## Data Ingest and Management Information Workflows In the Metadata Store

## Architectural Context

The metadata store is implemented as a set of enhancements to the existing ANU data commons.

It is not conceived as a standalone application – rather it adds functionality to an existing platform and reuses existing workflows and code where appropriate.

This document provides supplemental information regarding the data ingest and Management Information workflows. Infromation is provided within the technical documentation that allows the development of client applications.

The major advantage of this approach is extreme flexibility and ease of integration with existing college and university systems. The flexibility allows the easy integration of additional data sources – it also provides a single source of data allowing developers to access an aggregated source of data rather than having to combine disparate data sources.

The metadata store is exactly that – a set of databases containing data aggregated from a range of sources.

## Data Sources

Aries – researcher identification, publication, and grants held

LDAP – researcher name, affiliation, status

Data Commons – datasets deposited, people, activities and services

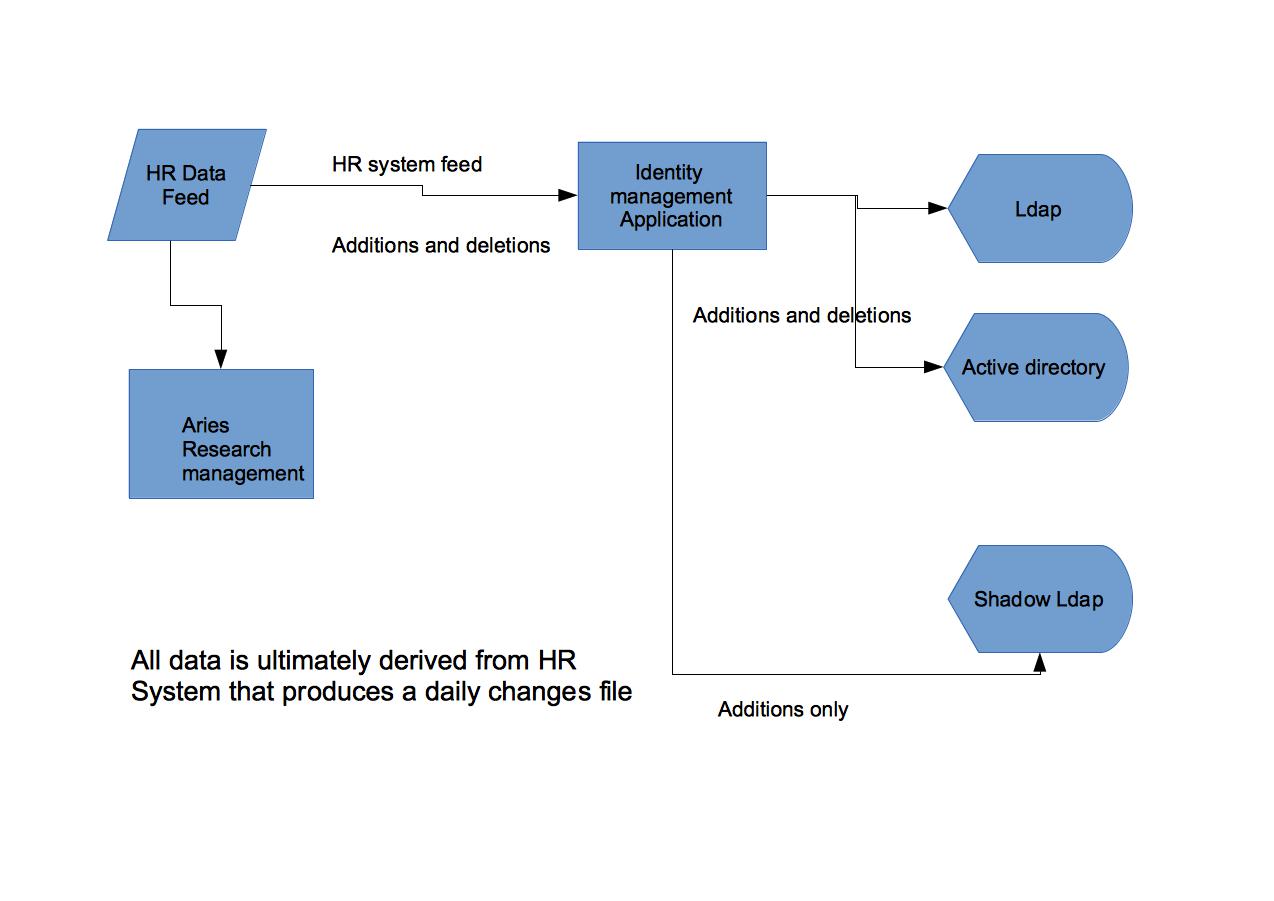
Digital Collections – preprints and other material

The data sources selected were chosen to allow us to build a comprehensive picture of research activity. The LDAP service was chosen rather than the University's identity management service as the identity management service was under review at the time as part of a major review of the University's identity management strategy.

It was judged that whatever changes were made to the identity management system, the LDAP service would remain as a downstream service from the identity management application.

