SOWING SEEDS IN THE DIGITAL GARDEN

Murray Henwood
Susan Hanfling
Rowan Brownlee
Tristan Gutsche
Belinda Pellow
University of Sydney

Increasing interest in learning, teaching and research in the plant sciences has led to a growing demand for flexible reference, research and teaching tools for use in the laboratory and field. *eFlora*, an electronic compendium of the plants of the Sydney region and *eBot*, a digital repository of botanical objects are two partnership projects that support these activities.

eFlora is a taxonomically comprehensive description of the vascular plants of the Sydney region. In addition to interactive keys for species identification, *eFlora* will have an illustrated, interactive glossary of technical terms. *eBot* provides a sustainable, standards-based repository for digital objects sourced from fieldwork, existing collections and via conversion from physical formats. It acts as a source of content for the development of a range of innovative and dynamic digital services. For example, *eFlora*, and its associated glossary of terms, will access objects within *eBot* to aid in plant identification.

eBot and *eFlora* will provide an authoritative resource to a diverse community of users worldwide. These types of standards-based resources, working within an interoperable framework, support cross-disciplinary application and the growth of knowledge.

Introduction

It is an exciting period of transition for the botanical community as they move from the use of mainly print-based research tools to the creation and application of digital solutions for archiving, management and sharing of information.

Throughout history, botanists have created and summarised the information derived from collected plants and observations of plant communities to produce an account of the plants growing in a specific geographic area. These accounts, known as 'Floras' are supported by

catalogued collections of dried plant specimens held in herbaria. Researchers have used these Floras, including printed taxonomic keys to identify plants included in their botanical research. In more recent times, these tools have increasingly been digitised, firstly for local use then moved onto distributed networks and the Internet. Data associated with herbaria are currently being digitised and uploaded into web-enabled databanks; for example the Australia's Virtual Herbarium (Australian National Herbarium, 2006) and PNGplants (Conn, Banka, & Lee, 2006). Floras from diverse geographical regions have started to appear online,[1] and the development of software and schema has supported the presentation of electronic keys in a range of formats.

In a similar way to other research communities, botanists have been developing digital standards over an extended period of time, using an intensely collaborative process with diverse geographical representation. Discussions about digital standards in the botanical sciences are taking place worldwide via groups such as TDWG (International Union of Biological Sciences Taxonomic Databases Working Group, 2006a), HISCOM (Herbarium Information Systems Committee, 2006) and GBIF (Global Biodiversity Information Facility, 2006). Core standards are beginning to emerge and an important characteristic of most botanical digital projects is the flexibility to adapt to new guidelines as they are produced by the key collaborative advisory groups.

These new foundation digital datasets and tools, and the shared standards that underpin them, will provide enormous opportunities for botanists and allied researchers. They will result in a corpus of resources that is globally accessible and interoperable and, in time, will support the overlay of datasets from other disciplines to produce new knowledge.

This paper describes two projects currently underway at the University of Sydney which seek to add to the global foundation resources for botany: *eFlora* of the Sydney Region and *eBot*: a database of objects for the plant sciences.

eFlora of the Sydney Region

One of the main outputs of plant taxonomic research is the production of keys to the identity of the plants of a particular geographic region. The keys are often accompanied by descriptions and illustrations of the species under consideration, and they sometimes provide an indication of the

relationships between and within the various levels of the taxonomic hierarchy.

Traditionally, keys have employed a dichotomous structure comprising a series of parallel and mutually exclusive choices or couplets. When identifying a plant, a user chooses the most appropriate description of a characteristic at each couplet and proceeds through the key until they reach the name of their unknown plant. Dichotomous keys are often grouped together with illustrations, maps of distribution and descriptions of each species in what is known as a regional *Flora*. The Flora of a particular geographic area, then, consists of a repository of taxonomic knowledge in the form of descriptions and a tool by which users can identify the plants from the area.

Floras differ from other means of plant identification – such as guides or manuals - by providing a comprehensive coverage of all species for a particular geographic region. In addition, Floras can contain information on flowering time, chromosome number, common name(s), uses, conservation status and ecology. In this way, floras provide core data for field scientists but are also powerful learning tools as students move through an experiential pathway to gain new knowledge.

The hierarchical nature of plant classifications provides a structured framework that can be readily displayed in a web environment. Since individual dichotomous keys are usually arranged according to these classifications, these keys can also be effectively presented via the internet. Online interactivity enables users to traverse hyperlinked pathways through the keys in a way that is much easier than moving from one section to another within a printed book. The learning potential of the web-based keys can be further enhanced by incorporating associated learning objects such as high quality illustrations of plants and their defining characteristics, and by providing immediate access to the technical language of botany by way of electronic glossaries.

