Biodiversity Informatics, 7, 2010, pp. 130 – 136.

LEVERAGING THE FULLEST POTENTIAL OF SCIENTIFIC COLLECTIONS THROUGH DIGITIZATION.

ROGER BAIRD
Collection Services Division, Canadian Museum of Nature, Ottawa, Ontario Canada
rbaird@mus-nature.ca

Abstract -Access to digitized specimen data is a vital means to distribute information and in turn create knowledge. Pooling the accessibility of specimen and observation data under common standards and harnessing the power of distributed datasets places more and more information and the disposal of a globally dispersed work force which would otherwise carry on its work in relative isolation, and with limited profile and impact. Citing a number of higher profile national and international projects, it is argued that a globally coordinated approach to the digitization of a critical mass of scientific specimens and specimen-related data is highly desirable. An action plan of this scale is required to maximize the value of these collections to civil society and to support the advancement of our scientific knowledge globally.

Key words. biodiversity; metadata; digitization; science infrastructure

Natural and Human History collections are a part of the wider scientific collections infrastructure. Currently, the Earth is estimated to be home to approximately 11.3 million species, but less than 2 million have been formally described by science (Chapman, 2009). Whether these species are exploited for commercial gain, or conserved for ethical, aesthetic and scientific reasons, the limited knowledge that we have about the biological diversity of the planet is of serious concern internationally. Scientific collections can be characterized as assemblages of natural history specimens as well as human history artifacts that have been sufficiently documented, at the time of acquisition or through the course of analysis, as to have lasting value as part of a broad research infrastructure.

HOW ARE SCIENTIFIC COLLECTIONS AND THEIR DATA USED?

CULTURAL COLLECTIONS:

Scientific collections of material culture have been important resources for archaeologists, anthropologists and ethnologists, facilitating studies of ancient and living cultures as well as comparative analysis between cultures. The ability to supplement personal networks of colleagues and, in part, to overcome their ephemeral nature, is a key benefit from the creation of datasets that serve a global audience. The data allows institutions to perpetuate the knowledge created through the life work of a researcher, and to broaden the reach of its holdings for the benefit of others.

Equally, data on these human history collections is of interest to other individuals who may be related to, or representing, the very people studied. Interest in traditional techniques, a desire to connect with the past, an objective to undertake a physical or “virtual” repatriation of one’s culture are all potential motivators for the broader audience for digitized material culture collections.

NATURAL HISTORY COLLECTIONS:

The traditional users of natural history scientific collections are invariably taxonomists - identifying, naming and classifying species- and systematists who study the diversity of life on the planet’s past and present as well as the relationships among living things through time. These specialists make it possible for the comparative science of biology to flourish. Without this essential work, a large portion of anatomy, physiology, biochemistry and microbiology, and ecology could not be realized. This work is essential to ensure economic well-being, to preserve natural resources, to maintain health, and to guard against invasive species.

Whether they were amassed historically or compiled recently in response to pressures from development or exploitation of a site, scientific collections are invaluable resources to answer science-based questions far beyond the reach of a single individual. The collections themselves are sub samples of the world as it once was and as we know it today, and can also provide critical predictive modeling data on what the future could be. Providing digitized access to the information that is inherent in these collections and having this inherent information converted into insight and knowledge by appropriate specialists allows a researcher or policy maker to verify a range of questions related to a wide variety of subjects:

BIODIVERSITY AND ENVIRONMENTAL CHANGE

Collections offer evidentiary value for documenting the biological diversity of life and in doing so, can also demonstrate changes in the environment that have taken place through time. The presence or absence of a species in a geographic region, as well as extensions and retractions in the distributional range of a species over the course of time are documented by examining the records associated with these collections. A habitat recovery program can be demonstrated to be successful if species once thought to be extirpated from the area or endangered are documented as being re-established, through observation records as well as physical specimens. Scientific collections that are well documented and deposited can offer proof that mitigation measures were successful, can provide evidence by proxy that climate change has taken place, and can help to effectively monitor rare, threatened and endangered species.

INVASIVE ALIEN SPECIES

Examining the causes of biodiversity loss, digital mapping techniques have revealed that invasive alien species are second only to the threat posed by habitat destruction. In a 1993 report, the U.S. Office of Technology Assessment[1] estimated cumulative economic losses of $100 billion in the US due to noxious weeds, invasive insect pests, introduced aquatic species from ballast water, and other non-indigenous species (Simberloff, 1996). The ability to differentiate native from non-native species is necessary and made possible by analyzing the specimen holdings in natural history collections, which provide that baseline data on which these analyses are built. Collection-based science is essential to prevent, detect, and to rapidly respond to and manage these threats.

