ANNUAL BUSINESS PLAN 2008-2009

for

The National Collaborative Research Infrastructure Strategy's Research Capability

known as

5.2 Integrated Biological Systems: 5.2.3 Biological Collections –

The Atlas of Living Australia

June 2008

Contents

1. Executive Summary

The ALA is now moving from planning into the implementation phase. Core staff have been recruited and an informational web site (with email newsletter) is in place.

The ALA is actively exploring the range of potential users and requirements for the ALA through a detailed user needs analysis to be completed in October, after a workshop at the TDWG annual conference. This user needs analysis is expected to provide core direction and priorities for many areas of ALA activity.

The ALA has identified a suite of core services which should be developed and will begin work on these in 2008-2009. Most of these services are free-standing tools with value to the wider community in their own right, but the combination of them will provide the foundations for the ALA to build tools to meet the needs of different user groups.

The ALA has been active developing relationships with other NCRIS capabilities, government departments, Australian and international NGOs, research groups and amateur bodies. Additional funding has been secured from the Platforms for Collaboration NeAT programme to assist the ALA with development of some of its core services, and the ALA expects to be able to organise collaborations with many of these other bodies.

The budget for the project has been revised to reflect the changed circumstances arising particularly from the delayed start of work and to reflect increased understanding of the wider community's needs for services which the ALA can deliver.

It is expected that the ALA will finish 2008-2009 with a visible presence and a range of tools of value to Australian researchers.

2. Project Contents

2.1 Status of ALA Project

The project is now well under way. A small core team has been recruited to coordinate activity from Canberra. The Project Director and Executive Officer have been in place since January 2008. Contracts are being developed for the Metadata Curator, Java Developer and Technical Architect positions to be filled early in 2008-2009.

The ALA web site was launched in May 2008 and is serving a range of information on the project partners and progress with implementation. The first edition of the ALA newsletter links to the web site and has been widely distributed by email. This has led to a significant number of contacts, most of whom have been polled for the ALA user needs analysis.

The ALA Management Committee has started a branding activity to develop a logo and look-and-feel for the project and expects to finish this process by August 2009, at which point the web site will be restyled.

The ALA has completed both an initial review of biodiversity data resources and a review of tools and software components which may contribute to the development of the Atlas. The tools review was developed by a team from the Australian Museum as an online resource which will subsequently be integrated into the ALA web site. A further contract is under way with the Australian Museum to carry out the ALA user needs analysis and again to make this available as a dynamic online resource. The summary work plan is included as Attachment 4.

A contract has been placed with ABRS to provide web services exposing data from the Australian Faunal Directory and Australian Plant Census. These resources will provide the core organisational framework for Australian biodiversity data. The summary work plan is attached as Attachment 5.

The ALA has made progress with the other NCRIS 5.2 Integrated Biological Systems capabilities in planning expenditure of NCRIS funds allocated for integration between the three capabilities. Job advertisements have been prepared for recruitment of bioinformaticians to be based with APN at ANU in Canberra and with APPF at ACPFG in Adelaide. These bioinformaticians will be responsible for identifying data standards and best practices for data storage to enable the ALA to provide suitable data discovery and management services on behalf of APF and APPF.

The NCRIS 5.16 Platforms for Collaboration NeAT programme has allocated up to $1,000,000 over 2.5 years to support development of a Metadata Repository and Annotation Services for the ALA (*Data Integration and Annotation Services for Biodiversity* – DIAS-B*)*, with the goal of developing models and software which can be reused within other NCRIS capabilities. CSIRO is proposing to provide further resources under its eSIM programme to contribute to the development of this Metadata Repository. The project plan for these activities is under development but the draft work plans are attached as Attachment 6.

The ALA has had contact with an extremely wide range of government departments, research groups and agencies and has developed or initiated relationships with a number of significant partners or potential partners, in particular:

- Global Biodiversity Information Facility (GBIF) – the ALA will manage data integration for Australian data providers to GBIF (i.e. ALA will assume many of the responsibilities of the current ABIF node). The ALA will reuse software components from GBIF to support this goal.
- Encyclopedia of Life (EOL) – the ALA has developed a Memorandum of Cooperation with EOL. This will be signed early in 2008-2009. EOL is also interested in the ALA being formally recognised as a cornerstone "institution" within EOL (i.e. one of the major components contributing to the development of EOL – there are currently six EOL cornerstone institutions). Discussions are continuing about the practicalities and feasibility of such an arrangement. EOL and the ALA also plan to co-fund development of Online Identification Services (diagnostic keys to species, etc.).
- Taxonomic Databases Working Group (TDWG) – TDWG is holding its annual conference in Perth in October 2008 and has chosen to devote a full day (Wednesday October 22) to presentations and discussions relating to the ALA. This day (and training sessions arranged by TDWG on the previous evening and following morning) will therefore be treated as a core ALA meeting. Those involved in the ALA will be encouraged to attend these discussions and will consequently also have the opportunity to meet with a wide range of international representatives from related projects (GBIF, EOL, OBIS, LTER, etc.). A workshop is also planned at the conference to enable biodiversity projects to share requirements and plans for metadata management.
- EarthWatch – the ALA is developing a Memorandum of Cooperation with EarthWatch to collaborate in developing tools to collect and manage observational data and to pilot their use with members of the public. The ALA already needs to develop or source much of the software required to make this possible to address its own data management needs for such data. EarthWatch is planning its ClimateWatch project as a pilot activity for engaging schools and members of the public in recording natural history observations. These two goals are highly complementary and the ALA Director is now serving as a member of the ClimateWatch

technical advisory group. GBIF New Zealand have also indicated an interest in collaborating with the ALA in developing infrastructure for managing observational data from the general public.

- Murray-Darling Basin Commission (MDBC) – the ALA has met twice with groups within the MDBC and the two organisations identified two key areas in which to begin collaboration. MDBC requires better access to expert information on the names and classification of organisms found in the basin in order to improve management of all the MDBC's biological data. In addition the MDBC is developing informational pages on many of these organisms and wishes to be able to reuse content wherever possible. The ALA regards MDBC as a highly suitable partner for trialling many early services.
- Birds Australia (BA) – the ALA has met with BA to discuss options for integrating bird observational data within the ALA framework and for collaboration in development of infrastructure to collect and manage observational data. Similar discussions have also been held at a more local level with the Canberra Ornithologists Group.
- Taxonomic Research Information Network (TRIN, the CERF project formerly known as the Taxonomy Hub) – TRIN has software developers based at the Australian National Herbarium in Canberra and working on tools and portals to support taxonomic working groups within Australia. The ALA plans to collaborate closely with this team to avoid duplication of effort and to ensure compatibility of software products.
- TERN and ABIN – the ALA has met with representatives from these developing NCRIS capability areas and is developing understanding of possible ways for the ALA to meet some of the infrastructure needs for these capabilities.

2.2   Project Outlook

The ALA user needs analysis (see Attachment 4) is under way and is already receiving input from many different groups. After the preliminary round of input, a core set of around 10-12 scenarios will be identified for further investigation. The Australian Museum team will work with data providers and users to understand how biodiversity information is currently sourced and used within these scenarios and will document the information flow as an online resource. The team will select a smaller number of these detailed scenarios for presentation and discussion at the ALA day to be held during the TDWG conference in October, with the aim of engaging the broader Australian and international community in identifying information resources, tools and collaborators with which the ALA can address these user needs.

The ALA Management Committee will make decisions based on the final results from the user needs analysis, along with the previous tools survey and the survey of biodiversity data resources, to establish priorities for content and tools within the rest of the funded ALA period.

The ALA already recognises several core tasks which will certainly need to be performed, regardless of the specific recommendations and priorities arising from this process. Activity will begin immediately in these areas:

- Development of web services to enable the AFD and APC to serve as a taxonomic backbone for Australian data (see Attachment 5)
- Deployment of the GBIF data portal software as an interim framework to manage a central cache of specimen and observation data and to relate these data to the AFD/APC taxonomy
- Development of an interim database to store metadata for ALA data resources while work proceeds on the NeAT DIAS-B project (see Attachment 6)
- Review of metadata standards, ontologies and controlled vocabularies of relevance to the ALA and other IBS capabilities (to be carried out by the ALA Metadata Curator and by the APF and

APPF Bioinformaticians as these are recruited, and to include input from the metadata workshop to be held at the TDWG conference)

The NeAT DIAS-B project is expected to start early in 2008-2009 and to contract two developers to work on the ALA Metadata Repository and two developers to work on the ALA Data Annotation Services. The draft work plans are attached as Attachment 6. A workshop will be held at the start of the project to review requirements for both areas and to revise the project plan.

The current ALA web site (http://www.ala.org.au/) will be developed to incorporate web databases such as the ALA user needs analysis and tools survey, to include online collaboration and discussion areas, and to serve as the access point for the ALA biodiversity data services as these are developed. The ALA aims to bring together as many of these services and as much content as possible for a more public launch of the web site on or around 12 February 2009 (the 200th anniversary of Charles Darwin's birth).

Early in 2008-2009, the ALA Management Committee will review the results from the first round of content development using participant cash and make preliminary recommendations for participant activity during the year. When the user needs analysis is completed, the Management Committee will establish priorities for the rest of the year and for the remainder of the funded ALA period. Consequently this Business Plan does not include any specific commitments from participants on the activities to be carried out during 2008-2009 using participant cash.