A project that combines these attributes has been developed by the University of Sydney and the University of Wollongong, supported by funding from the New South Wales Environmental Trust. The printed *Flora of the Sydney Region* (Carolin & Tindale, 1994) is being revised, updated and digitised to form the centre-piece of a multifunctional, web-based resource - the *eFlora* - for the presentation of information on botanical diversity. An integral part of the *Flora of the Sydney Region* is an illustrated glossary of technical botanical terms.

The *eFlora* comprises keys to all 3000 vascular plant species found in the geographic region from the Shoalhaven River to the Hunter River, and west to the Great Divide. The text of the *Flora of the Sydney Region* will be marked-up in XML. This approach has a number of benefits including the long-term storage and management of the text, the ability to generate a word index to enable fast keyword searching and the capability to support online interactivity and generate outputs for a variety of uses including print and web publishing. One of the most significant advantages of this approach is the ease with which the content of the *eFlora* can be updated to reflect the latest concepts in plant taxonomy.

Production of the *eFlora* has required the investigation and enhancement of standards for the presentation of Floras in a web environment. The Australian Biological Resources Study (ABRS), for example, has developed a schema (Australian Biological Resources Study, 2004) for use with its multi-volume publication, *Flora of Australia* (Australian Biological Resources Study, 2006). ABRS are currently reviewing their schema with reference to work being undertaken by TDWG (International Union of Biological Sciences Taxonomic Databases Working Group, 2006b). The project team for *eFlora* are planning discussions with TDWG to ensure that the current projects are compliant with emerging standards being developed across the international botanical community.

One of the associated aims of *eFlora* is that it should reflect the content of the printed version of the *Flora of the Sydney Region*, but with enhanced operability. To this end, the glossary component of the *Flora of the Sydney Region* will be marked-up in XML and integrated with the text of *eFlora* by way of automatically hyperlinked keywords from the glossary. This format will ensure that the glossary can be used independently of *eFlora* (for stand-alone teaching purposes) or can be integrated with other botanical information services.

eBot: a database of objects for the plant sciences

The source of illustrations for both *eFlora* and the accompanying electronic glossary will be a digital repository of research and learning objects (*eBot*) currently being developed by the University of Sydney. By linking digital objects in *eBot* to plants in *eFlora,* the research and learning capabilities of both *eFlora* and *eBot* will be further enhanced.

Visual media have long been used in the plant sciences to support research and learning. Plant taxonomic decisions, for example, are often accompanied by a description, an illustration of the plant, and a photograph of a representative herbarium specimen. *eBot* will contain digital objects ranging from scanned dried herbarium specimens through to complex digital files such as those produced using cutting-edge technology like tomography.

eBot is being implemented using iSpheres (iSpheres, 2005), an open standards application developed by the University of Sydney as part of its role in the APSR (Australian Partnership for Sustainable Repositories, 2006). As a locally developed open source application, iSpheres offers the opportunity to tailor functions to specific requirements but also to share developments with the wider research and information technology community. iSpheres was selected for the *eBot* project as it has been tested on several other projects and it can be configured so that the broad range of functionality requirements can be managed. *eBot* requires an application to manage a large collection of images and other objects; enable development of customised user interfaces and support object and description upload from anywhere in the world. It also requires support for discipline-specific descriptive metadata and automatic extraction of technical metadata. Flexibility and ease of use are also important outcomes that must be maintained during the development of *eBot*. Using iSpheres, each object in *eBot* will have a unique URL that can be linked to learning management applications, web pages and a range of other digital resources. Users will be able to create an album of images, download objects in a format tailored on-the-fly to the user's specifications and apply for a high resolution copy of an object for publication via the *eBot* website.

The *eBot* access model allows any user to search for and download medium quality images. Authentication and authorisation is required to perform more advanced functions such as uploading images or editing metadata records. Access for users from the University of Sydney is currently managed through a web-based single sign-on system developed at the University of Sydney. However, the recommendations of the Internet2/MACE project Shibboleth (Internet2, 2006), in consultation with the Australian Meta-Access Management System (MAMS) project (Macquarie University, 2006) are being considered. This would enable remote *eBot* users who are members of partner institutions to login once through their local organisation and gain immediate access to *eBot* without the need for a second login account.

The project team sought advice from a range of imaging and preservation agencies on standards for image conversion.[2] *eBot* colour slide conversion specifications are based on requirements for digital preservation, high quality print reproduction, and accurate representation of colour. Slides will be scanned at 4000 dpi to create 24 bit master files in uncompressed TIFF format. Copies in JPEG format will be created for transmission over networks for use within services such as *eFlora*.