PUBLIC HEALTH AND WILDLIFE DISEASE

Important scientific collections are also managed outside of museum environments. Viruses, cultures, tissues and pathology samples are an important subset of this scientific infrastructure to be found in research laboratories and biological resource centers or BRCs.

Infections from H1N1 Influenza, West Nile Virus, Lyme disease, tuberculosis, chronic wasting disease, and SARS are all medical conditions that manifest themselves in human and non-human populations. For these and other zoonoses, approximately 70% of new or newly important diseases affecting human health are believed to have a wild animal source (Blancou, 2005), and can have profound impacts on urban, rural, and human health, culture, and global economy.

Collections data over time can reveal biodiversity threats through knockout effects on ecosystems e.g. high mortality of UK rabbits after introduction of myxomatosis led to declines in predators such as stoats, buzzards, and owls (Sumption, 1985, 2008); the reduced grazing pressure by rabbits on heath lands in turn removed the habitat for an ant species that assists developing butterfly larvae, leading to extirpation of populations of the endangered large blue butterfly.

ECONOMICS, BIOSECURITY AND REGULATORY FRAMEWORKS

The predominance of global trade and marketing in our modern world requires an internationally coordinated infrastructure to share expertise derived from scientific collections. Individual customs or border security officials who exercise levels of control on the movement and transport of imported goods and products have a reliance on knowledge derived from these collections on a daily basis. For example, the invasive pest *Agrilus planipennis* or Emerald Ash

1 http://www.fas.org/ota/reports/9325.pdf

Borer is considered to have been introduced to North America through infested wood crating materials in 2002, and has come to rival Dutch Elm disease in its impact on tree populations[2]. Timely access to information can counter distribution of fungal or insect infections through horticultural trade, and can reduce economic loss from unwarranted delays in customs inspections and quarantines (Renaud, 2008). As well, genetic species barcodes linked to taxonomic collections have been demonstrated to be of great significance in detecting market substitutions involving over-fished species or market fraud such as *Lutjanus campechanus* or Red Snapper (Wong, 2008).

ACCESS AND BENEFIT SHARING

Material culture collections of ancient or indigenous cultures are predominantly held by developed countries, while parties related to the cultures studied are most prevalent in underdeveloped countries. Equally, biological collections and related scientific expertise are held disproportionately within developed countries, although the greatest portion of the world's biological diversity is found in countries which are presently not as economically advantaged.

Under the Convention for Biological Diversity[3] sharing information and assisting international development through sustainable development practices is considered a global imperative, and facilitating access to data and knowledge on material collections supports the perpetuation of knowledge from traditional cultures.

Documentation from archaeological sites has also been relevant to the resolution of land claims by aboriginal groups, by evidencing traditional use and occupation of geographic areas. Scientific collections both historical and modern also have the potential to assist in the isolation and identification of biopharmaceuticals. For example, Taxol as a treatment for ovarian cancer has been derived from *Taxus brevifolia* (Pacific Yew) (Stierle, 1994) and such ethno botanical sources or "traditional" medicines have become the subject of discussion for recognition and possible compensation under patent regimes.

These kinds of advances in information technologies and tissue sampling techniques for molecular biology and genomics are creating new applications for traditional collections beyond their original intents as objects of study and as vouchers that allow for the verification of research results. It is unfortunate therefore that the preservation of, and providing access to, scientific collections is not fully seen as the "big science" that it truly is on an international scale. Researchers are spread out across the globe, searching for the new and unexpected, rather than working together on a single project or at a major new facility. The resulting lack of basic knowledge puts at risk other research and development investments in areas such as biotechnology, genomics, agriculture, forestry, fisheries and aquaculture, and public health. The promise of new information technologies, from genomics to geographic information systems, is that research results can be captured in a more systematic way, and will contribute to a greater understanding of the whole.