The Identity management data is directly derived from the HR system, which is also under review as part of the University's identity management strategy

The information flow can be represented as:



This gives the university a unified namespace to uniquely identify staff and students on the basis of their university identifier and associated LDAP attributes.

A complication is that the LDAP directory service does not have a memory of people who were formerly members of the University and who have now left. Thus if a university identifier occurs in Aries but not in LDAP we make the assumption that the individual is no longer associated with the University.

As the ANU has now moved to issuing academic transcripts digitally, and as the transcript delivery system contains a shadow LDAP system that retains details of expired accounts, a future enhancement will be to refine the process to involve a query of the shadow LDAP to confirm that the account is a valid expired account, and that the party involved had an association with the ANU.

There remains the unresolved issue that unless an author has created an external identifier such as an NLA party identifier prior to leaving the university there is no way we can map a publication or dataset to a current identifier.

## Project outputs

As the University’s knowledge management strategy is developed, this project would support a capability building exercise that would accelerate outputs of the system which align with the ANDS metadata store program.

In particular this project assists with:

· Provision for researchers to inform the University at the appropriate stage of the research data cycle on the existence and basic provenance of a research data collection; for example, to assist judgments about the need to retain data and its citation rate

· Integration of information held in the research management system, Digital Collections repository and Data Commons repository to allow matching of research papers to published datasets and the inclusion of grey literature and working papers as appropriate

· Integration of information held in the Digital Collections repository and Data Commons repositories with the University’s identity management systems to tie both data and publications to individual researchers and further integrate it with existing Research Management systems to tie the publication of research papers and the deposit of data with research grants held thus giving a comprehensive view of scholarly activity

## Architecture and implementation

In essence the system is very simple in terms of design:

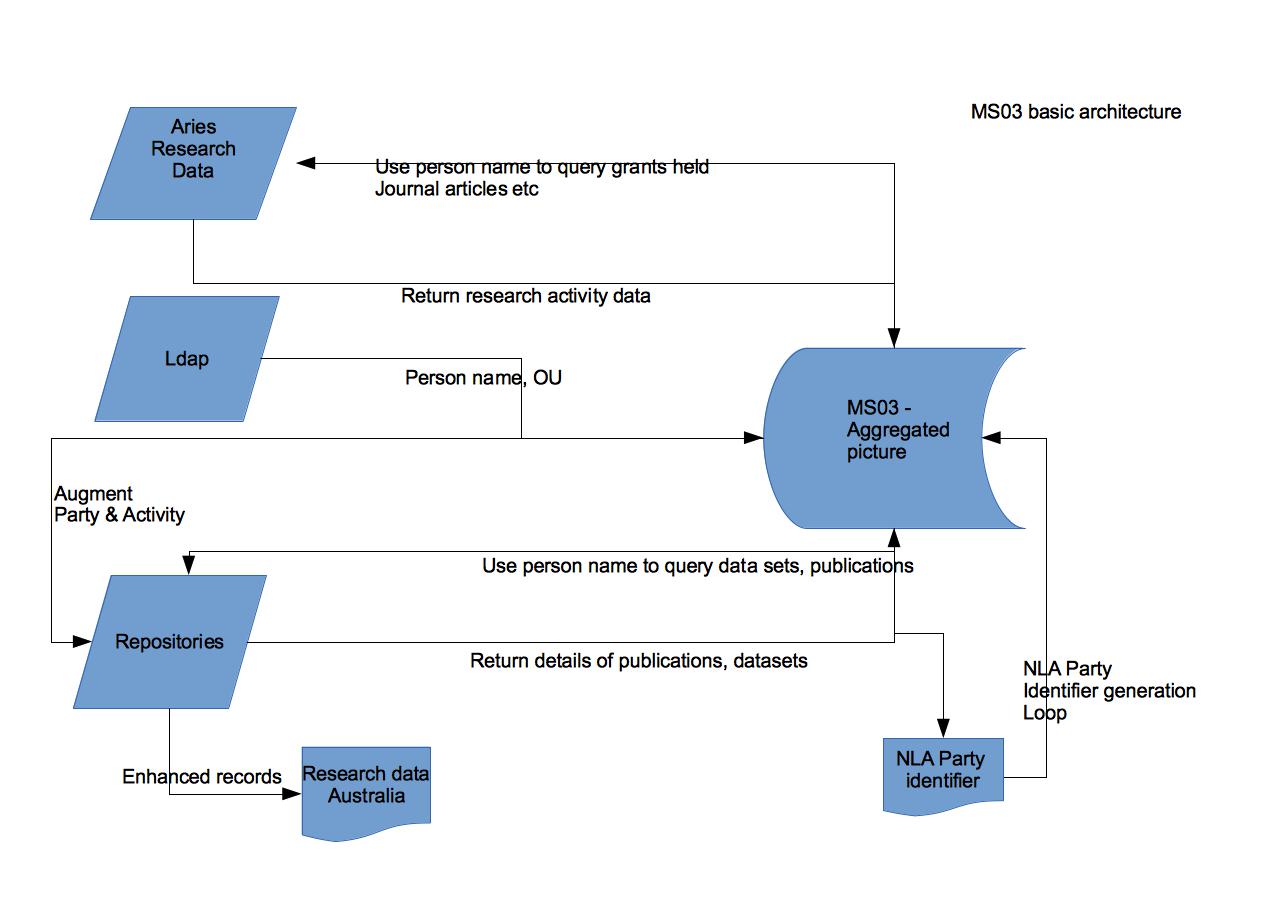
Information is taken from a range of sources