The ALA is working to establish and clarify the relationship between the Atlas and existing Australian biodiversity informatics projects. In general it is not intended that the ALA should replace these activities but should consolidate the services and infrastructure required to enable them to succeed. The following relationships will be particularly significant:

- Australian Virtual Herbarium (AVH), Online Zoological Collections of Australian Museums (OZCAM), Australian Plant Protection Database (APPD) and Australian Microbial Research Information Network (AMRIN) – the ALA will work with these existing network projects to ensure that ALA software developments (especially Data Provision Services) are fully supportive of the needs of each project; ALA activities should reinforce the development and identity of these networks, which will be key contributors of content to the ALA.
- Australian Biological Resources Study (ABRS) – the ALA will work with ABRS to develop web delivery mechanisms (especially web services) for data sets produced through ABRS activities
- Australian Biodiversity Information Facility (ABIF) – the ALA will assume responsibility for the integration of Australian data resources with GBIF, and will work with the Australian GBIF delegation to ensure that mechanisms and processes are in place for making decisions on the content Australia should offer at this level.
- Ocean Biogeographic Information System (OBIS) – many data sets of relevance to the ALA also relate to OBIS; the ALA will coordinate activity with OBIS to ensure compatibility and efficient use of resources; this relates in particular to Data Provision Services and the Metadata Repository, but linkages are likely for all the ALA components identified below.

As the components described under *Research Infrastructure* below are developed, the ALA will deploy them through the ALA portal and will then identify partners with whom to develop more specifically targeted web sites for different communities. Such web sites will organise combinations of these and other components to provide functions required by a particular group of users.
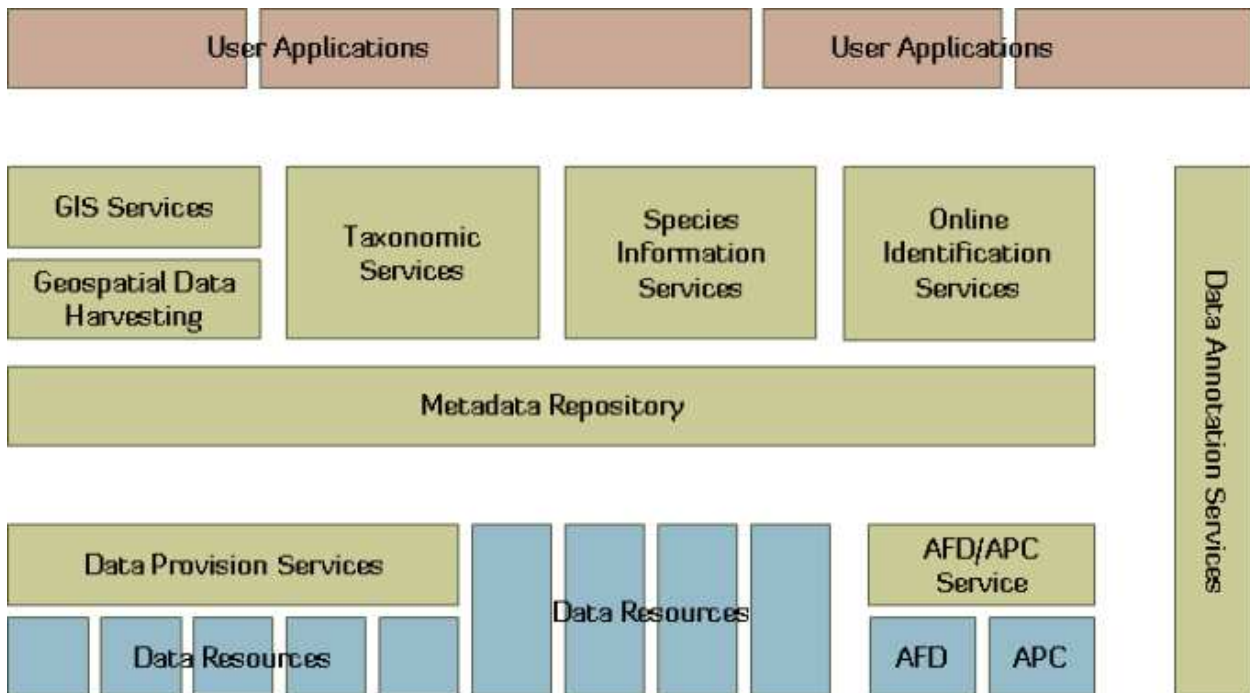
2.3   Research Infrastructure

Figure 1 - Overview of ALA services

During 2008-2009, the ALA will aim to develop the core systems and services needed for managing biodiversity information. These services will provide the foundation for future developments, including collaborative development of portals to support specific user groups.

Figure 1 provides an overview of the key components to be delivered as part of this core. Data resources have been represented in blue. Components developed by the ALA using NCRIS funds have been represented in green.

The following sections provide more detail on the ALA components in this figure.

2.3.1 Data Provision Services

The ALA will integrate a wide range of data resources which have already been shared online and which can be accessed without any additional work on the part of the data providers.

However many institutions and individual researchers are hindered from making biodiversity data available by the complexity of the existing tools. Some of these tools require significant understanding of database and web technologies. In all cases they require the data provider to have access to a web server and permission to run executable code on such a server. This is an insurmountable technical threshold for many research groups.

The ALA will therefore invest resources during 2008-2009 and beyond in improving the usability of existing tools and in developing alternative means for data providers to share their data. The key elements in this activity will be:

1. Developing installable software bundles including web server and database components and all prerequisite software necessary to run standards-compliant wrapper software. (At present data providers are often obliged to download and even compile some of these prerequisite components.) The following features are desirable:

a. A fully configured web server implementation (without providers being required to find prerequisite packages themselves) for Windows and Linux.

b. A simple default database (e.g. MySQL) with a basic predefined structure into which data can easily be imported from spreadsheets, etc.

c. Interfaces to import data into the default database.

d. Preconfigured TAPIR (and any other required interfaces) offering Darwin Core and ABCD interfaces against the default database (developed in conjunction with AVH and OZCAM to ensure that they also support their needs).

e. Wizard interfaces for mapping other databases and exposing their data using the same Darwin Core and ABCD interfaces.

f. Basic HTML user interface for users to browse or search the data – with simple configuration to apply basic styling, logos, etc. (in other words to allow the installation to serve as part or all of a web site for the provider).

g. Google Maps (or similar) map interface integrated into the HTML user interface (including clear instructions on getting the necessary key from Google to use their maps).

h. Gallery (or similar, open source) image management integrated into database to make it easy to associate images with specimen data

i. Documentation and tutorial materials for installing and using this software

2. Enhancing ALA Metadata Repository and Data Harvesting components to support harvesting of data directly from flat-file, spreadsheet formats (CSV dumps, Excel, etc.). This will allow data providers to share data simply by placing a file on a web server and registering associated metadata in the ALA Metadata Repository.

3. Exploring collaborative opportunities for the ALA and third-party organisations to develop online repositories to which data providers may upload biodiversity data. Such a repository should give users of all kinds (from lab researchers to amateur naturalists) to upload and view data (especially ecological and observational data sets) through a web browser interface. The interface should provide all of the features listed above for the installable software bundles and should give users control over the visibility and publication of data.

The ALA will collaborate with interested partners itn meeting these goals. Potential partners include:

- CHAFC - ABRS has recently funded some work to web-enable databases at several Australian natural history collections
- CHAH – several herbaria need software to establish a dynamic connection with the AVH
- GBIF shares most of these requirements and will also be allocating development resources to this area during the time period
- Birds Australia – the ALA has had initial discussions with Birds Australia, during which it was recognised that amateur birdwatchers need good online tools for uploading observations
- EarthWatch – early conversations with EarthWatch have focused on web tools to support the development of citizen scientist networks – this is an opportunity for the ALA to promote the use of web infrastructure to integrate observational data

The ALA aims to work with these and other partners to share development resources and to ensure

that software tools are easy to use and genuinely interoperable.

The ALA will contract the equivalent of a full-time developer throughout 2008-2009 to work on developing and bundling software to address these needs.

2.3.2 AFD/APC Service

Management of biodiversity data depends on the existence of an authoritative taxonomic framework to which any data item including a scientific name can be related. The ideal taxonomic framework would contain all published scientific names, all common names, an up-to-date classification based on the best available taxonomy and indicating whether each name is accepted as the current name for a species or considered to be some kind of synonym, and a set of mappings between this classification and other checklists of importance to the community (e.g. classifications used in state floras, red lists, CITES lists).

There is no online source including all of this information. Indeed relevant information is scattered through 250 years of printed literature. However the ALA will be able to manage the vast majority of available data by making use of the Australian Plant Census (APC) and the Australian Faunal Directory (AFD) as the most complete and current taxonomic checklists for Australian plant and animal species. Neither of these resources includes all relevant taxa, but work continues through ABRS to address the gaps. In the mean time, the ALA expects to make use of other resources to help to organise data on the remaining species (see *0 When the* ALA has developed a cache for geospatial data relating to Australian organisms (see Geospatial Data Harvesting), it aims to provide services a range of services allowing these data to be explored and integrated within geospatial analyses.

The GIS Services component will offer the following services:

- Mapping tools allowing biodiversity data for any species or multiple species to be combined with other geospatial data layers (e.g. geology, elevation, soils, vegetation, climate and land-use) as well as relevant divisions of the continent and adjacent marine areas (e.g. local government areas, water catchment areas, protected areas)

- Links to tools for modelling and analysis based on the available biodiversity data (e.g. niche modelling, analysis of gaps in recording effort)

- Web services for incorporating these data in other GIS systems, including OGC Web Feature Service (WFS) and Web Coverage Service (WCS) interfaces

- Reports for each division (local government area, water catchment area, protected area, zip code, marine area) on the organisms recorded from the division, including links to the original data sources, information on the basis for each record (specimen, observation, literature) and the most recent recording date, etc.

The *BioMaps* project already addresses some of these requirements and the ALA expects to work closely with the Australian Museum and Rio Tinto to coordinate activity in this area. The ALA has reserved funds to engage a GIS Services developer via contract. More work is required in this area to identify requirements more clearly and to develop relationships with partners able to make available the data layers or services necessary for visualising and analysing biodiversity data. As TERN becomes active, it will be a major partner in this area.

