

Data Integration and Annotation Services in Biodiversity

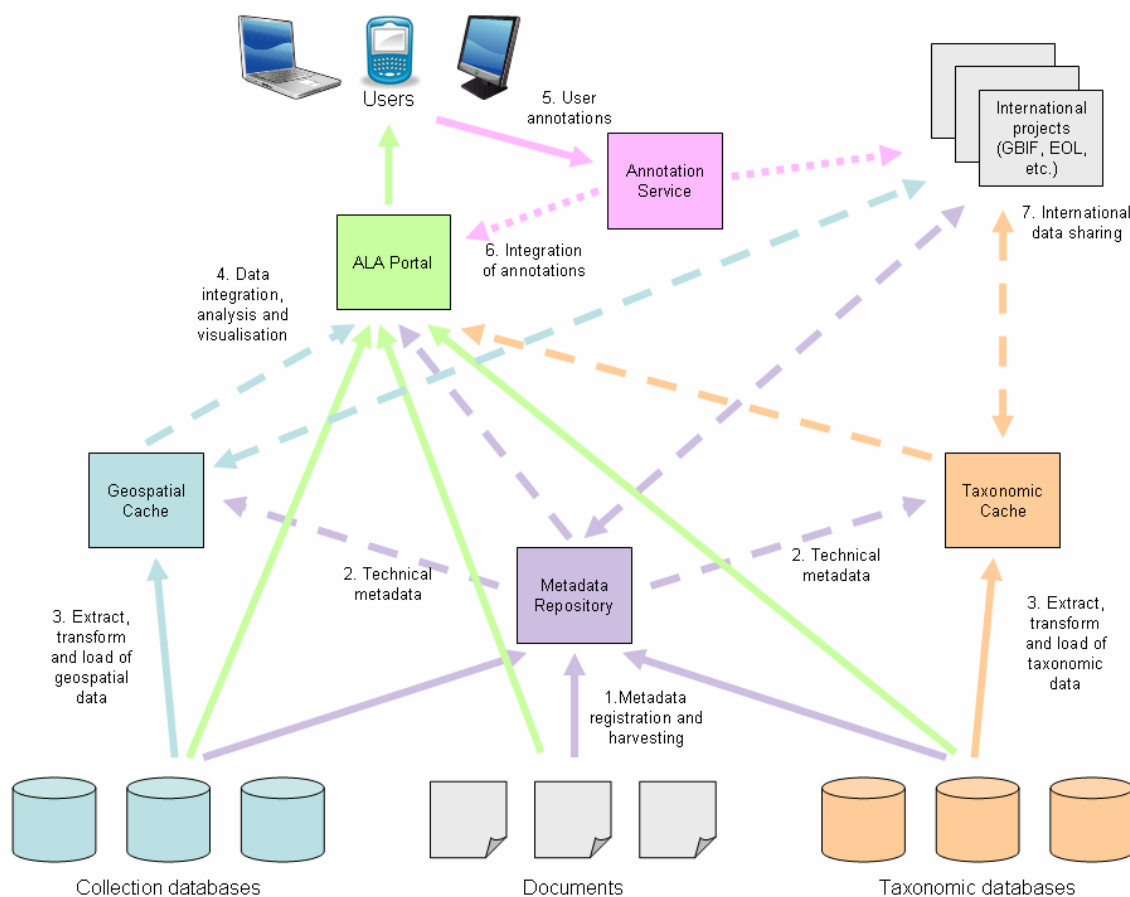
Overview *Summarise the context and objectives of this project.*

The DIAS-B project addresses both the data publishing and data use aspects of data integration across biodiversity data holdings and also considers a two-way flow of information between those functions.

It consists of two related sub-projects that are of immediate benefit to NCRIS 5.2.3, the Atlas of Living Australia (ALA) and more widely to NCRIS 5.2 Integrated Biological Systems (IBS).

However, the services developed for these sub-projects are expected to have much broader applicability.

The following diagram is a simplified illustration of some components of the ALA and indicates the role and significance of these sub-projects within the ALA.



