NeAT Business Plan Component
Data Integration and Annotation Services in Biodiversity (DIAS-B)

1. Service Description

1.1. Description of a research community and the eResearch service need

The Atlas of Living Australia (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 – from many different data providers. The user community for the ALA is very broad, encompassing 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.

In order to provide discovery and interoperability across many and varied biodiversity data sets, the ALA requires needs best practices for metadata management, including adoption of relevant vocabularies and ontologies, and the ability to map between different metadata models. A Metadata Repository is required to enable metadata registration and harvesting for all available digital resources of biodiversity information.

The quality and consistency of the ALA data is crucial for its use. There is a need for an authenticated annotation service that will allow users or automated data analysis tools to provide information to users and feedback to data providers by annotating data records and resource metadata with comments on data quality and suggested corrections.

1.2. Description of the proposed service solution and how it meets that need

Data quality services:

- Annotation service allowing human and machine users to store and retrieve annotations relating to any data record within the ALA to record possible errors.
- Reporting service that alerts data provider/owners of possible quality issue.

Data integration services:

- Catalogue of mandated and supported data standards, vocabularies, ontologies for use within the ALA.
- Metadata repository and metadata registration software for registration of all Australian biological data resources and for relating data sets to supported vocabularies and ontologies.
- Search interfaces (including web service interfaces) to search the metadata repository using terms from the supported ontologies.

Together, these services will enable data providers and researchers to actively participate in the creation and use of the Atlas of Living Australia.

These services will initially be hosted by ALA, however if these services can be made more generic they may be hosted by ARCS on behalf of ANDS.

2. Benefits and proposed measures

2.1. Benefits to the user community and associated measures

Easier and faster integration of new data sources into ALA, measured by:

- Level of provision of metadata to metadata repository

- Number of data sets accessible via ALA
- Number of records accessible through ALA.
- Usage of ALA

Improved data discovery and federated search, measured by:

- Direct use of metadata services from the repository by other networks and repositories
- Percentage of data resources with metadata entries including references to ontology terms
- Extent of discovery through ontologies

Improvement of data quality via use of annotation services, measured by:

- Number and range of annotations in annotation database
- Number of responses from data providers
- Direct use of services (UI and web services) for providing annotations (other than through the ALA portal UI and ALA data validation tools services)
- Direct use of services (UI and web services) for accessing annotations
- Number of data records for which annotations have led to corrections in source data

Improved level of user experience, measured by:

- 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.
- An online survey tool to allow users to document their experience in using the ALA infrastructure. This survey tool will be continuously available as a data capture method. This survey will explicitly determine success in using data quality annotation.
- Analysis of web logs to determine whether users are guided to relevant information.

The Project Plan and the ALA Business Plan will specify some quantitative goals for these metrics for each year of the project.

2.2. Benefits to ANDS or ARCS (or other provider) and associated measures

There are two key benefits being sought. Firstly, an ability to integrate several different data sets with different schemata or ontologies, enabling researchers to find things despite having different knowledge lenses – this will be measured using the surveys described above. Secondly, we wish to understand how well an annotation service supports improvement of data quality – it will be measured using the survey described above.

2.3. Expected flow-on benefits to others

As articulated above, this affects all disciplines where commentary on other work is important. More specifically, the services developed within this project would also be of benefit to:

- NCRIS 5.12 Marine Sciences and Climate
- Social Sciences, ASSDA, AustLit
- NCRIS 5.3 Microscopy and Microanalysis
- NCRIS 5.8 Bio-security
- NCRIS 5.11 Terrestrial Ecosystem Research Network

3. Resources and commitments

3.1. Resources provided by the user community

The Atlas of Living Australia will have approximately 5 EFTs working on software development related to this project. The Australian Biological Resource Survey (ABRS) and the Australian Museum also have existing developers whose products will be contributing directly to the development of the Atlas.

3.2. Resources provided by ANDS or ARCS

NeAT funding of $400K p.a. for 2 years and nominally $200K for a third year. The actual amount for the third year will be dependent on the outcome of project reviews and available NeAT funds.

It is anticipated that the NeAT funding for this project would be used primarily to hire software developers at sites with relevant expertise, possibly including CSIRO, SAPAC, ANU and UQ. This will be decided by the Project Committee and specified in the Project Plan.

3.3. Resources provided by others

There will be significant related international effort in standards development by the Taxonomic Data Working Group (TDWG) and open source software development effort from members of TDWG, GBIF and EoL (the Encyclopedia of Life). It is expected that some of these standards and software will be utilized in the NeAT project. This effort is difficult to quantify and not included here.

Total project resources and commitments are summarized in the following table.

	Y1		Y2		Y3		Total	
	Cash	EFT	Cash	EFT	Cash	EFT	Cash	EFT
ALA		5		5		5		15
ANDS	200K		200K		100K	1	1.0M	1
ARCS	200K		200K		100K			

4. Governance

4.1. Governance processes to be applied to the project

- The ARCS/ANDS agreed governance mechanism for NeAT projects, defined in the ARCS and ANDS Business Plans.
- The Project Committee will meet quarterly via phone and/or agreed electronic medium.

4.2. Quality assurance processes to be used by or applied to the project

The ARCS/ANDS agreed arrangements will apply.

4.3. List of names against key governance and project management roles

Project Committee:

- ARCS Executive Director, Professor Anthony Williams, or nominee
- ANDS Executive Director, or nominee
- Director of the Atlas of Living Australia, Donald Hobern
- Chair to be appointed.

The Project Manager will be appointed by the Project Committee and specified in the Project Plan.

5. Project Summary

5.1. Deliverables / Milestones

Metadata Repository Activities - Year 1:

- 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.

Metadata Repository Activities - Years 2-3:

- 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 Activities - Year 1:

- 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.

Annotation Service Activities - Years 2-3:

- 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.

5.2. Overall risk assessment

Risk	Mitigations
Take up is slow	• Measure take up • Project leader to be responsible for identifying projects and collaborations to provide and consume metadata and annotations
Poor software	• Good people/place – specifically oversight of experts from UQ, SAPAC and ANU
Other approaches are more attractive	• Keep watching and be adaptive in project development to make best use of developing standards and practices

5.3. Review points

Quarterly reviews by ANDS and ARCS, six monthly written reports from the Project Manager to the Project Committee, and a yearly review each September (starting 2009) by NeAT.