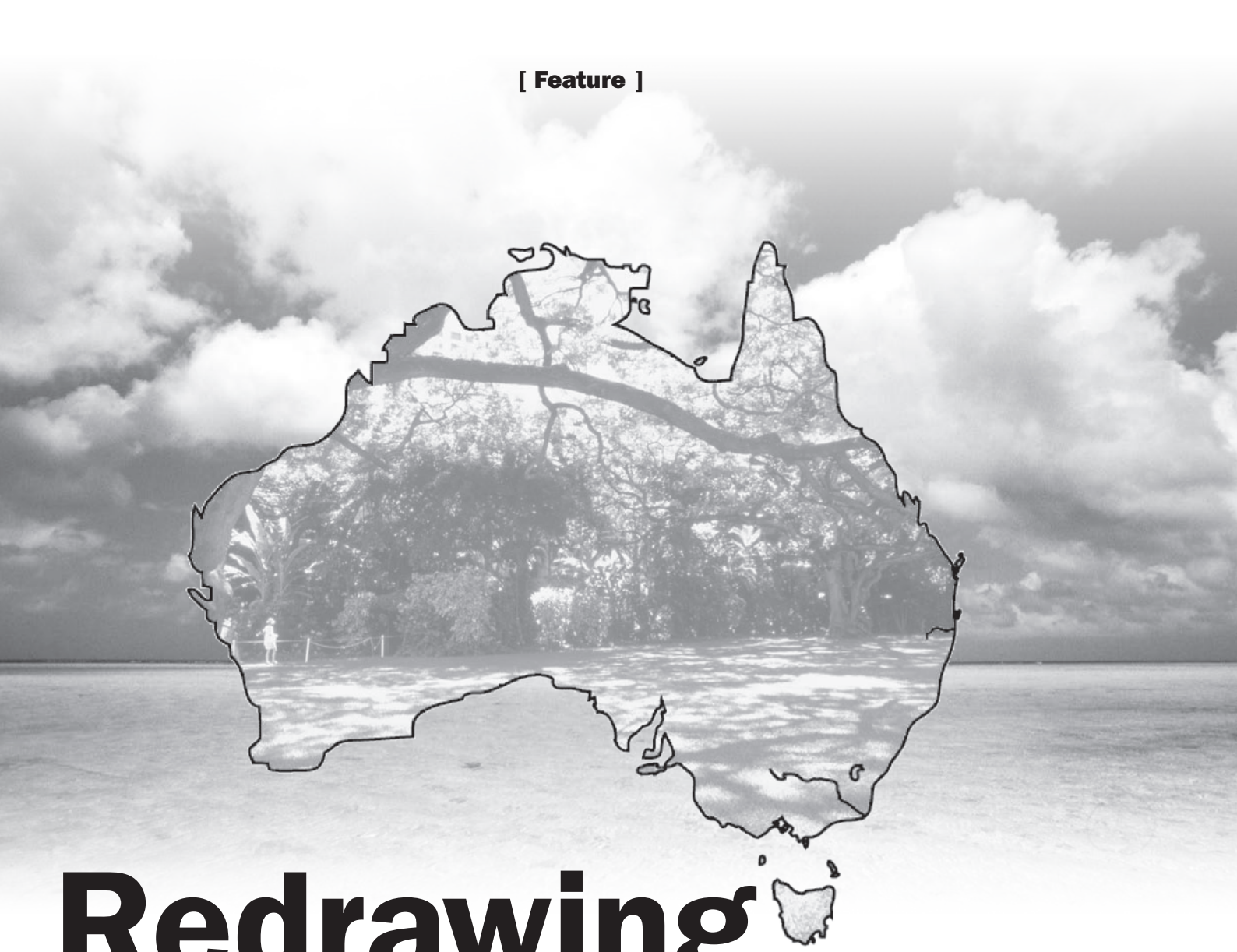

Redrawing the map

Donald Hobern takes a look at the challenges of managing biodiversity data

Many groups require better access to up-to-date and comprehensive information on biological organisms. Biosecurity officers need to identify specimens intercepted at borders. After identifying the organism, they need information on the biology, life cycle, host-prey relationships and habitat requirements of the species concerned, and often, information on how to control it.

Conservationists and land-use planners need to know which species occur in a given area, whether those species are threatened and how to conserve them. Other scientists requiring such information include those working in crop development, natural resource management, health and medicine, development of biomaterials and forensic science. The general public also wish to identify organisms and better understand their place in natural and modified ecosystems.

