

Project Plan for *Data Integration and Annotation Services in Biodiversity*

Project Charter

- *Overview of the research community need (why the project has been started)*
- *Description of the proposed service solution and how it meets the need identified above*

This project serves a dual purpose:

1. To provide core services required by the NCRIS 5.2.3 *Atlas of Living Australia* (ALA) capability to support management and discovery of biodiversity data resources.
2. To use the ALA requirements to develop tools and best practice models for data management which can be reused in other NCRIS capabilities and to facilitate use of data between NCRIS capabilities.

NCRIS 5.2.3 ALA Needs:

The ALA has been conceived as a project to provide data integration services for a wide range of data resources relating to Australian plants, animals and microorganisms (in particular specimen and observation databases, names and classification, digital literature, DNA and protein sequences, descriptive data and online keys, images and other multimedia).

The most fundamental requirement within the ALA is to improve the management of metadata describing existing resources. It should be possible to manage metadata with variable levels of granularity. For example a *resource* could correspond to a single image or document (with detailed descriptive and technical metadata relating to the single object) or to an entire database of images, documents or observations. In the latter case the metadata may remain at the level of a general description of the entire database and how it can be accessed, or the database-level metadata may provide global information which is then augmented with record-level metadata harvested e.g. using OAI-PMH. The granularity of the metadata captured and managed in different contexts will be determined by the capabilities of the data providers, by the readiness of the data providers to share detailed metadata, by the needs of end users and software applications, and by the findings from this project on the most appropriate approaches for different situations. In general the metadata management system should be able to support documentation for individual objects or (potentially nested) collections of objects and to integrate metadata from multiple levels of granularity to meet user expectations.

The ALA is seeking to describe each resource not only with descriptive metadata fields (e.g. Dublin Core) but also by reference to a range of ontologies and standard vocabularies to improve the effectiveness of search tools for end users and to facilitate mining of the associated data by intelligent software agents. The ALA does not exist in a vacuum and needs to ensure that its metadata management is consistent and compatible with approaches taken by other biodiversity informatics projects around the world and by other research infrastructure within Australia. In particular it needs to ensure that resources combining biodiversity data with elements relevant to other domains (e.g. ecological data sets including geochemical or atmospheric measurements) should be discoverable by users within those domains.

It is also recognised that it is difficult to promote the production of high-quality metadata for each data resource. Data providers rarely have the resources required to spend significant time authoring metadata – and the expertise may be lacking to produce a full and accurate description for legacy data sets. The ALA therefore wishes to explore the use of community annotation services to improve description and quality control for biodiversity data resources.

NCRIS Needs:

The requirements identified for the ALA are representative of data management needs shared by many other NCRIS capabilities. Each capability will need tools and models to ensure that increasingly large amounts of data are well catalogued and that users are able to discover relevant resources using

Data Integration and Annotation Services in Biodiversity (ALA: DIAS-B) National eResearch Architecture TaskForce Project

appropriate classifications. Significant benefits can be expected both in development costs and in long-term sustainability from ensuring that such tools and models can be reused across capabilities. Even greater benefits can be expected from ensuring that users and software agents can use standard discovery services to locate and integrate resources from multiple domains.

NCRIS also requires that data integration and annotation services exploit the developing services of ANDS, ARCS and AAF. The DIAS-B project will provide standard implementations which conform in particular with the AAF authentication frameworks.

Proposed services solution:

This project will provide services in two key areas:

1. Data Integration Services, including:
 - a. Operational metadata repository for biodiversity data resources, including registration, discovery and annotation services.
 - b. Reusable software implementation for use by other NCRIS capabilities, etc.
2. Annotation Services, including:
 - a. Operational annotation repository for annotations relating to biodiversity data (but potentially open for use by users in any domain), including services to create, recover and harvest annotations
 - b. Reusable software implementation for use by other NCRIS capabilities, etc.

Project Approach

How will the required services be delivered? This will influence the resources and staff required.