The diagram illustrates the following flows:

1. **Metadata registration and harvesting** – The ALA needs to maintain a **Metadata Repository** for metadata (technical and descriptive) for all available digital resources of biodiversity information. Within the ALA this repository will support the development of a range of secondary information tools (see 2, 3 and 4 below) and provide metadata to a range of international projects. The first subproject described in this document relates to the development of such a repository and the processes and tools required to integrate it into the wider Australian IT environment.

2. **Technical metadata** – Various components within the ALA will query the **Metadata Repository** to discover relevant resources to incorporate in secondary information resources such as GIS caches or integrated taxonomic tools.
3. **Extract, transform and load** – ALA components will interact with resources to build caches of specific classes of data using an ETL (Extract, Transform and Load) process. This process will be triggered by the addition or modification of metadata in the **Metadata Repository** or on a regular schedule (in part determined by whether each resource is under active development or maintenance). Extraction will make use of protocols and standards including those developed by the Taxonomic Databases Working Group (TDWG) and the OAI Protocol for Metadata Harvesting. The initial implementation will be based on the open source software developed by GBIF, but later phases of this project will investigate the use of more general ontology-driven tools to map between different data models.
4. **Data integration analysis and visualisation** – The ALA will develop a range of user interfaces to guide different user groups to discover, analyse and visualise data from the registered resources. These interfaces will use the various ALA caches to optimise data discovery and to provide rapid overview of data, but will also link to the original resources as required (to retrieve additional data fields, etc.).
5. **User annotations** – Users should be able to annotate data records and resource metadata (and ideally arbitrary sets of search results) with comments (and ultimately structured corrections). The proposal is for an **Annotation Service** to be developed for this purpose, to maintain such annotations and the identifiers (URNs, URLs, etc.) for the original records.
6. **Integration of annotations** - The **Annotation Service** should provide interfaces for other services to relate annotations to the referenced data and to treat them as an integrated whole. The diagram shows only a small subset of the components which might use this service. For example, the original data providers could connect to the **Annotation Service** to retrieve and process annotations to their data.
7. **International data sharing** – The ALA will act as a bidirectional gateway between Australian and international biodiversity projects and will manage transfer (and where necessary transformation) of both metadata and data to and from these projects.

Most of the elements in this architecture are well understood and working well within a number of biodiversity informatics projects, including the Australian Virtual Herbarium (AVH), the Online Zoological Collections of Australian Museums (OZCAM), the Global Biodiversity Information Facility (GBIF) and the Ocean Biogeographic Information System (OBIS).

The **Metadata Repository** and the **Annotation Service** are both services which are required within the ALA but which will also need to be developed for other NCRIS capabilities and for use more widely within ANDS. For this reason it is proposed that a NeAT project is initiated to develop these services to address the needs of the ALA and simultaneously develop standard models, tools and services which can be scaled more widely. The timing of this project may necessitate the development and deployment of ALA-specific instances of these services, but the goal should be for these to be candidates for wider adoption and for long-term support through ANDS, or at least that they should be designed to fit into a federated framework of peer implementations.

Users *Identify the research communities and resource providers that this project serves; and the potential number of users. This should include some NCRIS capabilities or other data federating or collaborating research groups, and any institutions that will participate through setting requirements for or steering this project.*

The project draws on data providers and research data users covering:

- NCRIS 5.2.3 Atlas of Living Australia (ALA)

- NCRIS 5.2.1 Australian Phenomics Network (APN)
- NCRIS 5.2.2. Australian Plant Phenomics Facility (APPF)
- NCRIS 5.11 Terrestrial Ecosystem Research Network (TERN)
- NCRIS 5.8 Australian Biosecurity Information Network (ABIN)
- NCRIS 5.12 Integrated Marine Observing System (IMOS)
- Biodiversity data providers (museums, herbaria, state government departments, universities) and users (taxonomists, botanists, zoologists, environmental scientists)

The specific data publishers within the project are:

- CSIRO (including collection databases from the Australian National Insect Collection, Australian National Herbarium, Australian National Fish Collection and Australian National Wildlife Collection, as well as image databases, digital literature resources and dynamic identification keys)
- State museums and herbaria (including collection databases, as well as a range of species information resources and images)
- Southern Cross University (including collection and sequence data from the Australian Plant DNA Bank)
- Australia's Virtual Herbarium (an index of collection databases from Australian herbaria)
- Online Zoological Collections of Australian Museums (including collection databases from Australian zoological collections, as well as a range of species information resources and images)
- DEWHA (including taxonomic databases and a range of species information resources and images)
- DAFF (including the Australian Plant Pests Database)
- Australian Microbial Resources Information Network (including collection data and strain information from Australian culture collections)
- Global Biodiversity Information Facility (including taxonomic, collection and observational data relating to Australian biota and held in overseas institutions)
- Australian Phenomics Facility (including mouse phenomics databases from the Australian Phenomics Network)
- NHMRC Australian PhenomeBank (including plant phenomics databases)
- Australian National University (including plant phenomics databases)

The specific research data users within the project include all of the data publishers listed above as well as the wider community of taxonomists, botanists, zoologists, environmental scientists, land-use and conservation planners, and biosecurity officers. The Australian Centre for Plant Functional Genomics will be a specific user of plant phenomic data mediated through the project.

The range of current and potential users of biodiversity data within Australia is very broad. A paper is available (Chapman, [Uses of Primary Species-Occurrence Data](http://www.gbif.org/prog/digit/data_quality/UsesPrimaryData), http://www.gbif.org/prog/digit/data_quality/UsesPrimaryData) detailing a wide range of examples of such uses and users. The ALA will be focusing particularly on developing tool sets for biosecurity (including identification of pest species, modeling of environmental requirements and access to literature on control and biology), land-use management (including integration of geospatial data with climate, vegetation, soil and geology, and information on the conservation or pest status of different taxa) and taxonomy (integration of primary information of all classes for each species). However the ALA is currently in the process of contracting a more detailed user needs analysis which will be used to establish specific priorities for the next few years and to identify core research collaborators.

Needs *Describe the needs of the research communities or resource providers that this project seeks to address.*

Metadata Repository

The ALA needs to support integration of a wide range of different types of biodiversity data – taxonomic data (e.g. taxon names and synonyms), specimen and observation data, species descriptions and associated images, diagnostic keys, genomic data, etc.

The project aims to maximise the discoverability and interoperability of all of these data by managing a metadata repository (or repositories) and a range of data indexes or caches. These services should be optimised to address the needs of currently recognised user groups, but should also as far as possible support hitherto unrecognised approaches to searching and combining the data.

The ALA therefore needs to adopt best practices for metadata management, including adoption of relevant vocabularies and ontologies. The Taxonomic Databases Working Group (TDWG, <http://www.tdwg.org/>) acts as a standards development body for the collections community and biodiversity informatics in general. TDWG has been reworking a range of existing standards to serve as a framework ontology for managing biodiversity information. This includes a significant amount of domain expertise gathered from many years of modelling collection data, taxonomic and nomenclatural information, and species descriptions. It also incorporates a general-purpose high-level categorisation of species information suitable for organising biological information resources (Species Profile Model, SPM). TDWG standards are used by many international and Australian projects, including:

- Global Biodiversity Information Facility (GBIF, <http://www.gbif.org/>)
- Bioversity International (<http://www.bioversityinternational.org/>)
- Mammal Networked Information System (MaNIS, <http://manisnet.org/>)
- Ocean Biogeographic Information System (OBIS, <http://www.iobis.org/>)
- Encyclopedia of Life (EOL, www.eol.org)
- Australia's Virtual Herbarium (AVH, <http://www.chah.gov.au/avh/>)
- Online Zoological Collections of the Australian Museums (OZCAM, <http://www.ozcam.gov.au/>)
- IdentifyLife (<http://www.identifylife.org/>)

Adoption of these standards will therefore allow the ALA to share data resources more widely and to use information from a global pool. Mappings are also available or being developed between TDWG standards and related data formats (e.g. HISCOM for herbarium specimens, Delta for species descriptions).