 Taxonomic Services below).

ABRS and the Australian National Botanic Garden have been funding the development of online web services to make the AFD and APC accessible to web users, including the ALA. No funding is available for the completion of this work during 2008-2009, so the ALA will provide funds for two

developers for six months to secure online access to this essential resource.

It is expected that both AFD and APC will subsequently serve as sources of a wide range of additional species information to the ALA.

2.3.3 Metadata Repository

The most fundamental component of the ALA infrastructure will be a Metadata Repository which can not only document the existence of biodiversity data resources but also provide a flexible and scalable framework for managing information which can assist users of all types in finding relevant information.

The ALA has received additional funding of up to $1,000,000 in the period 2008-2011 from the *NCRIS 5.16 Platforms for Collaboration* capability's NeAT programme to assist in the development of its Metadata Repository and Data Annotation Services. It is intended that the ALA Metadata Repository will provide an environment in which tools and best practices can be developed which may subsequently be applied to other NCRIS capabilities.

The following functional requirements have been defined for the ALA Metadata Repository:

1. Storage of metadata documents, including:
   a. Support for Dublin Core
   b. Support for ISO 11179
   c. Support for arbitrary RDF properties
   d. Support for tagging metadata with ontology (OWL/OBO) terms

2. Metadata documents to describe:
   a. Online databases and web services
   b. Text documents (including PDF, Word, etc.)
   c. Images and other multimedia resources

3. Harvesting of metadata from other repositories
   a. OAI-PMH.

4. Pluggable framework for tagging metadata with terms derived from data content
   a. Support for terms derived from occurrence data
   b. Support for terms derived from taxonomic data
   c. Support for terms derived from detecting scientific names in text documents (see uBio services)
   d. Support for terms derived from image metadata (including XMP metadata)

5. Publication of metadata to other tools and repositories
   a. OAI-PMH (offering ontology terms as OAI-PMH Sets?)
   b. Other SOAP and REST services as required

6. Data provider interface
   a. Register/update/delete data provider
   b. Register/update/delete database/service/document/image from data provider
   c. Register/update/delete OAI-PMH feed from data provider
   d. Accept/reject annotations

7. Administrator interface
   a. Annotate metadata documents with arbitrary RDF properties and ontology terms
   b. Accept/reject annotations

8. End-user interface
    a. Full-text search
    b. Browse by data provider
    c. Browse by ontology terms
    d. Faceted search via multiple ontologies
    e. Propose annotations to metadata documents with arbitrary RDF properties and ontology terms

9. Access control
    a. Integration with Shibboleth/PKI (to exploit AAF infrastructure and services)
    b. AAF authenticated access to data provider, administrator and end-user interfaces
    c. AAF-mediated restrictions on visibility for some metadata documents (possible – may not be necessary)

10. Wider compatibility (may be ensured by other requirements)
    a. Compatibility with requirements within other NCRIS capabilities and within Platforms for Collaboration
    b. Compatibility with other developing international biodiversity metadata repositories from projects such as GBIF, EOL, OBIS, and LTER

The ALA expects to use the NeAT funds to recruit two developers based at the CSIRO ICT Centre in Canberra to work on the development of the Metadata Repository. CSIRO IM&T has also identified resources it wishes to allocate to this development activity. See Attachment 6 for more details.

2.3.4 Data Annotation Services

The ALA is a project to support integration of data from a wide range of different sources. This activity can be significantly enhanced if users have the opportunity to store comments of various kinds on data items. Examples could include

- Plain text annotations providing comments or proposed corrections for any data item

- Structured annotations proposing corrections for data items with well-known structures and formats

- Annotations providings links to other data items or vocabulary terms

- Responses from data providers or other users to any annotation

The ALA has received additional funding in the period 2008-2010 from the NCRIS Platforms for Collaboration capability's NeAT programme to assist in the development of its Metadata Repository and Data Annotation Services.

The aim behind such services is to provide a common and consistent set of services for storing and accessing annotations. Each annotation will consist of a block of text or structured data and will reference the original data item via a globally unique identifier which reliably allows the item to be accessed again. Annotations will be stored in a central service and can be retrieved by user applications and tools (including web sites) by supplying the identifier for the original data item.

Uses for such a service include:

- Providing a consistent model for associating any comment with any data item and for managing threads of such comments

- Allowing users (or automated data validation tools) to propose structured corrections to a data item which appears to be in error, and allowing other users (or the original data provider) to decide whether to keep the original values or to adopt the corrected values in their own analyses, etc.

- Allowing users to store additional information about a particular species. For example a user could store a small piece of structured data indicating that a particular species (identified via a globally unique identifier) feeds upon another species (also identified via a globally unique identifier). Similarly a user could make a link between a species and properties in an ontology (e.g. to provide descriptive data for use by Online Identification Services).

The ALA will use NeAT funds to recruit a developer based at the University of Queensland to work on the development of Annotation Services and will use NCRIS funds to recruit a second such developer. See Attachment 6 for more details.

2.3.5   Geospatial Data Harvesting

The ALA will manage access to many classes of information by providing users with links connecting to the original provider web sites. Geospatial data are a special case because users typically require the ability rapidly to search and view all available data on the occurrence of a particular organism or the biodiversity present in a particular area, regardless of how many different data resources may contribute to these views.

For this reason, the ALA will maintain a central cache of occurrence data and will seek to populate this cache with a core subset of fields from every specimen and observation record shared through the Atlas infrastructure (i.e. all occurrence data from resources registered in the ALA Metadata Repository).

This cache should accommodate occurrence data from any relevant sources and include links to metadata allowing users and tools to determine whether particular records are suitable for use in a particular context. The relevant sources include specimen-based data from natural history collections and herbaria, observational data from ecological field work and recorder networks, and potentially information held in the literature. In each case the metadata should allow a user to understand the basis for the record, the expertise of the parties recording the information, and any standardised methodology underlying the recording activity (atlasing projects, transects, long-term monitoring etc.) since such methodologies can provide a framework for interpreting the statistical significance of the data from each resource.

GBIF has developed open source software for building and maintaining such a cache and uses it to operate its global cache of species occurrence data (see the GBIF Data Portal). The ALA intends to reuse this code to harvest data from Australian data resources and to make this cache available to GBIF as an efficient way to integrate Australian data into the global cache. At the same time, the ALA will retrieve data on Australian biodiversity harvested by GBIF from data providers in other countries, and will integrate these into the ALA cache.

This activity will be handled by the ALA Technical Architect and Java Developer.

2.3.6   GIS Services

When the ALA has developed a cache for geospatial data relating to Australian organisms (see Geospatial Data Harvesting), it aims to provide services a range of services allowing these data to be explored and integrated within geospatial analyses.

The GIS Services component will offer the following services:

- Mapping tools allowing biodiversity data for any species or multiple species to be combined with other geospatial data layers (e.g. geology, elevation, soils, vegetation, climate and land-use) as well as relevant divisions of the continent and adjacent marine areas (e.g. local government areas, water catchment areas, protected areas)

- Links to tools for modelling and analysis based on the available biodiversity data (e.g. niche modelling, analysis of gaps in recording effort)

- Web services for incorporating these data in other GIS systems, including OGC Web Feature Service (WFS) and Web Coverage Service (WCS) interfaces

- Reports for each division (local government area, water catchment area, protected area, zip code, marine area) on the organisms recorded from the division, including links to the original data sources, information on the basis for each record (specimen, observation, literature) and the most recent recording date, etc.

The *BioMaps* project already addresses some of these requirements and the ALA expects to work closely with the Australian Museum and Rio Tinto to coordinate activity in this area. The ALA has reserved funds to engage a GIS Services developer via contract. More work is required in this area to identify requirements more clearly and to develop relationships with partners able to make available the data layers or services necessary for visualising and analysing biodiversity data. As TERN becomes active, it will be a major partner in this area.

2.3.7 Taxonomic Services

The ALA needs to support the integration of data resources which may vary greatly in their use of scientific names (and common names) and classifications. Some resources will include scientific names no longer in use (e.g. because they are regarded as synonyms of earlier names, or the species has been moved to a different genus, or the species has been combined into another species) or will follow a taxonomic hierarchy which differs from that accepted by other resources for the same species.

The ALA will use the names and classification shared through the AFD/APC Service as the primary taxonomic structure to which it will relate all Australian biodiversity data (although at present neither AFD nor APC is complete).

To make this possible, the ALA will need tools to:

- Determine whether a species or higher taxon exists in the AFD or APC to which each data item can be related (even if there are differences in classification or if the AFD or APC consider the supplied name to be a synonym)

- Otherwise, determine the best point to which to connect data items for species not included in the AFD or APC hierarchy (e.g. to connect a new species to an existing genus, or a family to an order) - this is important to allow the ALA to give overviews of all data relating to a higher taxon

- Link alternative taxonomies and additional sources of data on the scientific and common names and classification for Australian and adventive species

- Offer web services for users and web sites to be able to explore the names and classification of Australian organisms and to integrate this information into their own systems

- Allow users to browse or search for data using either the AFD/APC taxonomy or another classification or checklist to organise the information - for example to find data using a red list, CITES list or state checklist

The GBIF Data Portal software which the ALA intends to reuse for its Geospatial Data Harvesting already includes code to address many of these needs. The ALA therefore expects to use this code as the foundation for developing its Taxonomic Services.

The ALA Technical Architect and Java Developer will coordinate this activity.

