxonomy, identification tools provide users with the resources to find taxon names and subsequently access to other knowledge about the organism. Such complete systems, able to manage descriptive data and to make identifications, will become more commonly used by both producers and users of biodiversity information.

For over twenty years, Xper has provided a complete environment dedicated to managing taxonomy. Different tools for the management of descriptions, interactive free-access identification keys, or other purposes (*e.g.*, to generate static keys and diagnoses) have been developed in order to facilitate a taxonomist's work. The new generation of this software, Xper<sup>2</sup>, is now available and offers new functionality that focuses on data exchange, such as the export of HTML files (for an easy and quick online distribution of the contents of a knowledge base) or the import/export through table format (CSV) and structured descriptive data (SDD) format (providing full compatibility between different platforms).

Here, we will present an overview of some collaborative work we are doing both in Zoology and in Botany, highlighting the practical and efficient links to other resources (*e.g.*, direct links to collection or observation databases, or to the GBIF portal). We will also focus on the flexibility and potential use of the software in other research fields, *e.g.*, as a diagnostic tool in human and veterinary medicine, control of protected species, education, etc.

Xper<sup>2</sup> is probably the most user-friendly tool in its category. It is one of the tools short-listed to appear on the cyberplatform being developed by the EDIT project on cybertaxonomy. EDIT will include many tools, giving taxonomists world-wide an open platform for exchanging ideas and data in a new, delocalized taxonomy.

### Reference

Polaszek A., 2005. A universal register for animal names. Nature 437: 477.

### 19.39. The SCAR Antarctic Biodiversity Database of terrestrial and limnetic data

### Dave Watts

Australian Antarctic Division

The Antarctic Biodiversity Database (<u>http://data.aad.gov.au/aadc/biodiversity/</u>) developed and hosted by the Australian Antarctic Data Centre (AADC), holds data and information on Antarctic and sub-Antarctic flora and fauna in support of the Scientific Committee on Antarctic Research (SCAR) Life Sciences programme on Evolution and Biodiversity in the Antarctic (EBA, <u>http://www.eba.aq/</u>).

The database catalogues details of collections containing details such as description, acknowledgments, data sources and lists of taxon types and regions. The samples and/or observations from each collection are classified into one of three possible habitat domains: terrestrial, limnetic or marine. Links are provided, where possible, to metadata records held in the Antarctic Master Directory (http://gcmd.nasa.gov/KeywordSearch/Home.do?Portal=amd) at the Global Change Master

Directory (GCMD). These records contain information on data quality, access constraints, collecting and analysis methods and the collection custodians.

Because the Antarctic Biodiversity Database is a central repository for EBA, users can query across all taxa, places, collecting dates, latitude and longitude, bounding box coordinates, elevation and collection names. The data can also be exported, or explored and queried from links within a collection or taxon record.

SCAR is an associate member of the Global Biodiversity Information Facility (GBIF) and encourages the publication of data into this global initiative.

Current plans include the improvement of the online mapping tools and to utilisation of the Antarctic Digital Database (<u>http://www.add.scar.org:8080/add/</u>), for background topography.

Support is acknowledged from: Australian Antarctic Division

# **19.40.** An Evaluation of Taxonomic Name Recognition (TNR) in the Biodiversity Heritage Library

Qin Wei<sup>1</sup>, Chris Freeland<sup>2</sup>, P. Bryan Heidorn<sup>1</sup> <sup>1</sup> University of Illinois, <sup>2</sup> Missouri Botanical Garden

The Taxonomic Name Recognition (TNR) algorithm – identifying a text string as a taxonomic name or not and recognizing the boundaries of the name – is very important in the BHL digitization project for determining whether the users/researchers could find the materials they want efficiently. The BHL has incorporated TaxonFinder, a taxonomic name finding algorithm and service (provided by uBio.org), into its portal for the identification and verification of taxonomic name strings found within the digitized BHL corpus. An eight-week evaluation was performed to determine the factors affecting the accuracy of the results returned. Our findings are not only valuable for BHL but also for other digital projects that would like to do text mining on their collections. In this evaluation project, we explored and analyzed the factors influencing the performance of: 1) Optical Character Recognition (OCR) for transforming images into text, 2) TNR matching algorithms for identifying taxonomic names from texts, and 3) the completeness of NameBank, which is used as an authority file for name verification.

We randomly selected 392 pages from the BHL database, which contained 4,843,619 pages at the beginning of our project. This sample included 3,003 valid names (2,610 unique names), which were identified manually by a group of biologists. For this sample, the OCR error rate for name strings was 35.16%, meaning that among the 3,003 valid names the OCR software packages correctly output 74.84% of them.