The ALA will also make use of a number of Australian and international reference standards which will act as reference vocabularies for different parts of its domain. The Australian Plant Names Index (APNI) and the Australian Faunal Directory (AFD) are being brought together to provide a single reference service for all Australian organism names and to serve as a reference classification. International nomenclatural and taxonomic services (IPNI, Index Fungorum, ZooBank, Catalogue of Life, etc.) will augment these services where required. The ALA will seek out reference gazetteers, and protected and pest species lists to assist with management of Australian data. Metadata and identifiers for natural history collections will be integrated into the Biological Collections Index, a new international database being developed to bring together information on all the world's collections. Under TDWG's guidance most or all of these projects are currently assigning resolvable globally unique identifiers (LSIDs – Life Science Identifiers) with RDF metadata to these reference lists.

The two 5.2 phenomics capabilities similarly need to be able to relate their data to international standards such as the Plant Ontology (<http://www.plantontology.org/>), various mouse ontologies and the Mouse Genome Database. A key concern for these projects is to identify and adopt best practice standards for management of metadata to minimise disruption as these ontologies continue to develop.

The ALA therefore needs a world-class **Metadata Repository** to manage a wide range of metadata for all of these different data sets, including at least:

- Basic Dublin Core metadata describing each resource
- Technical metadata (endpoints, schemas and protocols for accessing each resource)
- Information on the origins and methodology for each resource (particularly for collection and observation databases, for which this information provides the essential context for interpreting bias)
- Terms from taxonomies, gazetteers, controlled vocabularies and ontologies which can serve to characterise the data and allow user applications to recognise the most relevant resources for a given purpose
- Access metadata (IPR, etc.)
- Logos and credit statements for inclusion in user interface presentations of the data

This service should support access and search interfaces to meet the needs of Australian researchers through ANDS and also to meet the needs of international data sharing activities such as GBIF and EOL.

It is expected that metadata may enter the repository through a number of routes, including:

- Primary registration of a resource by the providing institution (normally including ingestion of an existing metadata format in a known format)
- Secondary discovery of resources through a harvesting protocol such as OAI-PMH
- Manual or automated annotation and enhancement of registered metadata by:
 - A Metadata Curator (ALA is recruiting someone in this role)
 - Analytic tools – for example other ALA components may be able to annotate metadata with summary information derived from ETL processes)

The core of the **Metadata Repository** will depend upon the ability to map between different metadata models (both when populating the repository and when exporting metadata documents). The project will therefore investigate the use of ontology-driven tools to manage such mappings. In later phases, the use of these tools may also be extended to handle transformation of data records between alternative formats.

These requirements are, in the main, highly generic and shared with other data integration projects. For this reason it is clearly preferable for the **Metadata Repository** to be developed as a service or federation of services operating under ANDS.

Annotation Service

An additional requirement is for tools that enable authenticated community annotation of data. This should include the capability to flag errors or possible errors in the data, and mechanisms for this information to be passed on to the data custodians, and for the people who flagged the error to be notified about the outcome of their query (similar to a helpdesk ticketing system). This system should also support services for automated checking of data quality and consistency of data across multiple data providers.

The detailed implementation of this service would seek to address the following user scenarios:

- A user wishes to annotate a data record with some comments that can also be viewed by other users.
- A user considers that a particular data record (or some more systemic feature of an entire data resource) is somehow incorrect (e.g. the identification of the specimen to a particular taxon is dubious, or the coordinates for a eucalyptus specimen place it in the ocean). The user wishes to flag the problem and for this information to be fed back to the data provider for further investigation. This is problematic because data from the same provider may be presented to the user through many different portals, e.g. ALA, GBIF, EOL, state museum herbarium information systems, etc. A further enhancement would be to allow the user to edit a structured copy of the record to propose some required changes.
- Data validation software can automatically identify some problems in a data set (e.g. inconsistencies between coordinates and locality name) and could provide feedback to data providers as in the previous scenario. This implies a need for a standard web services API for providing annotations and feedback.
- A data provider will require some tools to view feedback and annotations. Such tools could notify the supplier of the comments that they are being processed or provide a response (that the record has been corrected or was in fact valid). In cases in which the original record was valid, it will be important to maintain the comments and responses to inform future users. The tools could also assist providers with importing structured corrections into the source database.
- Regardless of whether or not a data provider processes the feedback and annotations, there should be services allowing other web applications to discover them and include them for display to other user (including where applicable any responses from the data provider).