SIGNIFICANCE OF SCIENTIFIC COLLECTIONS

The importance of scientific collections is exemplified to great effect in the United States of America, where the Office of Science and Technology Policy (within the Office of the President) has recognized scientific collections as critical scientific infrastructure since 2005. An Interagency Working Group on Scientific Collections was created to conduct a survey of scientific collections held by Federal agencies and collaborated with the National Science Foundation to document collections not owned by the Federal government. Among its findings the working group has identified the need for: a) a comprehensive, government-wide mechanism for the responsible management of Federal collections that addresses short- and long-term preservation issues both within and between agencies, and, b) the establishment of an information clearinghouse, for agencies to share policies, procedures, and other information[4] The digitization of this

2 http://www.emeraldashborer.info/

3 http://www.cbd.int/abs/

4 http://www.whitehouse.gov/sites/default/files/sci-collections-report-2009-rev2.pdf

infrastructure is a key element in making these specimens accessible, searchable and distributable to present and future researchers wherever they may be physically located.

WHAT IS CURRENTLY BEING DIGITIZED AND TO WHAT EFFECT?

Access to digitized data at the specimen-level has been a vital means to distribute information and to create the potential for that information to generate knowledge. The Global Biodiversity Information Facility[5] is widely recognized as a leading example of the multiplier effect that is created by pooling the accessibility of specimen and observation data under common standards and harnessing the power of distributed datasets. And yet, the very strength of such a model is almost equally its weakness. The task at hand is made so daunting because of several factors: the sheer number of specimens, the variety of analogue and digital formats in which associated data are held, the fact they are dispersed in repositories large and small in all corners of the world, and the reality that they are amassed over decades and even centuries. The development and application of data standards which provide common descriptive language and common binary formats are critical tools to facilitate this access.

Metadata can be used to capture the essential information describing the general nature, as well as the how and when and by whom, of a particular set of data, as well its format. This has the effect of characterizing larger groupings of information into formats that are discoverable by users and are sufficiently described to stimulate interest in the grouping by the end user. The metadata is in effect the finding aid or catalogue that can lead the researcher onto further discovery. Metadata standards and their application are therefore critical to rapidly characterize and quantify specimen information, but impose limits on what can be discovered by the data user. Metadata can lead to promising paths of discovery but will rarely reach beyond the function of way finder.

The scientific collections that can be accessed through metadata standards are not limited to biological or natural history holdings. A number of national and international initiatives have also been examining and highlighting the significance of material culture and their related data, in addition to biological material, and these are enumerated below:

1) The Australian Research Council's Network for Early European Research (NEER) is one of many digitization initiatives (Burrows, 2006) which have made use of the power of metadata –data about data. *Europa Inventa* (the Australian Collections Service) was envisaged as a gateway to all Early European items in Australian collections (Burrows 2008), including manuscripts, papers, artworks, maps, furniture, fabrics, scientific instruments, and other material culture, with a focus on unique items and those with specific associations rather than on mass-produced items such as early printed books. Links from descriptive catalogue records to digitized images held on the distributed servers of the holding institutions provide this virtual access, and commissioning of digitization work in these repositories was also envisaged as part of the project.

2) Colombia has engaged in a large scale systematic program of digitization under policy initiatives established by the Ministry of Culture. At an International Workshop on Digital Preservation and Copyright organized by the World Intellectual Property Organization (WIPO) in Geneva in July 2008, participants heard general details of two important national initiatives. In recognition of 200 years of independence in 2010, the Bicentennial Digitization Program will digitize and publish on the web the most important documents of the time from Colombia's National Library, National Archives, the National Museum and The Central Bank's Library. The resulting standards and best practices will be leveraged to give form to a National Digitization Plan. As well, the Colombian Digital Library[6] has been established as a joint program between thirteen universities. Working with *Colciencias* (The National Science Committee) and *RENATA*[7] *(*The National Academic Web of High Technology), they will define standards and mechanisms for

5 http://www.gbif.org/

6 http://www.bdcol.org:8080/

7 http://www.renata.edu.co/

digitization. This collaboration may overcome economic and legal issues that impede preservation of the works and access by researchers and the general public.

3) In the realm of biodiversity-related collections, twelve Federal Departments and Agencies within the Government of Canada have been collaborating in a horizontal initiative known as the Federal Biodiversity Information Partnership (FBIP)[8]. The 'natural capital' of the country's biological resources - from molecules and genes to organisms and ecosystems – are being positioned as resources that have the potential to yield benefits ranging from bio economic and environmental to social and human health. The key to unlocking these benefits is the creation of integrated bio-information systems of scientific information with interoperability of databases and national specimen repositories.