In digitization projects such as the BHL, TNR must also be able to find names even if they have OCR errors. So our evaluation standard included taxonomic name strings that were identifiable by humans as being names even when they had OCR errors. We assessed two TNR matching algorithms: TaxonFinder and FAT (Find All Taxonomic Names), which are widely used within the biodiversity community. The performance was evaluated by two measures: Precision (P) and Recall (R). Precision is the proportion of algorithm identified strings that are valid names. In our case, the precision means the capability of the algorithm to identify the valid names as well as exclude the non-valid name at the same time. Recall is the proportion of valid names in the sample that are recognized by the algorithms, meaning the capability of finding all valid names from the collection. TaxonFinder found 1540 names; 674 of them were correct names. FAT found 1603 names; 517 of them were valid names. The precision for TaxonFinder and FAT are 43.77% (=674/1540) and 32.25% (=517/1603) respectively. The recall for TaxonFinder is 25.82% (=674/2610) and FAT is 17.21% (=517/3003).

For TaxonFinder, the NameBank omission rate is 5.4%, *i.e.*, 5.4% of the real names found by TaxonFinder were not in NameBank. This demonstrates that names missing from the NameBank

authority file are not the major source of information loss in converting the data from images to a structured searchable database.

Our results indicate that improving the performance of TNR algorithms is the main challenge for producing an index to taxonomic names within digital library projects like BHL. The future work should determine what names are not found by the algorithms and the reasons why they were missed.

Support is acknowledged from: Biodiversity Heritage Library

# **19.41.** The Global Compositae Checklist – integration of nomenclature and taxonomic concepts from multiple data sources

Aaron Wilton<sup>1</sup>, Kevin Richards<sup>1</sup>, Jerry Cooper<sup>1</sup>, Christina Flann<sup>2</sup> <sup>1</sup>Landcare Research, <sup>2</sup>Wageningen University

The Compositae, or Asteraceae, is one of largest families of flowering plants containing an estimated 25,000 species – approximately 10% of the global flora. A GBIF seed funding project was initiated in 2006 with the intention of compiling a global checklist of the Compositae by integrating already existing electronic data sources (*e.g.*, Euro+Med, SANBI, IPNI, Tropicos, Conabio, Flora of Japan) that document most of the centres of high Compositae diversity.

To support the development of a global checklist, the project had two primary informatics aims (in addition to other objectives):

- Develop the informatics tools necessary to manage the integration of multiple existing checklists into a single consensus view, whilst maintaining the linkages to original records in the individual checklist data sets.
- Create a baseline integration of key datasets from identified project partners.

To meet these objectives, the project developed a tool, the Checklist Integrator Tool (C-INT), that provides an editorial interface alongside managed and automated workflows for the importation and integration of multiple checklists. The main features of C-INT include

- Importation of data via xml (including the Taxon Concept Schema) or a local database.
- Integration of nomenclatural, taxonomic concept, and literature data using a flexible, rulebased matching algorithm that attempts to reduce computation time by resorting to computationally expensive steps (*e.g.*, calculation of Levenshtein distances) only when necessary.
- Calculation of consensus records, which are issued an LSID (Life Science Identifier), based on a simple majority of the linked, original records, whilst allowing for editorial contributions to have priority in a transparent manner.
- Management, integration, and presentation of other data components that are often included in checklists (*e.g.*, vernacular names, distribution data) using xml and xslt transformations.
- An associated website and web services that include reporting on data linkages and data differences for any original, contributed identifier or consensus LSID.

C-INT is now actively being used to create the Global Compositae Checklist. To date, the Checklist includes:

- o Data from 19 checklists' data sets ranging in size from 220 to nearly 160,000 records,
- o 373,052 contributed nomenclatural records, and
- o 263,462 contributed taxonomic concepts.

These have been integrated using C-INT to produce 136,377 and 211,156 consensus nomenclatural and concept records respectively.

This project now provides us with the opportunity to analyse this growing dataset for patterns such as the degree of overlap among datasets, levels of conflict, and the relative contribution of datasets.

Overall the data indicate that there is substantial overlap of names recorded in different checklists (ca. 2.7 contributed names per consensus name), but less overlap between the concepts recorded in those checklists. Thus the process of integrating separate regional checklists is creating a baseline database from these different sources, which expands the number of related concepts associated with each taxon.

# **20. Contributed Abstracts**

# 20.1. Managing Biodiversity Data and Species File Software

David C. Eades Illinois Natural History Survey

By developing recognized standards, TDWG plays a critical role in enabling the flow of information from a multitude of scattered databases to a small number of very large databases and websites. This has made much biodiversity information accessible to users who lack the time, expertise or resources to find and use the information without this concentration. But when we look at the overall biodiversity task to be done, the overwhelming bulk of the work is for taxonomic specialists, most of whom are unwilling to deal with technical computer issues. Some taxonomists have concluded that TDWG is irrelevant and a diversion of resources from "real" taxonomic work. That view denigrates the importance of providing useful access to taxonomic information for the wider community. At the same time, one might wonder if computer experts fully comprehend the magnitude of the task faced by taxonomists.