A complicating factor in all of these cases will be the sociological issue of how to handle what may be perceived as apparent criticism of data providers and their resources. Such considerations will need to inform the way that such information is presented and could necessitate additional workflow to allow comments to be reviewed before they are made fully public.

The ALA requires a service which will support all of these requirements. It is clearly preferable for such a service to be developed to become a general component in the ANDS infrastructure, for use by researchers in all areas.

Services *Describe the result of the project in terms of the service(s) that will be implemented and demonstrated by the project and which could be operated in an ongoing fashion; and the proposed operator of each service.*

The following includes description of several services which will be developed by the ALA outside this NeAT project but which are listed here because they provide the operational context for the proposed **Metadata Repository** and **Annotation Service**.

- **Catalogue** of mandated and supported data standards, vocabularies, ontologies for use within the ALA – the exact form of these services should be defined in collaboration with other NCRIS initiatives rather than uniquely within the ALA. In the short term ALA will host this catalogue as part of its own web infrastructure, most likely through CSIRO. As the project proceeds, this may be hosted by the ALA or by TDWG or through ANDS.
- **Metadata Repository** and metadata registration software for registration of all Australian biological data resources and for relating data sets to supported vocabularies and ontologies. In the short term it is expected that this will be hosted by the ALA, most likely through CSIRO. A goal in this project is however to understand the optimal model for managing metadata repositories for Australian research. It may be that ANDS will ultimately host an integrated metadata repository on behalf of all user groups, or that the ALA will continue to host its own

metadata repository as a peer within a network of metadata repositories. The **Metadata Repository** will include:

- User interfaces and web services for registering and modifying metadata for any resource
 - Search interfaces and web services to search the metadata repository using terms from supported ontologies (as identified in the Catalogue above).
- **Packaged data provider software**, based on TAPIR, but also incorporating other relevant standards such as OAI-PMH (<http://www.openarchives.org/OAI/openarchivesprotocol.html>), GML WFS (<http://www.opengeospatial.org/standards/wfs>) and EML (<http://knb.ecoinformatics.org/software/eml/>), that will support querying and accessing a variety of biodiversity data types using different standard or custom schemas. This software will be made available through the ALA web site, but may also be maintained as an open source community project (e.g. on SourceForge) in conjunction with GBIF, EOL, Bioversity International and others. The software will include interfaces for registering resources in the **Metadata Repository**.
 - **Annotation Service** allowing human and machine users to store information relating to any data record within the ALA (via LSIDs or other globally unique identifiers) and to accommodate data provider responses to this information. It is expected that some investigation will occur into supporting more structured annotation of data records, but the extent to which this will be possible within the time frame is not yet clear. In the short term this service will be hosted by the ALA, but it is expected that ANDS will ultimately deploy an implementation of it on behalf of all user groups. The **Annotation Service** will include:
 - Web service interface allowing web applications to retrieve and display any annotations and responses registered for any data record, hosted by the ALA.
 - Biodiversity search services which incorporate awareness of access frameworks and each user's rights to view any particular data record, hosted by the ALA, but reliant upon AAF frameworks.

All of these services are components which are core to the long-term operation of the ALA and can be maintained by the ALA as part of its future operation, unless it becomes clear that full responsibility for any service should be transferred to ANDS.

The ALA however expects that much of the software underlying these services will also be used by other international projects affiliated with TDWG and will aim to spread future development and maintenance costs by retaining these components as open source activities. Existing open source projects which could serve as homes for this software include the various TAPIR implementation projects and the GBIF data portal software codebase.

NeAT Characteristics

eResearch effect *What changes in behaviour and activity are expected from the project that will demonstrate the broader adoption of eResearch practice?*

The ALA is directed towards developing a digital information commons for biodiversity data in Australia. In the short term this work will focus on improving the accessibility and integration of existing digital data resources and ensuring that these resources are presented in forms which simplify their use in research and planning activities. As the ALA proceeds, it will develop tools and interfaces to ensure that new research data are automatically published within the data network. A key focus for the ALA will be to develop practices and tools which give incentives both to data providers and to data users to use the common infrastructure. The elements of this project are all fundamental components required by the ALA to make this possible.