GBIF New Zealand has expressed interest in collaborating around the development, deployment and integration of taxonomic services (in conjunction with their planned New Zealand Organisms Register). The ALA Taxonomic Services will also feed into the Global Names Architecture (see http://docs.google.com/View?docid=dm3q64p_106gjd66pfp), an international move led by GBIF and EOL to develop compatibility between taxonomic and nomenclatural data services and to develop tools and services based on these data resources.

2.3.8 Species Information Services

The ALA Metadata Repository will hold information on all biodiversity information resources accessible to the ALA and will seek to determine which Australian species are represented or included within each resource.

The ALA will build tools and services to present a species-oriented view into this collection of information. These services will include:

- A web page for every species (and higher taxon) providing organised links to relevant resources, including:

    o Information on names and classification via the ALA Taxonomic Services

    o Thumbnail images linking to original image resources

    o Overview map linking to ALA GIS Services

    o Links to other information resources, categorised as far as possible by major topic (biology, distribution, conservation, etc. - see the TDWG Species Profile Model)

    o Information in all cases on the data providers responsible for the content

    o Links to allow users to comment or supply additional information via the ALA Data Annotation Services

- Web services to allow external users and web sites to retrieve lists of links by species and topic and to embed this content into their information services

The ALA Technical Architect and Java Developer will perform this activity, potentially with additional support secured via use of contingency funds.

2.3.9 Online Identification Services

Users are normally only able to retrieve information for organisms which have already been identified. Identification is often a difficult process, and relevant materials are lacking for many Australian species.

A significant number of relevant tools and information resources are already accessible on the web for use in identification, and many more will become available over time (including resources being made available through the Atlas by its participants). These resources include dynamic online keys and text keys to separate groups of related species, digital versions (PDFs, HTML, etc.) of taxonomic or identification literature and images and other multimedia items. Information on distribution and biology can also be highly relevant.

The ALA therefore intends to work with other projects (e.g. IdentifyLife, EOL and KeyToNature) to assist users in discovering relevant identification resources and to provide user interfaces which integrate these resources in ways that allow users efficiently to reach the best possible identification. The ALA Online Identification Services could ultimately include:

- Tools to find available online identification tools and materials for any species or group

- Dynamic online interfaces to structured descriptive data resources, allowing users to identify organisms by selecting characters exhibited by an organism

- Access to existing sets of characters which have been used to describe different groups of organisms (allowing these characteristics to be reused in other identification tools)

- Interfaces to ALA Data Annotation Services allowing users to associate a species with appropriate characters (i.e. to support community-driven development of identification data)

- Tools to aggregate character-based data from different sources to populate identification interfaces for a list of species

- User interfaces to give quick access to relevant materials for use in cases where character-based data are not available (e.g. providing a synoptic overview of images, links to online literature resources and distribution data)

The ALA plans to contribute resources in 2008-2009 to joint development of Online Identification Services with IdentifyLife and EOL.

2.3.10 Hardware and networking

The ALA has established its web site (http://www.ala.org.au/) on web servers operated by CSIRO IM&T. During 2008-2009, this site will be moved to a cluster of virtual Linux hosts located in a CSIRO server farm in Melbourne. The ALA intends subsequently to replicate the site to a hosted environment in Perth (and potentially other Australian locations to ensure acceptable performance).

http://www.ala.org.au/ will serve as the home for ALA communications materials and for the suite of core data discovery services (Metadata Repository, Geospatial Data Harvesting, GIS Services, Taxonomic Services, Species Information Services, Online Identification Services, and potentially Data Annotation Services). All of these resources and services will be replicated as required. The underlying data resources will be managed by their providers and are in general unlikely to be replicated in this way. The core discovery services are intended to provide rapid responses for the vast majority of basic user requests and therefore to minimise the impact to users arising from the highly distributed nature of biodiversity data.

The ALA will also work with the APF and APPF to explore the most appropriate models to address their needs for large-scale long-term storage. It is expected that ARCS will meet much or all of this requirement and that the ALA will provide the infrastructure for cataloguing and managing discovery of APF and APPF data resources.

2.4   Governance

The ALA Management Committee expects to meet 4 times during 2008-2009, mainly by teleconference but with at least one meeting a year bringing together as many members as possible. The Committee receives regular information from the ALA staff and makes many decisions between formal meetings.

The Scoping Group receives regular updates and requests for input from the ALA staff, and has provided input into the planning of the ALA. The Management Committee is reviewing the operation and composition of the Scoping Group to ensure that it is able to provide timely technical direction to the project.

The ALA Key Performance Indicators document specifies a number of targets for 2008-2009 and also an external review of ALA services as they stand at the end of the period. The ALA Management Committee will be responsible for selecting reviewers and defining scope for their activity.

The ALA has proposed that the ALA Director should represent the ALA on the Integrated Biological Systems Steering Committee and has nominated another member to serve on this committee.

The ALA Management Committee is working to establish Memoranda of Cooperation between the ALA and the peak bodies for Australian natural history collections which are not covered by the ALA Participation Agreement:

- Council of Heads of Faunal Collections (CHAFC)
- Council of Heads of Entomological Collections (CHAEC)
- Australian Microbial Resources Research Network (AMRRN)

The ALA Management Committee is also developing a Communication Plan for ALA information materials to ensure that documents and discussions are disseminated to all relevant stakeholders.

2.5   Promotion

The ALA has already established an interim web site at http://www.ala.org.au/ (hosted on CSIRO infrastructure) for communication about the project's goals and progress. This site is a simple collection of web pages and does not support dynamic functions.

Early in 2008-2009, the site will be replaced with a new version with a dedicated web server. This version will incorporate wiki-based discussion areas and support for mailing lists and will integrate relevant tools and databases (including the ALA tool survey and the ALA user needs analysis) as they become available.

An article on the ALA has appeared in the *Canberra Times* and a short article has been written for the *Asia Pacific Biotech News.*

2.6   Access and Pricing

There are no access and pricing issues associated with the Atlas of Living Australia. The vision for the Atlas is to the greatest extent possible to provide free and open access to information. As the Atlas proceeds, it is likely that the same infrastructure will also serve for more restrictive point-to-point sharing of restricted data between data owners and authorised users. This scenario is however likely to result in reduced functionality and interoperability for the data resources so secured.

An Intellectual Property Rights Policy has been developed and was included as Attachment 5 to the 2007-2008 Business Plan. The Science Commons (http://sciencecommons.org/) has recently produced

a set of recommendations relating to the sharing of open access data. See the *Protocol for Implementing Open Access Data* (http://sciencecommons.org/projects/publishing/open-access-data-protocol). The ALA will explore the implications of this document and determine whether it is appropriate to revise the ALA IPR Policy.

2.7   Financial and Human Resources

At the beginning of the period, the ALA expects to hold approximately $2,490,000, representing the residue from the first two NCRIS payments and interest on these payments offset by payments expected under outstanding contracts (Australian Museum: user needs analysis; ABRS: taxonomic services). Final details will appear in the 2007-2008 Progress Report.

At the end of this reporting period, the ALA expects to hold approximately $2,700,000, representing the first two payments from NCRIS offset by an initial relatively low-cost phase in the ALA's existence while recruitment and initial planning is under way.

The following table summarises expected income for the period:

Balance forward	$2,490,000
NCRIS third payment	$1,951,000
Estimated interest	$100,000
Total NCRIS contribution (including interest)	**$4,541,000**
CSIRO cash	$1,000,000
Australian Museum cash	$100,000
Museum Victoria cash	$100,000
Queensland Museum cash	$100,000
Tasmanian Museum and Art Gallery cash	$70,000
University of Adelaide cash	$30,000
Southern Cross University cash	$50,000
Total participant cash contribution	**$1,450,000**
NeAT DIAS-B project	$400,000
Total other cash contribution	**$400,000**
Total income all sources	**$6,391,000**

CSIRO will credit the ALA with interest for the NCRIS funds held during the period. Interest will be credited for 2008-2009 during end of year accounting and will be reported in the 2009-2010 progress report.

The following table summarises **in-kind co-investment** by ALA participants:

CSIRO	$500,000
Australian Museum	$400,000
Museum Victoria	$850,000
Queensland Museum	$78,000
Tasmanian Museum and Art Gallery	$80,000
University of Adelaide	$18,000
Southern Cross University	$145,000
Australian Virtual Herbarium	$1,000,000
DEWR/ABRS	$321,000
Total participant co-investment	**$3,392,000**

See *Attachment 1: Revised ALA Budget 2006-2011* for details of planned expenditure within the period. This budget has been developed to replace that included in the Funding Agreement now that the project requirements and goals are better understood.

The ALA is pursuing opportunities for co-development and other in-kind collaboration with different partners which may lead to additional effective investment in the project.

During the period the following staffing positions will be funded using NCRIS funds:

Director	CSIRO Entomology
Executive Officer	CSIRO Entomology
Technical Architect	CSIRO Entomology
Metadata Curator	CSIRO Entomology
Java Developer	CSIRO Entomology
Mouse Phenomics Bioinformatician	ANU
Plant Phenomics Bioinformatician	University of Adelaide

Additional personnel will be engaged through contracts as follows:

Metadata Repository developer 1 (NeAT funds)	ANU
Metadata Repository developer 2 (NeAT funds)	ANU
Annotation Services developer 1 (NeAT funds)	UQ
Annotation Services developer 2 (NCRIS funds)	UQ
DIAS-B project facilitator (NCRIS funds – short contract)	ANU
Taxonomic Services developer 1 (NCRIS funds – part year)	ABRS
Taxonomic Services developer 2 (NCRIS funds – part year)	ABRS
Data Provision Services developer(s) (NCRIS funds, contracts)	Unknown
GIS Services developer(s) (NCRIS funds, contracts)	Unknown
Online Identification Services developer (NCRIS funds)	CBIT (IdentifyLife)

2.8   Milestones

Specific milestones are provided in Attachment 2.