Much of this information already exists on the Internet, and more is available all the time: specimen databases from natural history collections and herbaria, field observations from ecologists and amateurs, DNA and protein databases, online identification tools, catalogs of scientific names, taxonomic and phylogenetic hierarchies, and countless specialist databases from different research areas. Large projects are busy scanning the historical literature and serving it online. More recent publications and journal articles are often immediately available in digital formats. Scientists and photographers are creating immense libraries of natural history images.

For the user, much of this information can be found using standard web search engines, but it is difficult to gain a comprehensive view of the available materials. One reason for this is the complex history of taxonomy and nomenclature. Millions of books and articles have been published on living organisms over several centuries and much of this literature remains relevant today. Over this period, scientific understanding has developed and names and classifications have changed. Many species have been known at different times under many different names. A user needs to be aware of these changes to find the full range of information for any given species.

Cicadetta sp., Canberra, Australia, December 2007
Photograph Donald Hobern

Rainbow Lorikeet (Trichoglossus haematodus), Queensland, Australia, May 2008
Photograph Wolf Wanjura

Another challenge is the complexity of the biological domain at all levels of organization from molecules through cells, organs, organisms, species and communities up to ecosystems. Each of these levels represents a rich matrix of inter-related concepts with additional complexity arising from the connections between levels.

Most significantly, the majority of existing information exists as unstructured text. Even when information has been entered into a structured database, there may be no available standards to ensure compatibility with information held elsewhere. In general, it takes significant domain knowledge to understand how to integrate information from different sources.

The Project

The Atlas of Living Australia (ALA) is a project funded under the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS). NCRIS is an initiative to fund a collaborative development of key research infrastructure to support Australian science. One of the thematic areas included within this initiative is known as Integrated Biological Systems. This comprises the ALA and two research networks:

- *The Australian Phenomics Network* (http://www.australianphenomics.org.au/), a network of mouse production, cryopreservation, phenotyping, documentation, distribution and databasing facilities;

- *The Australian Plant Phenomics Network* (http://www.plantphenomics.org.au/), a partnership established to provide a continuous record and analysis of key physiological parameters throughout the plant life cycle for a range of plant systems.

The ALA itself is a partnership between CSIRO, the Australian Museum, Museum Victoria, Queensland Museum, the Tasmanian Museum and Art Gallery, Southern Cross University, the University of Adelaide, the Department of Agriculture, Fisheries and Forestry, the Department of the Environment, Water, Heritage and the Arts, the Council of Heads of Australasian Herbaria, the Council of Heads of Australian Faunal Collections, the Council of Heads of Australian Entomological Collections and the Australian Microbial Resources Research Network. All of these partners will be contributing resources and content to the development of the ALA as an integrated source for biological information.

The key goal of the project is to develop a biodiversity data management system, which will bring together information from Australia's scientific reference collections and other custodians of biological information (including the two phenomics projects included within Integrated Biological Systems).

The ALA partners are contributing data from many of the country's most significant natural history collections, as well as other information such as images, fact sheets and diagnostic keys. The ALA is developing relationships with other bodies holding ecological and other observational data, and the project plans to offer tools and interfaces that will simplify the task of sharing data for all these groups.

The ALA is promoting free and open access to data, but will ensure that the infrastructure can also support more restrictive sharing of data sets within specified user communities.

The timing of the project allows it to build on the results from other related national and international data management activities. Within Australia, the *Australian Virtual Herbarium* (http://www.anbg.gov.au/avh/) has been highly successful in integrating data on the specimens held by Australian herbaria and making these accessible in an integrated form for use in planning and research. The *Online Zoological Collections of Australian Museums* (http://www.ozcam.gov.au/) is a similar data integration network for zoological specimens. The *Australian Plant Census* (http://www.anbg.gov.au/chah/apc/) and *Australian Faunal Directory* (http://www.environment.gov.au/biodiversity/abrs/online-resources/fauna/afd/) provide a taxonomic framework for the project. The *Taxonomy Research & Information Network* (http://www.taxonomy.org.au/) project is well-aligned with the ALA and will provide significant additional content. Several other projects funded or planned under NCRIS will also connect with the ALA, in particular the *Networked Biosecurity Framework* (http://www.ncris.dest.gov.au/capabilities/networked_biosecurity_framework.htm) and the *Terrestrial Ecosystem Research Network* (http://www.ncris.dest.gov.au/capabilities/tern/).