Development

The services will be developed through collaboration between the ALA, CSIRO IM&T, the CSIRO ICT Centre in Canberra and the University of Queensland. This collaboration will be coordinated by the ALA.

The project will be executed as two separate but interrelated streams: Data Integration Services and Annotation Services.

Data Integration Services

The Data Integration Services relates to the development of the ALA metadata repository. The following resources will contribute to this development activity:

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 (Lynette Woodburn)
 - Adoption and promotion of metadata standards and practices
 - Adoption of biodiversity ontologies (including taxonomic hierarchies)
 - Metadata quality control
 - Integration of data resources
 - ALA Architect (Dave Martin)
 - Data architecture
 - Integration between metadata repository and other ALA components

Data Integration and Annotation Services in Biodiversity (ALA: DIAS-B) National eResearch Architecture TaskForce Project

- ALA Java Developer (Nick dos Remedios)
 - 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
 - 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 (Carsten Friedrich, ICT Centre in-kind)
 - Technical vision and oversight
 - Guidance and management of ICT Centre developers
 - Repository Developer 1 (NeAT funded, ICT Centre covering overheads)
 - Repository Developer 2 (50% NeAT funded, 50% ALA 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 Active Directory, 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)
4. University of Queensland staff – the Annotation Services team (based at the University of Queensland) is expected to interact closely with the Data Integration Services development team. Exact relationships will be defined during the initial project task refining the project plan.

Annotation Services

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

Data Integration and Annotation Services in Biodiversity (ALA: DIAS-B) National eResearch Architecture TaskForce Project

- ALA Metadata Curator (Lynette Woodburn)
 - Integration of annotation services with ALA Metadata Repository
 - ALA Architect (Dave Martin)
 - Data architecture
 - Integration between annotation services and other ALA components
 - ALA Java Developer (Nick dos Remedios)
 - User interfaces for annotation from ALA web pages
2. 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 (NeAT funded)
 - Development of annotation tools and services
3. CSIRO ICT Centre staff – the Data Integration Services team (based at the CSIRO ICT Centre in Canberra) is expected to interact closely with the Annotation Services development team. Exact relationships will be defined during the initial project task refining the project plan.
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.

Delivery

There are two facets to the delivery of the outputs.

- Software components will be made available via SourceForge or other appropriate open source code repositories. An information page will be developed for the project giving full details of how to access and use source code and compiled versions of software.
- The Data Integration Services and Annotation Services will be deployed as part of the ALA web services accessible through the ala.org.au domain. All services within this domain will be hosted by CSIRO IM&T in Melbourne and subsequently replicated to other locations in Australia as required to ensure adequate performance in all states. It should be noted that data volumes for both the Data Integration Services and the Annotation Services are expected to be moderate and should not present significant problems for replication.

NeAT funds will primarily be used to support the development of the software components at the CSIRO ICT Centre, Canberra (Data Integration Services) and the University of Queensland (Data Annotation Services). As part of its in-kind contribution the ALA will manage the transfer and deployment of these components as live web services, and will evaluate their performance and feed back requirements for enhancing and improving the software components.

The project will make use of a part-time project facilitator (funded by ALA) to organise and run a requirements refinement workshop to be held at the start of the project.

Project Scope

What exactly is included within the responsibility of the project?

Which work is included within the project and which work is outside the control of the project (but may still be required to meet parts of the objectives)?

Are there any interfaces with other projects?

Work within the responsibility of the project:

Data Integration and Annotation Services in Biodiversity (ALA: DIAS-B) National eResearch Architecture TaskForce Project

- This is described in detail in the Key Deliverables section.

Work outside the control of the project:

- Other ALA development activity, in particular:
 - Developing agreements on data standards
 - Packaging and deployment of data provision software, particularly in smaller institutions
 - Development of user interfaces to guide users to data resources

Interfaces to other projects:

- GBIF, EOL, OBIS and other international biodiversity information projects wishing to share access to Australian data resources – these relationships will be managed as part of the ALA's general activity and all projects will be kept informed of the goals and progress within this project and encouraged to align activities as appropriate

Governance and project management arrangements

Who is accountable for assessing project performance, what process will they apply?