2.9   Attachments

There is no confidential information attached, although several supplementary attachments are provided in appendices.

Donald Hobern
Project Director
26 June 2008

3. Attachments

3.1 Attachment 1: Revised ALA Budget 2006-2011

The following table is a revised budget for the ALA. The following should be noted:

1. Figures for 2006-2007 are actual figures reported by ALA participants for their own ALA-related activity.
2. The cash and in-kind income differs from the original budget in the following ways:
    a. The loss of Victoria DPI as an ALA participant reduced the projected cash allocation by $375,000.
    b. CSIRO originally projected $500,000 in cash for 2006-2007 and a further $500,000 in-kind activity through the CERF taxonomy hub project. The CERF project did not start until 2007-2008 and these funds were not available. The CSIRO cash allocation was also not made since the ALA started so late in the year. The ALA is exploring the possibility of spreading this cash sum over the remaining years of the Atlas.
    c. During 2007-2008 the NeAT programme has committed up to $1,000,000 for activities related to the ALA and these have been included here as in-kind funds.
    d. During 2006-2007 the cash and in-kind contributions of many of the participants deviated from the projections, although the net contributions significantly exceeded the projection.
3. In general small activities within the original project plan and budget have been associated with staff roles (Technical Architect, Java Developer, Metadata Curator) or broad development areas (Online Identification Tools, Data Provision Services, GIS Services).
4. See notes at the end of the table for information on individual line items.

	Income		Expenditure					
	Cash and In-kind	NCRIS	2006-2007	2007-2008	2008-2009	2009-2010	2010-2011	Total
TOTAL INVEST (original budget)	*26,553,255*	8,233,000						34,786,255
NCRIS by year			1,472,000	1,819,000	1,951,000	1,548,000	1,443,000	8,233,000
Cash and in kind			*6,125,243*	*5,632,081*	*5,139,194*	*5,146,539*	*4,510,198*	*26,553,255*
EXPENDITURE (revised budget)								
Output 1 Building the Atlas	*491,957*	5,183,000						
1.1 Recruitment and relocation costs								
1.2 Project Leader salary								
1.3 Project Leader Overheads (CSIRO in kind)								
1.4 Operating / travel etc 5 yrs								
1.5 User needs analysis								
1.6 Technical Architect								
1.7 Java Developer								
1.8 External review of usability and function								
1.9 Hardware (staff and development)								
1.10 Taxonomic Names Service								
1.11 Data hosting								
1.12 Data Provision Services								
1.13 Informatics support for 5.2.1, 5.2.2								
1.14 Contingency								
Total Spend			0	268,823	1,878,169	1,760,219	1,821,430	5,728,641
Cash and in kind			*0*	*69,320*	*134,064*	*140,767*	*147,806*	*491,957*

NCRIS Spend			0	199,503	1,744,105	1,619,452	1,673,624	5,236,684
Output 2 - Populating the Atlas	*27,663,685*	*0*						
2.1 Australian Virtual Herbarium cash								
2.2 Australian Virtual Herbarium in-kind								
2.3 CSIRO cash								
2.4 CSIRO in kind								
2.5 Australian Museum cash								
2.6 Australian Museum in kind								
2.7 Museum Victoria cash								
2.8 Museum Victoria in kind								
2.9 Queensland Museum cash								
2.10 Queensland Museum in kind								
2.11 Tasmanian Museum & Art Gallery cash								
2.12 Tasmanian Museum & Art Gallery in kind								
2.13 University of Adelaide cash								
2.14 University of Adelaide in kind								
2.15 DAFF (APPD) cash								
2.16 Southern Cross University cash								
2.17 Southern Cross University in kind								
2.18 ABRS in kind								
2.19 CSIRO External								
Total Spend			8,440,185	5,342,000	4,842,000	4,842,000	4,197,500	27,663,685
Cash and in kind			*8,440,185*	*5,342,000*	*4,842,000*	*4,842,000*	*4,197,500*	*27,663,685*
NCRIS Spend			0	0	0	0	0	0
Output 3 - Tools for data discovery	*1,000,000*	*1,250,000*						
3.1 Metadata Curator								
3.2 Tools survey								
3.3 DIAS-B project (NeAT in-kind)								
3.4 DIAS-B Annotation Services developer								
3.5 DIAS-B project facilitation								
3.6 Online Identification Services								
3.7 GIS Services								
Total Spend			0	35,000	967,840	956,733	646,070	2,605,643
Cash and in kind			*0*	*0*	*400,000*	*400,000*	*200,000*	*1,000,000*
NCRIS Spend			0	35,000	567,840	556,733	446,070	1,605,643
Output 4 - International Engagement	*0*	*850,000*						
4.1 GBIF Membership								
4.2 TDWG Membership								
Total Spend			165,200	165,600	165,600	165,700	165,800	827,900
Cash and in kind			*0*	*0*	*0*	*0*	*0*	*0*
NCRIS Spend			165,200	165,600	165,600	165,700	165,800	827,900
Output 5 - Governance & Management	*368,948*	*950,000*						
5.1 Operating / travel etc 5 yrs								
5.2 Branding								
5.3 Project Officer salary								
5.4 Project Officer Overheads (CSIRO) in kind								
5.5 CSIRO Line Management (10%) cash								
Total Spend			0	141,336	256,207	261,614	272,635	931,792

Cash and in kind			*0*	*55,236*	*99,147*	*104,451*	*110,114*	*368,948*
NCRIS Spend			0	86,100	157,060	157,163	162,521	562,844
ALL OUTPUTS							
Projected Income	*29,524,590*	*8,233,000*					
Total Spend			8,605,385	5,952,759	8,109,816	7,986,267	7,103,435	37,757,661
Cash and in kind			*8,440,185*	*5,466,556*	*5,475,211*	*5,487,219*	*4,655,419*	*29,524,590*
NCRIS Spend			**165,200**	**486,203**	**2,634,605**	**2,499,048**	**2,448,015**	**8,233,071**

Notes

1.1 Recruitment and relocation costs – includes advertising and international relocation for Director and Technical Architect

1.2 Project Leader salary – overheads carried by CSIRO as in kind (1.3 and 5.5)

1.5 User needs analysis – contract with Australian Museum

1.6 Technical Architect – partial year in 2008-2009

1.8 External review of usability and function – in line with Key Performance Indicators, budget originally allocated to usability testing has been allocated to two external reviews of the ALA services

1.9 Hardware (staff and development) – basic computing requirements for ALA staff and for development work

1.10 Taxonomic Names Service – contract in 2008-2009 with ABRS, may be extended subsequently to ensure integration of other species-based information held as part of AFD and APC

1.11 Data hosting – provisional budget for basic hosting of ALA servers

1.12 Data Provision Services – contracts to support bundling of software packages for data sharing and development of tools for hosting data

1.13 Informatics support for 5.2.1, 5.2.2 – primarily to be used to recruit bioinformaticians with APF and APPF

1.14 Contingency – to address additional opportunities as they arise

3.2 Tools survey – contract with Australian Museum

3.3 DIAS-B project (NeAT in-kind) – project funded by NeAT to support development of Metadata Repository and Annotation Services. Projection is to recruit two developers working on Metadata Repository at ANU and one on Annotation Services at UQ

3.4 DIAS-B Annotation Services developer – support a second Annotation Services developer at UQ

3.5 DIAS-B project facilitation – funds allocated to support initial project management and facilitation for the DIAS-B project

3.6 Online Identification Services – develop joint collaboration with Encyclopedia of Life and IdentifyLife to produce software and services for online identification tools (keys, etc.)

3.7 GIS Services – support developer working with BioMaps and others to produce GIS services.

4.1 GBIF Membership – annual contribution

4.2 TDWG Membership – original budget included funds for International Working Committee – this task is now being managed by using the TDWG online databases of projects, networks and events – funds partially reallocated to pay institutional membership of TDWG

5.2 Branding – domain name registration, branding work by Paper Monkey and funds reserved for developing materials based on branding

5.3 Project Officer salary – overheads carried by CSIRO as in-kind (5.4 and 5.5)

3.2 Attachment 2: Activities and Milestones for 2008/2009

Activities and Milestones for 2008/09	Achievement Date	Status
Project Implementation		
Output 1. Creation of ALA		
Technical Architect in position	6 October 2008	
Java Developer in position	21 July 2008	
Metadata Curator in position	21 July 2008	
Bioinformaticians in position	6 October 2008	
User needs analysis complete	31 October 2008	
AFD/APC web service active	31 December 2008	
Contract placed for Data Provision Services development work	31 August 2008	
Web site publicity around 200th anniversary of Darwin's birth	12 February 2009	
Output 2. Populating the Atlas of Living Australia		
Review of in-kind activity from 2007-2008	31 July 2008	
Prioritization criteria established (based on user needs analysis)	31 December 2008	
Key performance indicator content metrics for 2008-2009 achieved	30 June 2009	
Output 3. Tools for using biodiversity data		
DIAS-B project plan approved	15 July 2008	
DIAS-B project kick-off meeting	31 July 2008	
DIAS-B developers in position	31 July 2008	
ALA accepting registration of metadata for resources (interim Metadata Repository)	31 December 2008	
Contract placed for Online Identification Services project (with EOL and IdentifyLife)	31 August 2008	
Contract placed for GIS Services development work	30 September 2008	
Output 4. International Engagement		
ALA presented at TDWG conference	22 October 2008	
MOC established with EOL	31 July 2008	
MOC established with ClimateWatch	31 August 2008	
ALA web site acting as Australian GBIF node	28 February 2009	
Project Management		
Output 5. Governance and Management		
Branding selection process complete	30 September 2008	
Scoping Group review complete	30 September 2008	
Memoranda of cooperation established with CHAFC, CHAEC, AMRRN	30 September 2008	
ALA web site reworked to reflect branding	31 December 2008	
External review contracted for ALA portal and services	30 June 2009	

3.3   Attachment 3. Risk Management Strategy

This section updates the ALA Risk Management Strategy focussing particularly on the risks of particular relevance in 2008-2009. Changes and additions from the 2007-2008 Risk Management Strategy are in **bold**.