As *eBot* is intended to support archiving of high-value content, images in a preservation format such as uncompressed TIFF is optimal. However, it is understood that this requirement can not always be met, for some contributors might possess extensive collections of images of great research value in other formats. As these types of collections also require a secure and managed environment, *eBot* currently supports TIFF, PNG and JPEG image formats. If an image is accepted in a non-supported format, then the risks of format conversion will be evaluated before proceeding. If accurate conversion can not be guaranteed the original object will be retained and a converted copy will be made available for access. *eBot* will contain other media including sound, video and text. The project team are currently investigating support for audio and video standards.

Since metadata standards for botanical applications are still developing, the project team reviewed a range of general[3] and subject-specific[4] schema before finalising the core transitional metadata set for *eBot*. The resulting list of tags, which are suitable for current *eBot* functions and mappable to other schemas, will be refined over time as standards develop and the scope of the project expands. One development currently in progress is the mapping of the eBot descriptive metadata to ensure it aligns with the Australia's Virtual Herbarium (AVH) standard.[5]

The project team considered a range of international technical and preservation metadata initiatives undertaken by organisations such as the National Library of Australia (NLA), Library of Congress (LC) and the National Library of New Zealand (NLNZ).[6] Elements of the NLNZ preservation metadata schema were selected because it focuses on automatic collection of metadata and was designed to capture the most important information for digital preservation (Searle & Thompson, 2003). Schema development was informed by the work of other preservation agencies and its elements can be mapped to other technical schema. In addition to the NLNZ fields all other technical metadata that

can be extracted from the digital objects will be stored. This is an interim strategy while internationally agreed standards are developed.[7]

Conclusion

Continued collaboration amongst the botanical community in partnership with information and technology specialists will succesfully sow the seeds of a rich digital harvest. The opportunities offered by the digital mapping of data from dispersed sources, in diverse formats, will support global research community to gather, compare and use information in new ways. Common standards are the key to this development.

The digital environment is dynamic and the current standards are yet to be consolidated. Through cooperation and the use of flexible infrastructures it is still possible to build sustainable applications during the development phase. Current projects will act as testbeds and catalysts for the future development of sustainable standards. Deployment of emerging standards to practical applications allows useability testing and further refinement. The design of *eBot* and *eFlora* has been based on emerging standards, but with sufficient flexibility to be readily adapted to future enhancements.

Endnotes

[1] Including PNGtrees (Conn & Damas, 2005).

[2] Including LC (Library of Congress, 2006b), TASI (Technical Advisory Service for Images, 2006) and Western States (Western States Digital Standards Group, 2003).

[3] Including DC (Dublin Core Metadata Initiative, 2006) and MODS (Library of Congress, 2006a).

[4] Including ABCD (International Union of Biological Sciences Taxonomic Databases Working Group & International Council for Science: Committee on Data for Science and Technology, 2005) and work undertaken by SEABCIN (South East Asian Botanical Collections Information Network, 2006), and PIC (Plant Information Center, 2006).

[5] This schema is based on HISPID (Conn, 1996). This is a development by HISCOM (Herbarium Information Systems Committee, 2006). The HISPID standard forms the basis of the extensive Access to Biological Collections Data (ABCD) schema (International Union of Biological Sciences Taxonomic Databases Working Group & International Council for Science : Committee on Data for Science and Technology, 2005) currently being developed by TDWG and CODATA (International Council for Science : Committee on Data for Science and Technology, 2006).

[6] NLA Preservation Metadata for Digital Collections (National Library of Australia, 1999), NISO Metadata for Images in XML Schema (Library of Congress, MARC

Standards Office, & National Information Standards Organization, 2004) and the NLNZ Preservation Metadata Schema (National Library of New Zealand, 2006).

[7] Recent developments include publication by the APSR of a preservation metadata requirements statement (Lee, Clifton, & Langley, 2006).

References

Australian Biological Resources Study. (2004). *XML Schema for the Flora of Australia Online*. Retrieved October, 2006, from http://www.deh.gov.au/biodiversity/abrs/online-resources/flora/pubs/fl-australia-schema.html

Australian Biological Resources Study. (2006). *Flora of Australia Online*. Retrieved October, 2006, from http://www.deh.gov.au/biodiversity/abrs/online-resources/flora/index.html

Australian National Herbarium. (2006). *Australia's Virtual Herbarium (AVH)*. Retrieved October, 2006, from http://www.anbg.gov.au/avh/

Australian Partnership for Sustainable Repositories. (2006). *Australian Partnership for Sustainable Repositories (APSR)*. Retrieved October, 2006, from http://www.apsr.edu.au/

Carolin, R. C., & Tindale, M. D. (1994). *Flora of the Sydney Region*: Reed Books, Chatswood, N.S.W.