4) The European Community demonstrated its recognition of collection institutions as important research infrastructures by its support and funding of the European Distributed Institute of Taxonomy (EDIT)[9]. This "Network of Excellence" project is to integrate institutional policies and infrastructures of the participating 25 collection institutions. The continuing funding for the Synthesis of Systematics Resources (SYNTHESYS)[10] supports an initiative comprised of 20 European natural history museums and botanic gardens. Its objective is to create an integrated European infrastructure for researchers in the natural sciences through access and networking activities. Additionally, a preparatory project was granted for the preparation of a large scale and long term research facility ("LifeWatch")[11], uniting these networks with those in the areas of marine and terrestrial ecology.

5) As recently as August 2008 in the United Kingdom, the House of Lords Science and Technology Committee released its follow up report on Systematics and Taxonomy[12] (see para.2.13),

> *Measuring progress towards halting the decline in biodiversity is a key international obligation which cannot be achieved without baseline knowledge of biodiversity. Creating baselines and monitoring change is dependent upon the availability of taxonomic expertise across the range of living organisms. Systematic biology underpins our understanding of the natural world. A decline in taxonomy and systematics in the UK would directly and indirectly impact on the Government's ability to deliver across a wide range of policy goals.*

The conclusions of the report also emphasize the need for financial and infrastructure support in digitizing biological collections to facilitate the aggregation of collections data.

6) On the global scale, the need for large scale collaboration is equally reflected in the reports of the Megascience Forum of the Organization for Economic Cooperation and Development (OECD) that led to the foundation of GBIF[13], and the Global Taxonomy Initiative (GTI) of the Convention on Biological Diversity[14].

7) An initiative by the Netherlands in 2006 under the Global Science Forum of the OECD[15] led to workshops convened in Leiden (June 2007), and Washington D.C. (July 2008) with a steering committee guiding a proposal to establish an international coordinating mechanism supporting scientific collection-based institutions, in their collective role as part of a unique global research infrastructure (see pp.1-2),

> *Scientific collections are essential parts of the research infrastructure of all countries with scientific enterprises, and they are critical to many areas of science, from microbiology to space science.*

8 http://www.cbif.gc.ca/fbip/fbip_e.php

9 http://www.e-taxonomy.eu/

10 http://www.synthesys.info/

11 http://www.lifewatch.eu/

12 http://www.publications.parliament.uk/pa/ld200708/ldselect/ldsctech/162/162.pdf

13 www.gbif.org

14 http://www.cbd.int/gti/

15 http://www.oecd.org/topic/0,3373,en_2649_34319_1_1_1_1_37417,00.html

National governments share an interest in finding answers to basic research questions and many applied research challenges, and no one nation has all the assets to pursue major research challenges independently…The Mission of an International Coordinating Mechanism for Scientific Collections would include the following:

- *Enable global-scale research activities*
- *Promote an international culture of collections as large-scale distributed infrastructure*
- *Improve access to and mobility of collection objects and associated data, and the people associated with them; foster capacity-building*
- *Identify and integrate existing standards of community practice, and develop additional standards deemed necessary*

In order to fulfill this mission, this coordinating mechanism should undertake the following actions:

- *Create a research roadmap in coordination with the user community*
- *Create self-assessment tools for collections*
- *Set standards of practice*
- *Promote research on scientific collections and collections management*
- *Provide opportunities for the global collections workforce*
- *Provide a clearinghouse mechanism/interface between collection-based science and broader societal concerns/policies*

The report[16], submitted to the GSF in Krakow, Poland (OECD, 2008) elaborates an implementation plan for such a coordinating mechanism. Strong satisfaction was expressed by the GSF delegates for the work done to date, with Germany, France, Australia, Canada, Belgium, Holland, and the United Kingdom registering formal comments of support. Scientific Collections International or SciColl[17] has emerged as a nascent program, advancing on the strength of a well elaborated strategic plan, governance model and a program of outreach activities.

In conclusion, a globally coordinated approach to the digitization of a critical mass of scientific specimens and specimen-related data is highly desirable and required, to maximize the value of these collections to civil society and to support the advancement of our scientific knowledge globally. A more cohesive and inclusive approach to the digitization of scientific collections is highly desirable for all of these reasons and more.