3.3.1 Specific risks

NCRIS Investment Plan 5.2.3 Risk/hazard Identification and Management Strategy		
Specific risks of particular relevance to the NCRIS Investment Plan in the 2008-09 period		
Area	Specific risk/hazard	Management Strategy
Informatics technology	Lack of the Informatics capacity and vision to create a world class version of the ALA.	Development of the Atlas of Living Australia is not a trivial exercise, and no-one has done it in an acceptable way yet, and certainly not on an inclusive, comprehensive national scale. It will require a stringent iterative process around testing and standards, collaboration on a national and international scale, and some clear direction and vision. We will manage this through the creation of a steering committee with the appropriate talent and backgrounds, and through proactive contact with colleagues in biodiversity informatics.
Recruitment	Inability to recruit staff with key skills, particularly in the biological informatics domain.	In some cases we know of suitable candidates in Australia – either to hire or to serve as collaborators. In other cases we will probably need to recruit from a broader pool of candidates. We will use well defined job descriptions and selection criteria to recruit appropriate personnel. **The core ALA staff positions have been filled. The risk and management strategy both continue to apply particularly to the two bioinformatician positions**
Managing client/stakeholder relationships	Clients/stakeholders not having ownership of outcomes leading to lack of uptake of outputs	Uptake of outputs by stakeholders is critical to achieve project outcomes. Explicit relationship building with key stakeholders will be established to achieve agreed outputs. Stakeholders will be engaged throughout the life of the project to ensure from the start that end user needs are considered in all stages of project planning. **The ALA Newsletter is intended to serve as a key information tool to build stakeholder involvement. The ALA user needs assessment (and the associated ALA day to be held at the TDWG meeting) are also intended**

		to strengthen communication between the ALA and its stakeholders. Good communication and realistic setting of expectations will be essential since there are so many possible areas in which the ALA could contribute.
Project selection	Projects for creating tools or populating the ALA will need to be chosen to ensure meaningful outputs are achieved.	Governance and steering committees will be formed to make decisions on the projects to be included in the business plan. Project selection criteria will include feasibility and impact. **The ALA user needs analysis is intended as a key management strategy to give the ALA Management Committee the information needed to select priorities, and to provide criteria for measuring success in meeting user requirements.**
Technology Failure	Failure of technology to provide products essential to the delivery of outputs in the appropriate electronic environment.	Monitoring of existing technology, and working in conjunction with technology suppliers to ensure that suitable platforms are chosen for project delivery. Working with technology suppliers and informatics specialists to guarantee that where possible specific requirements can be built into existing application and products without needing large investment in programming costs.

3.3.2 General Project Management Risks

General Project Management Risks		
Area	**Specific risk/hazard**	**Management Strategy**
Resource management	Ineffective management resulting in slippage, non delivery, poor resource allocation	Effective project management, with realistic budgeting, milestones and workforce planning
Quality standards	Outputs not meeting agreed standards will jeopardize uptake	Project monitoring to ensure compliance with agreed global standards and stakeholder recommendations
Performance management; project team communication	Disenfranchised unhappy staff, resulting in poor performance	Clear and regular feedback. Performance agreements with clear , achievable objectives and appropriate training opportunities for staff at all levels
Managing team members in other institutions/ locations	Poor performance due to lack of communication, and proper controls to ensure delivery	Established framework for communication between all project members, with certain line management duties relegated to appropriate staff in other institutions.
Dependence on key staff	Loss of expertise essential to project delivery	Succession planning for life of project; project selection to ensure projects chosen

		where this is unlikely.
Skills & training	Lack of appropriate competencies can affect project delivery	Funding for necessary training and development built into project
Financial planning, control and reporting	Poor financial management can lead to cost overruns and insufficient budget for key tasks	Good project management, with regular review of budgets and expenditure
Data loss/backup/ storage	Data loss jeopardises delivery	Implement proper IT policy and procedures
Obsolete technology	Poor equipment and software will affect outcomes and quality	Sector knowledge and budget forecasting to ensure proper equipment and software
Technology transfer (publishing knowledge, website)	Failure to deliver outputs in appropriate format can lead to failed outcomes	Proper planning and monitoring to ensure outputs are to appropriate standards, which have been agreed between all parties.
Effective communication	Poor marketing of achievements would be detrimental to uptake of project results	Appropriate communication strategy, developed through engagement with stakeholders. This can include press releases, product launches, presentations at national and international conferences.

3.4 Attachment 4. User needs analysis work plan

The ALA user needs analysis will be a **targeted investigation of biodiversity information workflow**, including the following steps:

1. **Initial kick-off meeting (AM and ALA)** (0.5 days)
   Select initial contact groups and individuals to approach initially during subsequent steps.

2. **Quick capture of summary information** (10 days)
   Telephone and email contact with a wide range of people at the various institutions and development of a first pass summary of known user groups and uses of biodiversity data in Australia.

3. **Review of summary information** (0.5 days)
   Meeting between ALA and AM team to select approximately 12 scenarios for further investigation and refinement.

4. **Visits to capital cities or other key locations** (15 days)
   These visits (5-6 in all) will have two purposes:
   a. Interview data providers and data consumers involved in the selected scenarios to understand the full flow of information involved in reaching a conclusion (including digital and non-digital information resources).
   b. Hold 2-3 meetings to gather input and suggestions from a wider spectrum of ecologists, government officers, field workers, etc.

5. **Documentation of scenarios** (10 days)
   Produce a (wiki or web database) website with summary information for all scenarios identified 1.b. (and any additional scenarios identified in 1.d.ii.) and detailed information for selected scenarios explored in 1.d.i. The site should be planned to make it simple to add additional scenarios or to expand summary scenarios into detailed scenarios. The site should accommodate the following information:
   a. Brief summary description of scenario.
   b. User groups involved (contact details where permission is granted).
   c. Information used (databases, books, other resources).
   d. Specific questions or applications addressed.
   e. (Detailed scenarios only) Scenario presented in detail as a series of steps.

6. **Workshop exploring 2-3 scenarios** (1 day)
   A half-day workshop is planned at the TDWG meeting in Perth on 22 October 2008 to explore some of the ALA use cases. 2-3 scenarios should be presented at this meeting to organise discussion between content providers, IT experts and users on the best ways to address these needs.

7. **Summary report** (2 days)
   In addition to the website, a summary report will be produced to document the key findings from the investigation, including:
   a. Review of process, including any recommendations for how the process could be improved.
   b. Overview of findings from development of scenarios.
   c. Identification of key bottlenecks in information workflow which might be addressed by the ALA.

d. Any recommendations for ALA regarding the form and style of tools and interfaces to be developed, to improve their suitability for target user groups.
e. Recommendations for ways to publicise results.

3.5   Attachment 5. AFD/APC Service work plan

APC/APNI and AFD to provide open access to these taxonomic and nomenclatural data using generic interfaces.

The resulting project would continue to employ the existing java programmers and would begin in July 2008. The work required would deliver the following outcomes for the AFD, APC and ALA.

1. APNI and AFD enhancements to support TCS.
2. Implementation of LSID's within both systems.
3. Implementation of a TCS data store.
4. Development of a java TAPIR provider with TCS capabilities.
5. Unified web service interfaces allowing users to submit a request and retrieve TaxonName and/or TaxonConcept data from AFD and APC/APNI.
6. TaxonName and TaxonConcept records to reference Publication records wherever applicable.
7. Alternative interfaces based on TAPIR, SOAP and REST.
8. LSID resolution.
9. Support for response documents using TDWG TCS or TDWG TaxonName/TaxonConcept/Publication RDF vocabularies.
10. All TaxonName, TaxonConcept and Publication records to have LSIDs.
11. Interfaces to include:
    a. Find TaxonConcept(s) by scientific name
        Parameters including at least:
        - scientific name OR partial scientific name ("starting with")
        - (Optional) higher taxon to limit results (specified by LSID or name)
        - (Optional) flag to use only AFD or APC/APNI
        - (Optional) scientific name author and/or year
    b. Find TaxonName(s) by scientific name
        Parameters including at least:
        - scientific name OR partial scientific name ("starting with")
        - (Optional) flag to use only AFD or APC/APNI
        - (Optional) scientific name author and/or year
    c. Get TaxonConcept by LSID
    d. Get TaxonName by LSID
    e. Find child TaxonConcepts for TaxonConcept
        Parameters including at least:
        - LSID for parent TaxonConcept
        - (Optional) partial scientific name ("starame Services 08 09_letterhead.docting with")
    f. Get parent TaxonConcept for TaxonConcept (not strictly necessary since TaxonConcept documents presumably contain this)
        Parameters including at least:
        - LSID for child TaxonConcept
    g. Get Publication by LSID
12. Develop a foundation for future additional web service interfaces (TAPIR, SOAP, REST) to return SPM documents for any taxon identified by its TaxonConcept LSID.

3.6   Attachment 6. Draft DIAS-B work plans

3.6.1 DIAS-B work plan for ALA Metadata Repository

The following is a summary of the functional requirements for the ALA Metadata Repository.