Conn, B. J. (1996). *Herbarium Information Standards and Protocols for Interchange of Data Version 3 (HISPID3)*. Retrieved October, 2006, from http://plantnet.rbgsyd.nsw.gov.au/HISCOM/HISPID/HISPID3/hispidright.html

Conn, B. J., Banka, R., & Lee, L. L. (2006). *Plants of Papua New Guinea (PNGplants)*. Retrieved October, 2006, from http://plantnet.rbgsyd.nsw.gov.au/PNGplants/

Conn, B. J., & Damas, K. (2005). *Guide to Trees of Papua New Guinea*. Retrieved October, 2006, from http://plantnet.rbgsyd.nsw.gov.au/PNGplants/PNGtrees/

Dublin Core Metadata Initiative. (2006). *Dublin Core Metadata Initiative (DCMI)*. Retrieved October, 2006, from http://dublincore.org/

Global Biodiversity Information Facility. (2006). *Global Biodiversity Invormation Facility (GBIF)*. Retrieved October, 2006, from http://www.gbif.org/

Herbarium Information Systems Committee. (2006). *Herbarium Information Systems Committee (HISCOM)*. Retrieved October, 2006, from http://plantnet.rbgsyd.nsw.gov.au/HISCOM/

International Council for Science : Committee on Data for Science and Technology. (2006). *Committee on Data for Science and Technology (CODATA)*. Retrieved October, 2006, from http://www.codata.org/

International Union of Biological Sciences Taxonomic Databases Working Group. (2006a). *International Union of Biological Sciences Taxonomic Databases Working Group*. Retrieved October, 2006, from http://www.nhm.ac.uk/hosted_sites/tdwg/

International Union of Biological Sciences Taxonomic Databases Working Group. (2006b). *Standards*. Retrieved October, 2006, from http://www.nhm.ac.uk/hosted_sites/tdwg/standrds.html

International Union of Biological Sciences Taxonomic Databases Working Group, & International Council for Science : Committee on Data for Science and Technology. (2005). *Access to Biological Collections Data Schema (ABCD)*. Retrieved October, 2006, from http://bgbm3.bgbm.fu-berlin.de/TDWG/CODATA/Schema/default.htm

Internet2. (2006). *Internet2 Middleware Initiative*. Retrieved October, 2006, from http://middleware.internet2.edu/

iSpheres. (2005). *iSpheres*. Retrieved October, 2006, from http://ispheres.org/

Lee, B., Clifton, G., & Langley, S. (2006). *PREMIS Requirement Statement Project Report*. Retrieved October, 2006, from http://www.apsr.edu.au/publications/presta/

Library of Congress. (2006a). *Metadata Object Description Schema (MODS)*. Retrieved October, 2006, from http://www.loc.gov/standards/mods/

Library of Congress. (2006b). *Sustainability of Digital Formats : Still Images*. Retrieved October, 2006, from http://www.digitalpreservation.gov/formats/content/still.shtml

Library of Congress, MARC Standards Office, & National Information Standards Organization. (2004). *NISO Metadata for Images in XML Schema (MIX)*. Retrieved October, 2006, from http://www.loc.gov/standards/mix/

Macquarie University. (2006). *Meta-Access Management System (MAMS)*. Retrieved October, 2006, from http://www.melcoe.mq.edu.au/projects/MAMS/

National Library of Australia. (1999). *Preservation Metadata for Digital Collections : Exposure Draft*. Retrieved October, 2006, from http://www.nla.gov.au/preserve/pmeta.html

National Library of New Zealand. (2006). *Introduction to the Metadata Standards Framework for National Library of New Zealand*. Retrieved October, 2006, from http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#meta

Plant Information Center. (2006). *Plant Information Center (PIC)*. Retrieved October, 2006, from http://www.ibiblio.org/pic/

Searle, S., & Thompson, D. (2003). Preservation Metadata : Pragmatic First Steps at the National Library of New Zealand [Electronic Version]. *D-Lib Magazine*, 9. Retrieved October 2006 from http://www.dlib.org/dlib/april03/thompson/04thompson.html.

South East Asian Botanical Collections Information Network. (2006). *South East Asian Botanical Collections Information Network (SEABCIN)*. Retrieved October, 2006, from http://www.seabcin.org/

Technical Advisory Service for Images. (2006). *Technical Advisory Service for Images (TASI)*. Retrieved October, 2006, from http://www.tasi.ac.uk/

Western States Digital Standards Group. (2003). *Western States Digital Imaging Best Practices*. Retrieved October, 2006, from http://www.cdpheritage.org/digital/scanning/documents/WSDIBP_v1.pdf