REFERENCES

Blancou J, Chomel BB, Belotto A, Meslin FX. 2005. *Emerging or re-emerging bacterial zoonoses: factors of emergence, surveillance and control.* Vet Res. May-Jun;36(3):507-22. [18]

Burrows, T. 2006. *Network for Early European Research (NEER)Digital Services-Background*, Perth, Australia. [19]

Burrows, T. 2008. *Europa Inventa (Australian Collections Services)*, Perth, Australia. [20]

Chapman, A.D. 2009. *Numbers of Living Species in Australia and the World.* Report for the Australian Biological Resources Study, *2nd ed. 80 pp.* Canberra, Australia. [21]

_____, 2009 . *OECD Global Science Forum Second Activity on Policy Issues Related to Scientific Research Collections: Final Report on Findings and Recommendations* Submitted October 2010 to

16 http://www.oecd.org/dataoecd/7/58/42237442.pdf

17 www.scicoll.org

18 http://www.ncbi.nlm.nih.gov/pubmed/15845237

19 http://confluence.arts.uwa.edu.au/display/DIGITAL/NEER+Digital+Services+-+background

20 http://confluence.arts.uwa.edu.au/display/DIGITAL/Europa+Inventa

21 http://www.environment.gov.au/biodiversity/abrs/publications/other/species-numbers/2009/pubs/nlsaw-2nd-complete.pdf

the OECD Global Science Forum, Krakow, Poland. [22]

Martin, A. 2008 *International Workshop on Digital Preservation and Copyright*. World Intellectual Property Organization, Geneva. [23]

National Science and Technology Council, Committee on Science, Interagency Working Group on Scientific Collections, 2009. *Scientific Collections: Mission-Critical Infrastructure of Federal Science Agencies. Office of Science and Technology Policy*, Washington, D.C. [24]

U.S. Congress, Office of Technology Assessment, 1993. *Harmful Non-Indigenous Species in the United States*, OTA-F-565 U.S. Government Printing Office, Washington, D.C. [25]

Renaud, M.-A. 2008. *DNA Barcode, Trade, Plant Health and Quarantine in the Canadian Ornamental Industry*. Presented at Canadian Barcode of Life Network, 2nd Scientific Symposium, Royal Ontario Museum Toronto.

Stierle, A. et al. 1994. *Endophytic Fungi of Pacific Yew (Taxus brevifolia) as a Source of Taxol, Taxanes, and Other Pharmacophores*. Bioregulators for Crop Protection and Pest Control Chapter 6, pp 64–77 Chapter DOI: 10.1021/bk-1994-0557.ch006 ACS Symposium Series, Vol. 557 .

Sumption K.J. and Flowerdew, J.R., 1985. *The ecological effects of the decline in Rabbits (Oryctolagus cuniculus L.) due to myxomatosis.* Mammal Review,15: 151–186. re-published online 2008 at [26]

Sutherland of Houndwood et al. 2008. *Systematics and Taxonomy: Follow-up.* Report of the House of Lords Science and Technology Committee, London [27]

Simberloff, D. 1996 *Impacts of Introduced Species in the United States* Consequences Vol.2 No.2 United States Global Research Information Office, Washington [28]

Wong, E. H.-K. and Hanner, R.H. 2008. *DNA barcoding detects market substitution in North American seafood.* Food Research International,Volume 41, Issue 8, October 2008, p.p. 828-837 [29]

22 http://www.oecd.org/dataoecd/7/58/42237442.pdf

23 http://www.wipo.int/edocs/mdocs/copyright/en/wipo_cr_wk_ge_08/wipo_cr_wk_ge_08_www_105896.pdf

24 http://www.whitehouse.gov/sites/default/files/sci-collections-report-2009-rev2.pdf

25 http://www.fas.org/ota/reports/9325.pdf

26 http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2907.1985.tb00396.x/references

27 http://www.publications.parliament.uk/pa/ld200708/ldselect/ldsctech/162/162.pdf

28 http://www.gcrio.org/CONSEQUENCES/vol2no2/article2.html

29 http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6T6V-4SYJS3M-2&_user=8844459&_coverDate=10%2F31%2F2008&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_searchStrId=1439577567&_rerunOrigin=scholar.google&_acct=C000109520&_version=1&_urlVersion=0&_userid=8844459&md5=ca1f16fbe2a05e7b9fc1916102e565b3