1. Storage of metadata documents, including:
   a. Support for Dublin Core
   b. Support for ISO 11179
   c. Support for arbitrary RDF properties
   d. Support for tagging metadata with ontology (OWL/OBO) terms
2. Metadata documents to describe:
   a. Online databases and web services
   b. Text documents (including PDF, Word, etc.)
   c. Images and other multimedia resources
3. Harvesting of metadata from other repositories [1]
   a. OAI-PMH
4. Pluggable framework for metadata tagging [2]
   a. Support for TAPIR and TaxonOccurrence data
   b. Support for TAPIR and TaxonConcept data
   c. Support for scientific name detection in text documents (see uBio services)
   d. Support for image metadata standards (including XMP)
5. Publication of metadata to other tools and repositories
   a. OAI-PMH (offering ontology terms as OAI-PMH Sets?)
   b. Other SOAP and REST services as required (?)
6. Data provider interface
   a. Register/update/delete data provider
   b. Register/update/delete database/service/document/image from data provider
   c. Register/update/delete OAI-PMH feed from data provider
   d. Accept/reject annotations
7. Administrator interface
   a. Annotate metadata documents with arbitrary RDF properties and ontology terms
   b. Accept/reject annotations
8. End-user interface [3]

---

[1] It is expected that some resources will be registered directly into the metadata repository, but that there should also be an option to register OAI-PMH feeds from which streams of metadata documents can then be retrieved. The import of harvested metadata documents may be filtered as part of requirement 4). In other words, it could be that the harvesting process only imports documents for which at least one tag has been identified.

[2] As far as possible data providers will be encouraged to provide full metadata when registering data resources. However it is clear that many providers regard this as a burden and that most metadata projects have great difficulty in ensuring that resources are sufficiently well documented. It is therefore proposed that the ALA Metadata Repository should seek to automate generation of a range of basic metadata fields through direct inspection of the underlying data resources. Wherever possible such fields should be represented as RDF properties referencing standard ontologies. This approach will help to ensure a standard baseline for searching the repository. For online databases exposed through a protocol such as TAPIR, this approach may involve generating tags e.g. for every species represented in the database. For a text document it could mean using tools such as those developed by uBio to recognise scientific names and to generate tags from these names. Image metadata can be mined in the same way. A pluggable framework would allow new ontologies and new object types to be handled. Ultimately processors could be developed to classify text documents by looking for clusters of terms from different subdomains (e.g. medicine or conservation).

[3] One of the possible uses for the interface proposed in 8.e (*Propose annotations to metadata documents with arbitrary RDF properties and ontology terms*) could be to serve as a store for user annotations proposed through the Annotation Service also being developed as part of this NeAT project, although further investigation is required to determine the most appropriate place to store such information.

    a. Full-text search
    b. Browse by data provider
    c. Browse by ontology terms
    d. Faceted search via multiple ontologies
    e. Propose annotations to metadata documents with arbitrary RDF properties and ontology terms

9. Access control
    a. Integration with Shibboleth/PKI (to exploit AAF infrastructure and services)
    b. AAF authenticated access to data provider, administrator and end-user interfaces
    c. AAF-mediated restrictions on visibility for some metadata documents (possible – may not be necessary)

10. Wider compatibility (may be ensured by other requirements) [4]
    a. Compatibility with requirements within other NCRIS capabilities and within Platforms for Collaboration
    b. Compatibility with other developing international biodiversity metadata repositories from projects such as GBIF, EOL, OBIS, LTER

The following resources will contribute to the development of the ALA metadata repository:

1. ALA core staff - the following staff members are funded for the whole of 2008-2011 from NCRIS funds as part of the main ALA project. All of these will contribute expertise and development activity to the project. Other developers contracted under NCRIS funds will also contribute to tasks allocated here to the ALA Architect and the ALA Java Developer.

- ALA Director (Donald Hobern)
  - Project vision
  - Project management
  - Coordination with other ALA activities and international projects
- ALA Metadata Curator (to be appointed, June 2008)
  - Adoption and promotion of metadata standards and practices
  - Adoption of biodiversity ontologies (including taxonomic hierarchies)
  - Metadata quality control
  - Integration of data resources
- ALA Architect (to be appointed June 2008)
  - Data architecture
  - Integration between metadata repository and other ALA components
- ALA Java Developer (to be appointed June 2008)
  - Develop interim metadata repository (temporary database while this project develops ALA metadata repository)
  - Representation of taxonomic hierarchies as ontologies
- ALA-Mouse-Bioinformatician
  - Adoption and promotion of metadata standards and practices in NCRIS 5.2.1
  - Adoption of mouse ontologies
  - Metadata linkages to NCRIS 5.2.1 resources
- ALA-Plant-Bioinformatician
  - Adoption and promotion of metadata standards and practices in NCRIS 5.2.2

---

[4] The ALA is already in discussion with GBIF, EOL, OBIS and various LTER projects about management of biodiversity metadata. These organisations are planning a joint workshop to evaluate opportunities for shared development and for exchange of metadata. This workshop with take place at the TDWG annual conference in Perth in October 2008. This workshop will be an opportunity to identify collaborators and to understand wider interoperability needs.

- o Adoption of plant ontologies
- Metadata linkages to NCRIS 5.2.2 resources

2. CSIRO ICT Centre staff - the following staff members will be based at the CSIRO ICT Centre in Canberra and will be funded in part from NeAT funds and in part by the ICT Centre for the duration of the project.

- Supervisor (Kerry Taylor, ICT Centre in-kind)
  - o Technical vision and oversight
  - o Guidance and management of ICT Centre developers
- Repository Developer 1 (NeAT funded, ICT Centre covering overheads)
  - o Repository Developer 2 (NeAT funded, ICT Centre covering overheads)

3. CSIRO IM&T's eSIM project shares many interests and goals with the ALA Metadata Repository project and the IM&T staff have been offered to provide assistance in the following areas:

- Java software developer with GIS experience
- Authentication specialist (1 x AD, 1 x AAF), but both on part-time basis
- Repository Management (equiv 1 x FTE, but probably a few people on a part-time basis)
- Over the longer term - customer service and production support staff once repositories and AAF become part of the service suite offered by IMT (would be after this pilot and purely derived as a result from this pilot) - including Service Centre staff and Database specialists.
- Librarian/Records/curation expertise for work with meta-data and standards development (possibly up to 3 individuals, but not full time - dependent on library operational requirements) (Probably 1 FTE)
- Business Analyst/process specialist to understand lifecycle model/data management processes at ALA and develop system specifications – (1 FTE)

The following is a provisional work plan for the project:

1. Requirements refinement workshop (July 2008)

   One day workshop to review and refine basic requirements documented above (including identification of any external parties to be consulted).

2. Evaluation of metadata repository software (July – August 2008)

   4-week evaluation led by IM&T and ICT Centre staff, to review candidate metadata repository solutions and technologies (in particular Fedora as adopted by ANDS) and to recommend components for development of core metadata repository solution.

3. Selection of core metadata registry technologies (August 2008)

   Team workshop to review recommendation from software evaluation and develop detailed implementation/configuration plan, including refining timeline for task 4.

4. Implementation of core metadata registry (August – December 2008)

   Main development activity - exact details dependent on selected technologies and which functional requirements will be satisfied by team development. The immediate goal is to develop an

operational metadata store which can subsequently be enhanced with additional metadata tagging functions. This work will be carried out by the IM&T and ICT Centre developers. The dates are estimates and may be modified based on selected technologies.

5. Prioritisation of ontologies for tagging metadata documents (August – October 2008)

    ALA Metadata Curator and Bioinformaticians to identify priority ontologies for tagging metadata within the metadata repository (including taxonomies and possibly gazetteers).

6. International biodiversity metadata workshop (October 2008)

    Workshop to be held at the TDWG annual conference in Perth. Goal is to bring together a wide range of international projects managing biodiversity and ecological data sets to compare requirements and to identify immediate opportunities to collaborate in software development and/or metadata exchange.

7. Representation of ALA taxonomy as ontology (October 2008 – March 2009)

    Development of interfaces to represent ALA integrated taxonomy as an OWL or OBO ontology for use within the metadata repository.

8. Implementation of harvester components (January – March 2009, approx.)
    a) OAI-PMH – Metadata import

    Extension (if required) of core metadata repository to support registration of an OAI-PMH endpoint and to harvest metadata documents from external repositories. This work will be carried out by IM&T and ICT Centre developers.

9. Metadata tagging modules (April – June 2009, approx.)
    a) TAPIR – TaxonOccurrence
    b) CSV – TaxonOccurrence
    c) Text document – scientific names
    d) Images – scientific names

    Development of pluggable modules for evaluating content of data resources identified by metadata documents and for assigning corresponding ontology-based metadata tags. These will be developed by the ICT Centre developers with support and domain expertise from the ALA developers.

10. Metadata search interfaces (July-December 2009, approx.)
    a) Browse by ontology
    b) Ontology-based faceted search

    Development of user interfaces for browsing and filtering metadata documents using any of the supported ontologies as an organisational hierarchy or using all supported ontologies to present a faceted search interface. These will be developed by the ICT Centre and ALA developers.

11. Interfaces to retrieve data (January – December 2010, approx.)

    Building on metadata tagging in task 9, apply ontology-based metadata descriptions to structured data resources (typically relational or XML databases) in order to construct and execute queries to

retrieve attribute data from within a given data resource. Structured queries (as opposed to Boolean keyword or attribute-value-pair queries) will be phrased over metadata terms (such as those developed in task 10) and will be rewritten as queries for source repositories to retrieve the data elements specified in the query. The rewriting will be achieved through the use of expressive mappings that relate database content to metadata terms.

3.6.2 DIAS-B work plan for ALA Data Annotation Services

In the following text, all annotations are considered to be generated, stored and retrieved as *annotation documents*. In this context, an *annotation document* is considered to be an XML document conforming to a specified XML or RDF schema.