Describe the process by which project changes will be agreed.

Describe the authority structure over resources in the project.

Proposed arrangement:

- The ARCS/ANDS agreed governance mechanism for NeAT projects will apply (See Appendix A)
- The project will be managed by a Project Committee, consisting of nominated members from the key participants and users. It will meet quarterly via phone and or agreed electronic medium.
 - ARCS Executive Director, Professor Anthony Williams, or designate
 - ANDS Executive Director, or designate
 - ALA Director, Mr Donald Hobern (Chair)
 - University of Queensland Lead, Professor Jane Hunter
 - User representative, Professor Hugh Possingham, University of Queensland
 - Project Manager, Dr Lynette Woodburn, *ex officio*
- The ALA Metadata Curator will serve as Project Manager and will report to the Project Committee.
- The project will have a Reference Group of interested service providers, developers and users, which will act in an advisory capacity, as well as having an evaluation role, providing feedback on the project progress and outcomes. Initial members of the Reference Group should represent the following groups:
 - CSIRO ICT Centre, Dr Carsten Friedrich
 - CSIRO IM&T eSIM, Ms Tracey Hind
 - APSR, Chris Blackall
 - ALA Scoping Group
 - NCRIS 5.2.1 Australian Phenomics Network (APN – probably the bioinformatician being recruited as technical liaison between the ALA and APN)

Data Integration and Annotation Services in Biodiversity (ALA: DIAS-B) National eResearch Architecture TaskForce Project

- NCRIS 5.2.2 Australian Plant Phenomics Network (APPN – probably the bioinformatician being recruited as technical liaison between the ALA and APPN)
- NCRIS 5.8 Australian Biosecurity Information Network (ABIN)
- NCRIS 5.11 Terrestrial Ecological Research Network (TERN)
- NCRIS 5.12 eMarine Information Infrastructure (eMII)
- Global Biodiversity Information Facility (GBIF – Éamonn Ó Tuama, Programme Officer for Inventory, Discovery and Access)
- US Long-term Ecological Research Network (LTER-Net – Matt Jones, National Center for Ecological Analysis and Synthesis)
- Ocean Biogeographic Information System (OBIS, Tony Rees, CSIRO Marine and Atmospheric Research)

Accountability:

- Funding for the Data Integration Services subproject will go to CSIRO as the lead agency in the ALA, and CSIRO will manage employment of the two Repository Developers.
- Funding for the Data Annotation Services subproject will go to QCIF, who will sub-contract the University of Queensland eResearch group to execute this subproject and manage employment of the Annotation Services developers.
- ANDS and ARCS will pay their NeAT Project funds quarterly in arrears based on acceptable performance on a per EFT basis for each NeAT Project. The ANDS and ARCS quarterly reviews will be the trigger for either approving or withholding NeAT funds for that quarter from a NeAT Project or a component of that project as appropriate.
- The ALA will provide in-kind support for integration and deployment of the software as the basis for live web services hosted within the ALA domain.

Project Changes:

- The Project Committee will be responsible for changes in the project milestones and project outcomes.

Governance of resources in the project:

- Resources allocated to this project are managed by the Project Committee through the Project Manager.
- The Project Manager will make use of the wider reference group to ensure the project is consistent with international standards and community expectations.

Customer engagement

Describe the means by which customer satisfaction with the project's progress, quality and outputs will be measured.

- The ALA will promote the project through its own newsletter and through its own discussions with user groups and will solicit feedback on the suitability and usability of services as developed.
- The ALA will be partnering with other biodiversity informatics organisations in developing tools based around these services. In particular the Data Integration Services will feed into ALA's provision of data to projects such as GBIF and EOL, and the Data Annotation Services

Data Integration and Annotation Services in Biodiversity (ALA: DIAS-B) National eResearch Architecture TaskForce Project

are expected to be significant in collaborations with ClimateWatch and other user-oriented activities.