We can think of TDWG standards as defining a pipeline from original data to a form that meets the needs of users. Defining the pipeline and presenting the emerging data enjoy the most prominence. The bigger task of feeding data into the pipeline is not rewarded with comparable prominence. Nevertheless, this task is small compared to the basic taxonomic work to create the data. The computer experts in TDWG could see their work in three steps:

- 1. Develop standards and programs to convey the information through the pipeline and make it accessible.
- 2. Develop programs needed to harvest the store of existing data.
- 3. Develop tools that will make the work of taxonomists more efficient.

The greatest benefit toward our knowledge of biodiversity awaits the third step.

Species File Software <u>http://software.speciesfile.org</u> occupies a point in the middle of the pipeline. Fourteen species files are supervised by taxonomists with expertise in specific groups. This management of data at an intermediate stage provides substantial quality improvement for the data emerging at the end. The species files include 47,000 valid species, 220,000 citations, 57,000 images, 48,000 specimen records and keys with 2,600 end points.

A future objective for Species File Software is to add features that taxonomists will use as they work on revisions. Taxonomists' work would be more efficient; capturing data after publication would become simple. The data would be in species files but not open to the public until authorized by the author. This process could also recover incomplete work of authors who never publish.

It has been exhilarating to see the progress reported at TDWG meetings, but attempting to keep pace with the developing standards has been a frustrating task. We badly need collaborators in the development of Species File Software. Four software developers and fourteen data managers provide a group of volunteers willing to test and critique new standards.

# 20.2. PlantCollections – A Community Solution

Boyce Tankersley<sup>1</sup>, Greg Riccardi<sup>2</sup>, David Vieglais<sup>3</sup>, Min Cai Henderson<sup>1</sup>, Christopher Dunn<sup>4</sup>, Dan Stark<sup>5</sup>, Pam Allenstein<sup>6</sup>

<sup>1</sup> Chicago Botanic Garden, <sup>2</sup> Florida State University, <sup>3</sup> University of Kansas Natural History Musuem and Biodiversity Research Center, <sup>4</sup> University of Hawaii Lyon Arboretum, <sup>5</sup> American Public Gardens Association, <sup>6</sup> North American Plant Collections Consortium of APGA

PlantCollections provides access over the Web to plant records data held in botanic gardens and arboreta <u>www.plantcollections.org</u>. PlantCollections was developed by a community of eighteen botanic gardens & arboreta to provide access to records of over 89,600 taxa.

Survey results of the information needs from eight scientific disciplines were cross referenced with surveys of botanic garden and arboreta database fields to determine the 161 fields in the federated schema. A federated schema is a list of different types (fields) of information and a set of guidelines to insure all of the data are in the same format. Darwin Core and International Transfer Format (ITF2) for botanic garden records (<u>http://www.tdwg.org/standards/102/</u>) field definitions map to 37 and 44 fields respectively. One hundred and six fields are unique to the schema.

These records reflect environmental and biological data from collection sites and plant growth, aesthetics, propagation, mortality and weather data collected in botanic gardens and arboreta in the USA, China and the United Kingdom.

This project is the first instance of the use of Google Base (<u>http://www.google.com/base</u>) as a central Web repository for biodiversity data. Google Base has a quick response, caches large volumes of data, is free, extensively mirrored and requires data uploads monthly to eliminate serving out-of-date records.

The Portal for accessing the data integrates Web 2.O (<u>http://en.wikipedia.org/wiki/Web\_2.0</u>) advances and has been developed to access other data providers using RSS (<u>http://www.whatisrss.com/</u>) or ATOM (<u>http://en.wikipedia.org/wiki/Atom\_(standard</u>)) . Living plant collections, herbaria, DNA, image & seed storage databases all contribute data. A partnership with Morphbank (<u>www.morphbank.net</u>) has provided storage of digital images while Google Map/Earth can be used to create distribution maps.

Among the benefits to scientists are the creation of an application that bridges technology, nationality, language, professional affiliations & cost concerns. It supports the study of plant origins, characteristics of cultivated plants and provides a mechanism to contrast cultivated environments with those of natural populations. It identifies commercial sources for many botanic garden plants while providing insights in to plant propagation protocols. Data users may search records from one or more institutions to create reports, maps and images.

Support is acknowledged from: Institute of Museum and Library Services National Leadership grant in the Building Digital Resources category. Additional funding from an anonymous donor is supporting expansion of the project beyond the United States. Collectively, the 18 participating institutions have donated over \$666,453 in staff time and institutional resources towards the PlantCollections project.