Broader adoption *Which additional communities, resource providers or organisations would also be expected to benefit from the provision of the same or similar services should the project succeed?*

The services developed by DIAS-B and the associated software will be initially targeted at ALA and the other IBS capabilities, but could also be extended to support TERN and possibly IMOS. It is also likely that the Biosecurity capability would make use of the services that are provided.

The concepts and some of the software, particularly the data annotation mechanisms and integration with federated access mechanisms, should also be more widely applicable. The AusLit NeAT project will require a similar approach.

In particular, the use of a collections registry and metadata schema repository to assist with data discovery and integration is an approach that has broad applicability and will be adopted by ANDS, so this project provides a very useful demonstrator and exemplar for the rollout of this approach in the wider context of ANDS.

Value adding *Identify the components of the project that could be based around generic technologies or be implemented through shared services for which the project would provide an exemplar use case or requirement set.*

The Data Integration component of the project will reuse software developed by TDWG to implement the TAPIR protocol and to support LSIDs.

The Metadata Repository component will work with ANDS to adopt best practice models and standards and to serve as an exemplar and a test repository alongside other NCRIS-related metadata repositories for exploration of inter-capability metadata integration.

The project will also investigate using software for data annotation services and for collections and metadata registries that have been developed by the ARCHER and APSR projects, which are likely to be taken up by ANDS.

The shared services based upon federated access management will build on AAF frameworks and are expected to provide reference patterns for services in other NCRIS capabilities.

The annotation services could also become a shared component hosted by ARCS or serve as an exemplar for other capabilities.

Standardisation *Describe the global technology development or standardisation work that will be adopted, adapted or extended within the project and any risk reduction available by collaboration with similar activities occurring elsewhere in the world.*

The data integration component of this project will be based on international standards developed by the Taxonomic Data Working Group (TDWG) and software development efforts by TDWG, GBIF, EOL and ALA. Australia has been a leader in this area and a key contributor to TDWG over many years. Over the last two and a half years an Australian has served as Project Manager in modernising TDWG, and the Director of ALA, Donald Hobern, who formerly led software development for GBIF, is also currently the Chair of TDWG.

Project Scoping

Key Participants *Name any Pfc components, any NCRIS capabilities, or any other institutions or groups that will need to be involved in the project planning and execution.*

This project is mainly targeted at NCRIS 5.2 Atlas of Living Australia, and ALA will contribute significant effort to the project.

Within Pfc, it would mainly require input from ANDS, primarily the data integration aspects, with some overlap with ARCS, particularly in the provision of the different web services required for this project. AAF authentication is another component of the project, and this project requires authentication of data providers (who are often employees of government agencies) as well as researchers.

eRSA has already had significant engagement with the ALA community and Paul Coddington from eRSA is on the ALA Scoping Committee (the technical steering committee). As part of that interaction, eRSA has developed a new version of Australia's Virtual Herbarium that aggregates specimen data from TAPIR data providers, and provides a web portal for searching and accessing the aggregated data. As part of this work, eRSA has developed Tapirus, a Java library of simple tools for querying and parsing results from TAPIR providers.

Jane Hunter's group at UQ has expertise in ontologies and data integration, including applying them to ecological applications, and has developed a general collaborative data annotation system, and it is likely that these could be adapted and applied here.

A reference group will provide input to the project planning and execution. This will include members of the ALA Scoping Committee, who are key people with experience in biodiversity informatics who are providers and users of biodiversity data:

- Donald Hobern, Director of Atlas of Living Australia
- Kevin Thiele, Western Australia Herbarium
- Paul Flemons, Australian Museum
- John LaSalle, Australian National Insect Collection, CSIRO
- Steve Shattuck, Australian National Insect Collection, CSIRO
- Greg Whitbread, Australian National Herbarium
- Jim Croft, Australian National Herbarium
- Tony Rosling, Australian National Herbarium
- Ely Wallace, Museum Victoria
- John Morrissey, CSIRO
- Paul Coddington, eRSA

as well as the following additional people:

- Jane Hunter, UQ
- Hugh Possingham (Federation Fellow), UQ
- Lead software architect of the ALA, to be appointed
- Metadata Curator of the ALA, to be appointed
- Mouse Phenomics Bioinformatician, to be appointed

- Plant Phenomics Bioinformatician, to be appointed

The project plan would be developed by Donald Hobern with input from the reference group and guidance from the steering committee.