At its simplest such a document could be considered to comprise four elements:

- The (globally unique) identifier for the data object to which the annotation relates.
- An identifier or other link to metadata identifying the author of the annotation.
- The date and time at which the annotation was created.
- A body holding the annotation content as free-form text.

Other annotation documents may be more highly structured. For example, assume that a data object has the GUID urn:lsid:csiro.au:anic:1234 and is associated with a set of data elements taken from the TDWG taxon occurrence vocabulary (http://rs.tdwg.org/ontology/voc/TaxonOccurrence), including the following elements:

```
<rdf:RDF>
  <rdf:Description rdf:about="urn:lsid:csiro.au:anic:1234">
    <dc:identifier> urn:lsid:csiro.au:anic:1234</dc:identifier>
    . . .
    <to:locality>Reid, Canberra, ACT, Australia</to:locality>
    <to:decimalLatitude>149.138</to:decimalLatitude>
    <to:decimalLongitude>-35.280</to:decimalLongitude>
    . . .
  </rdf:Description>
</rdf:RDF>
```

An annotation document could represent a proposed correction to such a record in a structured form, in this case by reversing the values for decimalLatitude and decimalLongitude. The schema for such an annotation document might extend the general document structure described above to include within the body one or more properties from the taxon occurrence vocabulary, these being considered to be proposed overrides for any values in the original record.

Similarly schemas could be defined for annotation documents to address the following use cases:

- Tag a data object with a term from an ontology or set of more informal tags
- Associate an image (identified by URL) with a taxon concept (identified by an LSID referencing the Australian Plant Census or Australian Faunal Directory)
- Associate two taxon concepts (identified by LSIDs) via a specified relationship (predator-prey, host-parasite, plant-pollinator, etc.)
- Associate a taxon concept (identified by an LSID) with descriptive terms (e.g. "leaf shape: obovate") from an ontology of such terms

In all of these cases the schema defines an opportunity for significant information exchange between the annotator and final users of the data. The adoption of such schemas does not necessarily constrain the space of possible annotations but provides a framework for identifying and supporting these more

significant use cases.

1. Metadata schema repository for annotation document structures
   a. XML Schema for standard annotation document structures
   b. RDF Schema for standard annotation document structures
2. Metadata schema definitions for specific use cases
   a. General free text annotation for any data object
   b. General basic tag annotation for any data object
   c. Area of interest annotation for any image object
   d. Simple occurrence record (observation) annotation for any taxon concept
   e. Proposed data correction annotation for any occurrence record (TDWG vocabulary)
   f. Proposed data correction annotation for any data object based on any TDWG vocabulary
   g. Image link annotation for any taxon concept
   h. Inter-taxon relationship annotation for any taxon concept
   i. Descriptive term annotation for any taxon concept
   j. Response annotation for any annotation
3. Storage of annotations
   a. Central service for storing annotations
   b. Web service for storing annotations (single or bulk)
   c. Validation of annotations by schema
   d. Globally unique identifier (GUID) assigned to each annotation
   e. Integration with AAF identity services
   f. Annotations indexed by annotator, annotated data object, class of annotated data object, date of annotation
   g. Administrative interface to suppress all annotations from a given annotator (by annotator identifier)
4. Retrieval of annotations
   a. Web service to retrieve annotation by annotation GUID
   b. Web service to retrieve annotations for data object by data object GUID (includes retrieval of annotations as a thread)
   c. Web service to retrieve annotations by annotator identifier
   d. OAI-PMH service to harvest annotations (need to consider appropriate use of OAI-PMH Sets)
5. User interfaces
   a. "Widget" for entry of free-text comment for any data object (N.B. since an annotation is itself a data object, this may be all that is needed to manage response annotations)
   b. "Widget" for selecting ontology term for tagging any data object (N.B. with modification this could serve as the basis for a widget to annotate taxon concepts with descriptive terms or inter-taxon relationships)
   c. "Widget" for entering an occurrence record
   d. Form-based user interface for proposing corrected values for a data object based on a TDWG vocabulary
   e. "Widget" for uploading an image to an ALA image store and for storing an annotation relating the image to a taxon concept
   f. "Widget" for annotating an image area of interest

The following resources will contribute to the development of the ALA annotation services.

1. ALA core staff - the following staff members are funded for the whole of 2008-2011 from NCRIS funds as part of the main ALA project. All of these will contribute expertise and development activity to the project. Other developers contracted under NCRIS funds will also contribute to tasks allocated here to the ALA Architect and the ALA Java Developer.

- ALA Director (Donald Hobern)
  - Project vision
  - Project management
  - Coordination with other ALA activities and international projects
- ALA Metadata Curator (to be appointed, June 2008)
  - Integration of annotation services with ALA Metadata Repository
- ALA Architect (to be appointed, June 2008)
  - Data architecture
  - Integration between annotation services and other ALA components
- ALA Java Developer (to be appointed June 2008)
  - User interfaces for annotation from ALA web pages

2. CSIRO ICT Centre staff - a team will be based at the CSIRO ICT Centre in Canberra to work on the ALA Metadata Repository. This team will be funded in part from NeAT funds and in part by the ICT Centre for the duration of the project. This team is expected to interact closely with the Annotation Services development team.

3. University of Queensland staff – the following staff members will be based at the University of Queensland:

- Supervisor (Jane Hunter, 20% in-kind)
  - Technical vision and oversight
  - Guidance and management of UQ developers
- 2 Annotation Services Developers (1 NEAT funded, 1 ALA funded)
  - Development of annotation tools and services

4. Other collaborators - several other projects, including ClimateWatch (an initiative of EarthWatch) and the Murray-Darling Basin Commission have expressed interest in developing interfaces to allow observations and other user-provided data items to be stored and managed. The ALA expects to work with such projects to develop user interfaces which could connect to the Annotation Services.

The following is a provisional work plan for the project:

1. Requirements refinement workshop (July 2008)

   One day workshop to review and refine basic requirements documented above (including potential liaison with other groups). In conjunction with requirements refinement for Metadata Repository.

2. Development of implementation plan (August 2008)

   Team to develop detailed implementation/configuration plan, including refining timeline for tasks 3 and following.

3. Core implementation of metadata schema repository and annotation store (September – December 2008)

   Initial development activity to develop or configure tools for managing storage and indexing of a store for annotation documents. This work will be carried out by the UQ developer with support from the ALA core staff. The dates are indicative only and subsequent development will continue

in parallel with other activity.

4. Prioritised development of metadata schema definitions for key use cases (September – December 2008)

   ALA Metadata Curator and Architect to develop a series of metadata schema documents with support from UQ staff.

5. Implementation of retrieval services for annotations (January – June 2009)

   UQ developer to develop services to search, browse and retrieve annotations, with support from ALA core team. Dates are indicative.

6. Implementation of user interfaces (July 2009 – June 2010, approx.)

   UQ developer, ALA core staff and other collaborators to develop and deploy user interfaces for entry of annotations for prioritised metadata schema definitions.

3.7   Attachment 7. Acronyms

AAF	Australian Access Federation
ABCD	Access to Biological Collections Data
ABIF	Australian Biodiversity Information Facility
ABIN	Australian Biosecurity Information Network
ABRS	Australian Biological Resources Study
ACPFG	Australian Centre for Plant Functional Genomics
AFD	Australian Faunal Directory
ALA	Atlas of Living Australia
AM	Australian Museum
ANDS	Australian National Data Service
ANU	Australian National University
APC	Australian Plant Census
APF	Australian Phenomics Facility
APN	Australian Phenomics Network
APNI	Australian Plant Name Index
APPF	Australian Plant Phenomics Facility
AVH	Australia's Virtual Herbarium
BA	Birds Australia
CBIT	Centre for Biological Information Technology
CERF	Commonwealth Environmental Research Facilities
CHAFC	Council of Heads of Australian Faunal Collections
CHAH	Council of Heads of Australasian Herbaria
CITES	Convention on International Trade in Endangered Species of Wild Fauna and Flora
CSIRO	Commonwealth Scientific, Industrial and Research Organisation
CSIRO ICT	CSIRO Information and Communication Technologies
CSIRO IM&T	CSIRO Information Management & Technology
CSV	Comma Separated Value
DEWHA	Department of the Environment, Water, Heritage and the Arts
DIAS-B	Data Integration and Annotation Services for Biodiversity
EoL	Encyclopaedia of Life
eSIM	eScience Information Management
FTE	Full-time Equivalent
GBIF	Global Biodiversity Information Facility
GIS	Geographic Information System
GUID	Globally Unique Identifier
HTML	HyperText Markup Language
IBS	Integrated Biological Systems
IPR	Intellectual Property Rights
LTER	Long Term Ecological Research
MDBC	Murray-Darling Basin Commission
MoC	Memorandum of Cooperation
NCRIS	National Collaborative Research Infrastructure Strategy
NeAT	National eResearch Architecture Taskforce
NGOs	Non-government Organisations
OAI-PMH	Open Access Initiative Protocol for Metadata Harvesting
OBIS	Ocean Biogeographic Information System
OBO	Open Biomedical Ontologies
OGC	Open GIS Consortium

OWL	Web Ontology Language
OZCAM	Online Zoological Collections of Australian Museums
PDFs	Portable Document Format
PKI	Public Key Infrastructure
RDF	Resource Description Framework
REST	Representational state transfer
SOAP	Simple Object Access Protocol
TAPIR	TDWG Access Protocol for Information Retrieval
TCS	Taxon Concept Schema
TDWG	Taxonomic Databases Working Group
TERN	Terrestrial Ecosystem Research Network
TRIN	Taxonomic Research Information Network
uBio	Universal Biological Indexer and Organizer
UQ	University of Queensland
URL	Uniform Resource Locator
WCS	Web Coverage Service
WFS	Web Feature Service
XMP	Extensible Metadata Platform