- The existing ALA key performance indicators document (see <http://www.ala.org.au/documents.htm>) includes a number of criteria of relevance to this project (especially under 1.1.2, 1.1.3 and 1.1.4). The success of the ALA in meeting its own performance indicators will in large part depend on the success of this project. These indicators can therefore also serve as a basis for evaluating progress on this project.
- The ALA is also organising independent reviews (to be contracted at the end of 2008-2009 and at the end of 2010-2011) to document the experience of key target user groups and to compare the state of ALA infrastructure with other national biodiversity information platforms. The services to be developed under this project will be evaluated at the same time.

Performance Measurement

Which management systems and processes will be in place to ensure the scope, schedule and cost constraints are adhered to.

Management Systems

- The Project Manager is responsible for the ensuring the scope, schedule and costs constraints are adhered to, with the Project Committee ensuring accountability in these areas of the project.
- A number of performance indicators have been defined for the ALA, including targets for the number of records and data items to be integrated for different taxonomic groups and the range of different user groups supported. The success of the ALA in meeting these goals will depend on the successful development of the ALA Metadata Repository. These indicators can therefore also serve as measures for the DIAS-B project. They can be found at: http://www.ala.org.au/docs/ALABusinessPlan2007-2008_Attachment4_KeyPerformanceIndicators.pdf
- The ALA has scheduled external reviews of its services in 2009 and 2011 – the outputs from the DIAS-B project will be assessed as part of these reviews.
- The project team will investigate and document the benefits arising from the combination of Data Integration and Data Annotation Services by comparing the search results with and without use of annotation data (i.e. comparing plain full-text searches with annotation-driven searches) for real user search terms.
- The project team will also measure the effectiveness of the Data Annotation Services in improving data quality, by assessing:
 - The percentage of annotations which lead to changes in the original data records
 - The percentage of user requests for which annotation data refine the results (either by user-requested filtering of data records with identified issues, or by user-requested substitution of data elements with proposed corrected values)
 - The perceived value of these services to users (by questioning users as part of the ALA external evaluations in 2009 and 2011).

Work Breakdown Structure

Data Integration Services

The following is a provisional work plan for the project (to be refined during task 1):

1. Requirements refinement workshop (September 2008)

Data Integration and Annotation Services in Biodiversity (ALA: DIAS-B) National eResearch Architecture TaskForce Project

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

2. Evaluation of metadata repository software (September – October 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 (October 2008)
Team workshop to review recommendation from software evaluation and develop detailed implementation/configuration plan and QA (test) plan, including refining timeline for task 4.
4. 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.
5. Implementation of core metadata registry (October 2008 – March 2009)
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.
6. Prioritisation of ontologies for tagging metadata documents (November 2008 – February 2009)
ALA Metadata Curator and Bioinformaticians to identify priority ontologies for tagging metadata within the metadata repository (including taxonomies and possibly gazetteers).
7. Representation of ALA taxonomy as ontology (October 2008 – April 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 (April – September 2009, approx.)
 - a) OAI-PMH – Metadata importExtension (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 (September 2009 – June 2010, approx.)
 - a) TAPIR – TaxonOccurrence
 - b) CSV – TaxonOccurrence
 - c) Text document – scientific names
 - d) Images – scientific namesDevelopment 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.

Data Integration and Annotation Services in Biodiversity (ALA: DIAS-B) National eResearch Architecture TaskForce Project

10. Metadata search interfaces (September – June 2010, 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 (July 2010 – June 2011, 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.

12. Testing, tuning and enhancement of services (July 2010 – June 2011, approx.)

Testing and evaluation of services as they are deployed and to tune and enhance these services to improve their effectiveness and uptake.

Data Annotation Services

The following is a provisional work plan for the project (to be refined during task 1):

13. Requirements refinement workshop (September 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. Include representative from ANDS.