Governance

Outline the arrangements proposed to manage the contributions and user interaction. Eg: a project managed by ARCS, ANDS, or under another NCRIS capability governance, or by subcontract to a named lead agency, or a new J/V between the parties.

The project could be managed by a small steering committee comprising the following people or their representative:

- Donald Hobern, Director of Atlas of Living Australia
- Representative from each of the IBS mouse and plant phenomics projects
- Executive Director of ARCS
- Executive Director of ANDS

The project would be led by Donald Hobern, Director of ALA.

The contract for the project could be managed through ARCS.

ALA has a policy that all software they produce will be made open source and freely available.

Project Scale

Identify the overall scale expected in the project, eg. 1 to 3 years, total effort in any year, and nominate any parties that have indicated a willingness to participate through providing resources. (funded or in-kind, people or facilities).

The two subprojects will run in conjunction with other ALA development activity and are expected to run for the remaining 3-year duration of the current ALA funding period. The ALA, in conjunction with the APN and the APPF, expect to maintain a development team for this period including at least the following positions:

- Technical Architect
- Mouse Bioinformatician
- Plant Bioinformatician
- Metadata Curator
- Web Administrator
- 4 Programmers

The project expects also to engage additional development effort through participation in TDWG and GBIF open source development activities and through collaboration with EOL. ABRS and the Australian Museum also have existing developers whose products will be contributing directly to the development of the Atlas.

This team will work on all of the subprojects identified in this proposal.

NeAT funding is sought for 3-4 additional programmers for three years to complement this team and to focus specifically on the subproject areas (approximately two FTEs per subproject, but subject to actual requirements as the subprojects develop).

Major Steps

Identify the key steps that will be visible to users as the services develop. Note that some observable deliverable is needed every half year and projects may be reviewed based on the achievement of these steps.

These proposed steps are very preliminary. More detailed deliverables and milestones will be incorporated into the project plan. As ALA develops further, these will be refined by the ALA Director and Architect in consultation with the ALA Scoping Committee, PfC and key members of TDWG, EOL and GBIF.

Metadata Repository

2008H2

- Review of metadata management and requirements in related international biodiversity informatics projects (particularly GBIF and EOL)
- Review metadata standards and ontologies in use within relevant Australian and international projects, and mappings between them.
- Review available software options for a metadata repository.

2009

- Contribute to the development of the TDWG core ontology.
- Establish standards for the use of unique identifiers for data resources and data items.
- Develop user interfaces and web services for primary registration of data resources and for configuration of OAI-PMH harvesting.
- Develop user interfaces and web services for search and selection of data resources via ontology terms as well as free-text search.
- Develop alternative output metadata formats (based on review of metadata standards above).
- Investigate how to integrate outputs from the Annotation Service into metadata management.

Annotation Service

2008H2

- Investigate requirements for annotation services in other NCRIS capabilities and in ANDS.
- Investigate existing collaborative annotation systems and select the most appropriate solution.
- Investigate how to integrate it with the Metadata Repository and other components in the ALA system.

2009

- Develop an appropriate user interface that may need to be customised for structured annotation of different types of data.
- Test automated annotation of records by error-checking tools.
- Develop interfaces for management of obsolete annotations (e.g. after data record has been corrected for errors) and for threaded annotations (e.g. data provider responses to user comments)
- Test and refine the interface with a variety of users.
- Provide support for AAF authentication.

Dependencies

Identify dependencies that exist to activities or developments external to the project.

The key dependencies which are outside the control of the ALA relate to the adoption of appropriate standards and services as recommended models within NCRIS projects. In particular this project depends on the ability to follow best practice approaches in establishing its metadata repository and in interfacing with the AAF.

The project will also exploit developments in a wide range of other projects, particularly international biodiversity informatics activities. However the ALA does not expect failures in any of these projects to impact its ability successfully to deliver its core functions within the funding period.