In the international arena, projects such as the *Global Biodiversity Information Facility* (GBIF, http://www.gbif.org/) and the *Ocean Biogeographic Information System* (OBIS, http://www.iobis.org) have already demonstrated the feasibility of providing integrated access to data held by large numbers of distributed specimen and observation databases. Software components

used by these projects provide a starting point for development of the ALA. The *Encyclopedia of Life* (EOL, http://www.eol.org/) has recently been established to develop informational pages for every species on earth and the ALA expects to work closely with the EOL in developing tools to organize content.

Products

The ALA plans to develop three main products within the first couple of years of its existence: a metadata repository, species information pages and a regional biodiversity atlas. These are intended to be valuable services in their own right but also to provide building blocks which can contribute to the subsequent development of web sites for particular user groups such as biosecurity officers or conservation planners.

Metadata repository

The most fundamental component of the ALA infrastructure will be a metadata repository. This will be an online catalog of data resources which relate to Australian biodiversity. The resources to be cataloged include online databases (natural history collections, ecological and observational data, molecular data, etc.), literature (journal articles, scanned publications, web pages, etc.), images (both major image repositories and individual images) and other multimedia resources, and online identification keys.

The metadata repository will hold descriptions of each data resource, as well as URLs and other access information and details of ownership and copyright. These basic descriptions will be enhanced by the addition of key words and other terms of relevance to different user groups. These terms will include the scientific names of organisms referenced by each resource, the geographic and temporal coverage of the data set, and gene names and other ontology terms. For tabular data, the definitions of the columns will also be included (and related to ontology terms where applicable). Some of these terms will be inserted manually, but many of them can be determined by scanning the contents of the source data.

The use of multiple complementary cataloging approaches will enable different groups of users to determine which data resources are relevant for a given purpose. The *Metadata Repository* will also help users to identify combinations of complementary data sets which could be brought together to answer more complex questions.

Species Information Pages

The metadata repository will provide a catalog of biodiversity information resources accessible to the ALA. This will serve as the basis for a web site offering species information pages, each of which will provide an organised overview of the resources relating to a given species.

These pages will include information on the names and classification for each organism, thumbnail images (linking to original image resources), an overview map (linking to the ALA Regional Biodiversity Atlas - see below) and links to other information resources, categorized as far as possible by major topic (biology, distribution, conservation, etc.). In all cases, the page will identify the data providers responsible for the content and users will have the opportunity to comment or supply additional information.

External web sites and tools will be able to access information from these species information pages and embed it as part of their own information services.

Regional biodiversity atlas

The species information pages will link users to the original provider for each data resource. However, there are some situations in which users need rapid access to fully integrated data. A key example is in the management of geospatial data. Hundreds of different data sets may contain occurrence records for a particular species, often in a wide range of different formats. This makes it very difficult for a user to retrieve all of the relevant data for visualization or analysis.

The ALA will therefore develop a regional biodiversity atla*s*, which will incorporate core data elements (species, locality, date, etc.) from every specimen and observation record shared through the atlas infrastructure (i.e. all occurrence data from resources registered in the metadata repository) and will make them available for easy use via mapping and analytical tools. It will link every record back to information on the associated resource in the metadata repository. This will allow users to understand the evidence for the given occurrence, the expertise of the parties recording the information, and any standardized methodology underlying the recording activity (atlasing projects, transects, long-term monitoring etc.).

The ALA will reuse open source software from the GBIF Data Portal (http://data.gbif.org/) to harvest data from Australian data resources. This will simplify subsequent exchange of data with other countries within the GBIF network. The ALA will also work closely with the BioMaps project (http://www.biomaps.net.au/biomaps/, an initiative of the Australian Museum and Rio Tinto) to develop appropriate mapping interfaces.

One major use for these data will be to present them through a flexible mapping interface (and through web services which can be used by GIS software). The ALA will therefore map Australian occurrence data against relevant geospatial data layers (e.g. geology, elevation, soils, vegetation, climate and land-use) and significant divisions of the continent and adjacent marine areas (e.g. local government areas, water catchment areas, protected areas, zip codes, marine areas).

A second, related use will be to produce up-to-date online reports on the organisms recorded from each division of the continent or ocean. These reports will show what evidence exists for each species occurring within the division (specimens, observations, literature) and the most recent recording date. Such reports can feed directly into land-use planning.

Summary

The ALA is only just starting its work of providing services for better management of information on Australia's biodiversity, but is well-positioned to build on past developments and is keen to share ideas and resources with related projects around the region. ■

Donald Hobern spent 16 years working as a software developer and web architect for IBM before working to develop solutions for the Global Biodiversity Information Facility in Copenhagen, Denmark. He joined CSIRO in December, 2007 to take up the role of Director for the Atlas of Living Australia project.

Photograph Ciprian Vizitiu