14. Development of implementation plan (September – October 2008)

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

15. Prioritisation of use cases for capture of annotations (October – December 2008)

ALA Metadata Curator and Architect to document a set of priority annotation use cases (based in part on the outcomes from the ALA user needs analysis, to be completed in October 2008)..

16. Core implementation of metadata schema repository and annotation store (October 2008 – March 2009)

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 developers with support from the ALA core staff. The dates are indicative only and subsequent development will continue in parallel with other activity.

17. Prioritised development of metadata schema definitions for key use cases (January – March 2009)

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

Data Integration and Annotation Services in Biodiversity (ALA: DIAS-B) National eResearch Architecture TaskForce Project

18. Implementation of retrieval services for annotations (March – September 2009)

UQ developers to develop services to search, browse and retrieve annotations and to notify data providers of new annotations, with support from ALA core team. Dates are indicative.

19. Implementation of user interfaces (October 2009 – December 2010, approx.)

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

20. Testing, tuning and enhancement of services (July 2010 – June 2011, approx.)

UQ developers, ALA core staff and other collaborators to test and evaluate services as they are deployed and to tune and enhance these services to improve their effectiveness and uptake.

Key Deliverables

- *What are the key project deliverables?*
- *Who is responsible and accountable for each key deliverable?*

Project Deliverables at completion of this project
--

The project is divided into two related subprojects:

- The Data Integration Services component focuses on the development of a Metadata Repository with a range of services to enhance description of data resources and to support access by different groups of users.
- The Data Annotation Services component focuses on the development of tools to generate and store annotations on any data resource or individual data record.

The subprojects are related in the following ways:

- The Data Annotation Services will provide a pool of secondary metadata (from multiple sources) for use by the Data Integration Services.
- The Data Integration Services will use the Data Annotation Services to store annotations on data resources arising from data validation, data mining and user comments.
- The combined services will be used to explore approaches to enhance the discoverability and usefulness of data resources through annotation (e.g. with ontology terms)

The following sections outline the software deliverables from the project. The ALA undertakes to deploy these software components and to manage them as the basis for live web services (see *Delivery* under *Project Approach* above).

Data Integration Services

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.)

Data Integration and Annotation Services in Biodiversity (ALA: DIAS-B) National eResearch Architecture TaskForce Project

- c. Images and other multimedia resources
3. Harvesting of metadata from other repositories ¹
 - a. OAI-PMH
4. Pluggable framework for metadata tagging ² (uses the Data Annotation Services component)
 - 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 (uses the Data Annotation Services component)
7. Administrator interface
 - a. Annotate metadata documents with arbitrary RDF properties and ontology terms (uses the Data Annotation Services component)
 - b. Accept/reject annotations (uses the Data Annotation Services component)
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 (uses the Data Annotation Services component)
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) 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 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 will take place at the TDWG annual conference in Perth in October 2008 and will be an opportunity to identify collaborators and to understand wider interoperability needs.

Data Integration and Annotation Services in Biodiversity (ALA: DIAS-B) National eResearch Architecture TaskForce Project

- 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

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 (including metadata records) 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.

The following is a summary of the functional requirements for the ALA Data Annotation Services.

**Data Integration and Annotation Services in Biodiversity (ALA: DIAS-B)
National eResearch Architecture TaskForce Project**

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

Major Milestones and Target Dates

Milestones will be reviewed and refined as part of the requirements refinement workshop

Table 1: Data Integration Services milestones and dates

Milestone	Comments	Date	Staff member
Requirements refinement workshop	Held jointly with Data Annotation Services	August 2008	All
Implementation plan	Selection of core metadata registry	September	Project manager

**Data Integration and Annotation Services in Biodiversity (ALA: DIAS-B)
National eResearch Architecture TaskForce Project**

	technologies and detailed plan for implementation of components	2008	
Prioritisation of ontologies for tagging ALA metadata documents	Priority list of ontologies and vocabularies to be supported by ALA Metadata Repository	October 2008	ALA Metadata Curator
Core metadata repository implemented	Basic repository available as basis for development of other components, including basic registration of metadata	December 2008	ICT Centre developers (support from ALA developers)
Representation of ALA taxonomy as ontology	Web service providing access to names and classification of Australian organisms as an ontology	March 2009	External (ALA AFD/APC project)
Metadata harvesting components implemented	Harvesting of metadata from other repositories	April 2009	ICT Centre developers (support from ALA developers)
Metadata tagging components implemented	Interfaces (including web services) for tagging metadata documents with ontology terms	August 2009	ICT Centre developers (support from ALA developers)
Metadata search interface implemented	Search interfaces (including web services)	December 2009	ICT Centre developers (support from ALA developers)
Data retrieval interfaces implemented	Interfaces for intelligent discovery and access of structured data from resources catalogued in repository	December 2010	ICT Centre developers (support from ALA developers)

Table 2: Data Annotation Services milestones and dates

Milestone	Comments	Date	Staff Member
Requirements refinement workshop	Held jointly with Data Integration Services	August 2008	All
Implementation plan	Detailed plan for implementation of components	September 2008	Project manager
Annotation store and metadata schema repository implemented	Basic store available as basis for development of other components	December 2008	UQ developers
Prioritisation of metadata schema definitions	Priority list of XML document structures to support a range of user functions	December 2008	ALA Technical Architect
First priority annotation user interface implemented	User interface (widget) for first priority document structure released	June 2009	UQ developers (support from ALA developers)
Annotation retrieval services implemented	Web services and user interfaces for search and	June 2009	UQ developers (support from ALA

**Data Integration and Annotation Services in Biodiversity (ALA: DIAS-B)
National eResearch Architecture TaskForce Project**

	retrieval of annotations		developers)
Priority annotation user interfaces implemented	Interfaces (widgets) for all priority document structures released	June 2010	UQ developers (support from ALA developers)

Cost Estimates

The following table summarises planned expenditure of NeAT funds.

Table 3: Direct budget for subprojects for the next three financial years

Project	2008/2009	2009/2010	2010/2011	Total
Data Integration Services	\$200k	\$200k	\$100k	\$500k
Data Annotation Services	\$200k	\$200k	\$100k	\$500k
Totals	\$400k	\$400k	\$200k	\$1,000k

In the first 2 years, the NeAT funding will cover:

- 1.5 EFT for software developers at CSIRO ICT Centre to work on Data Integration Services
- 2 EFT for software developers of the UQ eResearch Centre to work on Data Annotation Services

In-kind resources will be provided through:

- The ALA will provide an additional \$70K each year to the Data Integration Services activity to enable support for two full-time developers, and up to \$30K each year to support a part-time project facilitator, as well as direct input from ALA staff (see *Project Approach* for more detail)
- CSIRO ICT Centre will pay overhead costs for the two Metadata Repository developers and will provide in-kind supervision for these developers
- University of Queensland will provide in-kind supervision for the Data Annotation Services developers
- CSIRO IM&T will make a range of skills available to contribute to the project (see *Project Approach* for more detail)

The continuation of the project in Year 3 and the amount of funding available will be subject to a project review. The costings for that year will be revised at that time.

Key Required Staff and Resources

Are there any individuals that are required for specific project activities. Are they currently available?

- All required staff are included within the project plan and/or are funded as part of the ALA project
- Integration of the project services within AAF frameworks depends on the overall progress and wider adoption of these frameworks

Risk Management Plan

What are the key risks to the successful delivery of the project?

How will the major risks to the project be managed?

**Data Integration and Annotation Services in Biodiversity (ALA: DIAS-B)
National eResearch Architecture TaskForce Project**

Table 4: Major project risks and their management strategy

Area	Specific risk/hazard	Management Strategy
Recruitment	Inability to recruit or keep development staff to work on project	All project elements will be developed with involvement of at least two team members to ensure that ownership is shared and loss of any individual is less significant
Managing client/stakeholder relationships	Clients/stakeholders not having ownership of outcomes leading to lack of uptake of outputs	The project will be publicised through the ALA's newsletters and through representation at the TDWG conference in October 2008. A project wiki will be established and stakeholders will be encouraged to participate.
Engagement of users	Data providers and users may not adopt the Metadata Repository; users may not adopt the Data Annotation Services	The key mitigation is to ensure publicity and understanding of project goals from an early stage (see previous risk). In addition the ALA team will work to identify partner projects which can incorporate the products of these services within their own infrastructure – this is central to the success of the ALA itself.
External technology development	Other projects may develop superior or more widely adopted technologies and standards which render this project obsolete	The project team and ALA Scoping Group will monitor related activities and will continually review these against the offerings from this project. The project should be prepared to reuse components from other sources if identified to be suitable and/or superior.

Constraints and Assumptions

What are the time, budgetary, resource and quality constraints on the project?

What major assumptions have been made in planning the project and estimating?

- The project assumes the availability early in 2009 of a web service interface to the Australian Plant Census (APC) and Australian Faunal Directory (AFD) data sets to serve as the basis for tagging metadata documents with standard taxon names. This project is also being funded by the ALA. In the case of a failure in this project, the ALA will use the Catalogue of Life global species list in its place.

Open Issues and Pending Decisions

Are there any open issues that need resolving before the project can start?

Are there any key decisions that the project team is dependent upon being made?

- None identified

APPENDIX A: NeAT Project Governance

All NeAT projects should aim to establish services that are useful both for the discipline involved and as potential national services. There should be only two levels of governance, where the distinction is clear between the governance and the deep technical and domain involvement needed for the project to succeed.

ARCS and ANDS have therefore discussed and jointly agreed on the management of NeAT Projects as follows:

- Each NeAT Project will have a NeAT Project Committee consisting of an ANDS representative (the Executive Director or delegate) and an ARCS representative (the Executive Director or delegate), representatives from any other institutions that would manage the enduring services provided by the NeAT Project, community nominated discipline representatives, a designated NeAT Project Manager (ex officio) and a prominent discipline leader as the NeAT Project Committee Chair. Where a suitable discipline Chair could not be found, the Chair will be either the ANDS or ARCS representative depending on whether the project was more ARCS or ANDS;
- Each NeAT Project will have a Project Manager selected by the relevant NeAT Project Committee;
- The Project Manager must be the person who manages the day to day work of the project;
- Project Managers must report to and be directed by the Project Committee;
- The governance structures of ANDS and ARCS will need to be satisfied with the Project Committee's management of the project in order to ensure the funds keep flowing, which provides the appropriate checks and balances and ensures accountability;
- At the start of the Project and subsequently once each quarter the Project Manager will attend a meeting chaired by the AeRIC Executive Director and attended by the Executive Directors of ANDS and ARCS and their nominees as well as the Project Managers of the other NeAT Projects;
- The Project Manager must meet no less than every four weeks with the Project Committee: in order to discuss the progress and evolution of the Project; to ensure that the Project is making optimal use of existing and planned services of project participants; and to ensure that the Project is being developed in a way consistent with the long-term delivery of the Services as per the project plan;
- Core responsibilities of each of the NeAT Project Committees include: overseeing and approving the design and implementation of an appropriate and relevant enduring service; and at the end of the Project identifying the key stakeholders and service providers to manage this enduring service into the future and to take over from the NeAT Project Committee.
- ARCS and ANDS will jointly review the progress of each NeAT Project every three months using their standard processes and the NeAT Project Committees would review their project every six months with a written report from the Project Manager. NeAT would review all NeAT Projects annually in September, beginning 2009, as part of the established NeAT processes.
- ANDS and ARCS will provide NeAT Project funds quarterly in arrears based on acceptable performance on a per EFT basis for each NeAT Project. The ANDS and ARCS quarterly reviews will be the trigger for either approving or withholding NeAT funding for that quarter from a NeAT Project or a component of that project as appropriate.