Information is aggregated in a set of postgres databases

Information is made available to a range of downstream systems

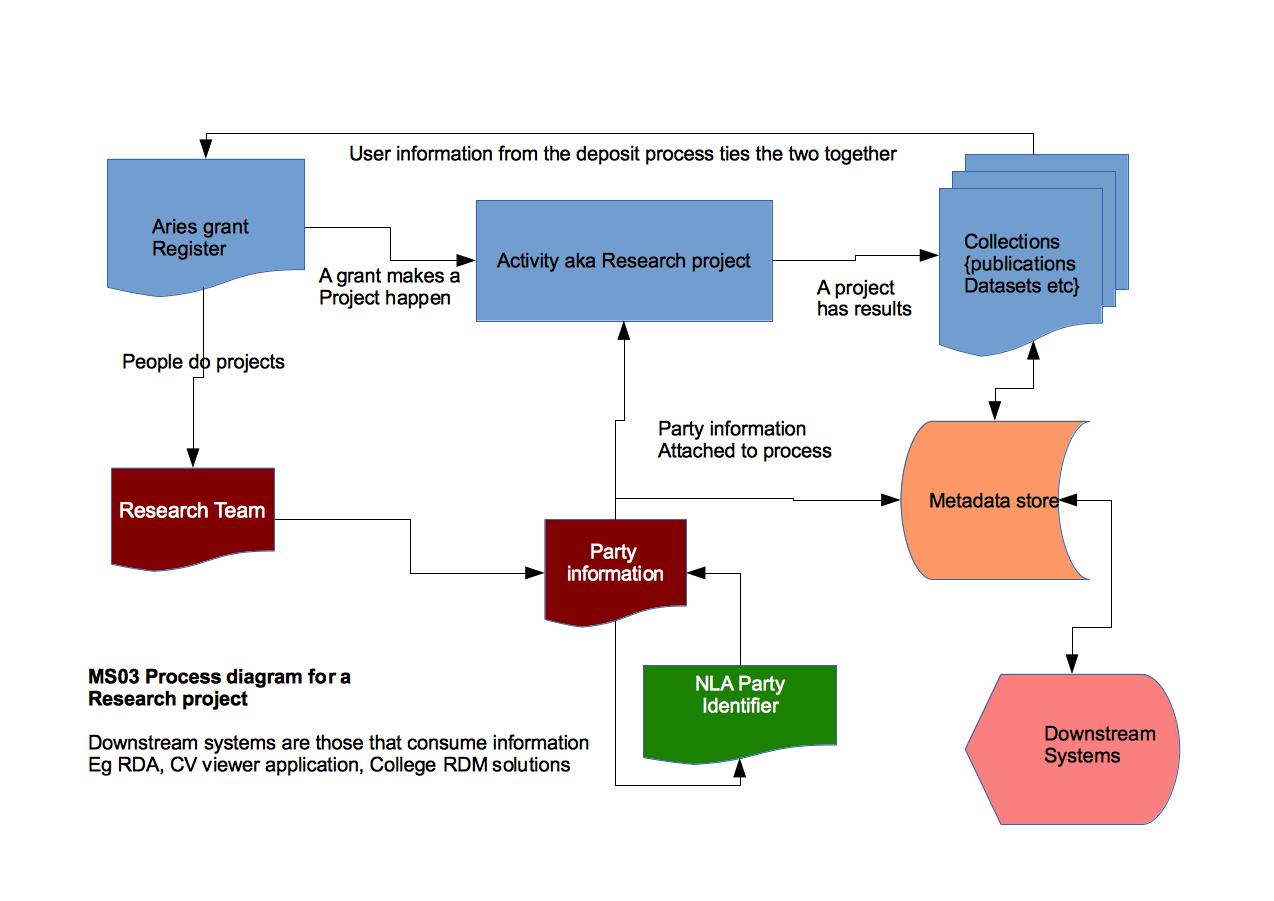
These downstream systems may be either client applications or other data aggregation/management systems

In order to maintain architectural simplicity data was harvested using either the data source's api in the case of LDAP, by a range jdbc queries in the case of Aries, and by polling the OAI-PMH feeds in the case of the Digital Collections print repository and the Data Commons.



## Collection registration workflow

In the initial design stages of the project we viewed the metadata store as a variant on a basic warehouse design but one in which included cross linkages between the various data sources to allow the building of a picture of scholarly data around research projects:

This workflow mimics existing workflows within the Research Services Directorate. Essentially a researcher cannot accept a grant without informing the RSD, as only the RSD can accept funding on behalf of a researcher and create the appropriate ledger codes to manage the funding.

Within ANU this is referred to as a project and is directly analogous to the ANDS concept of an Activity.

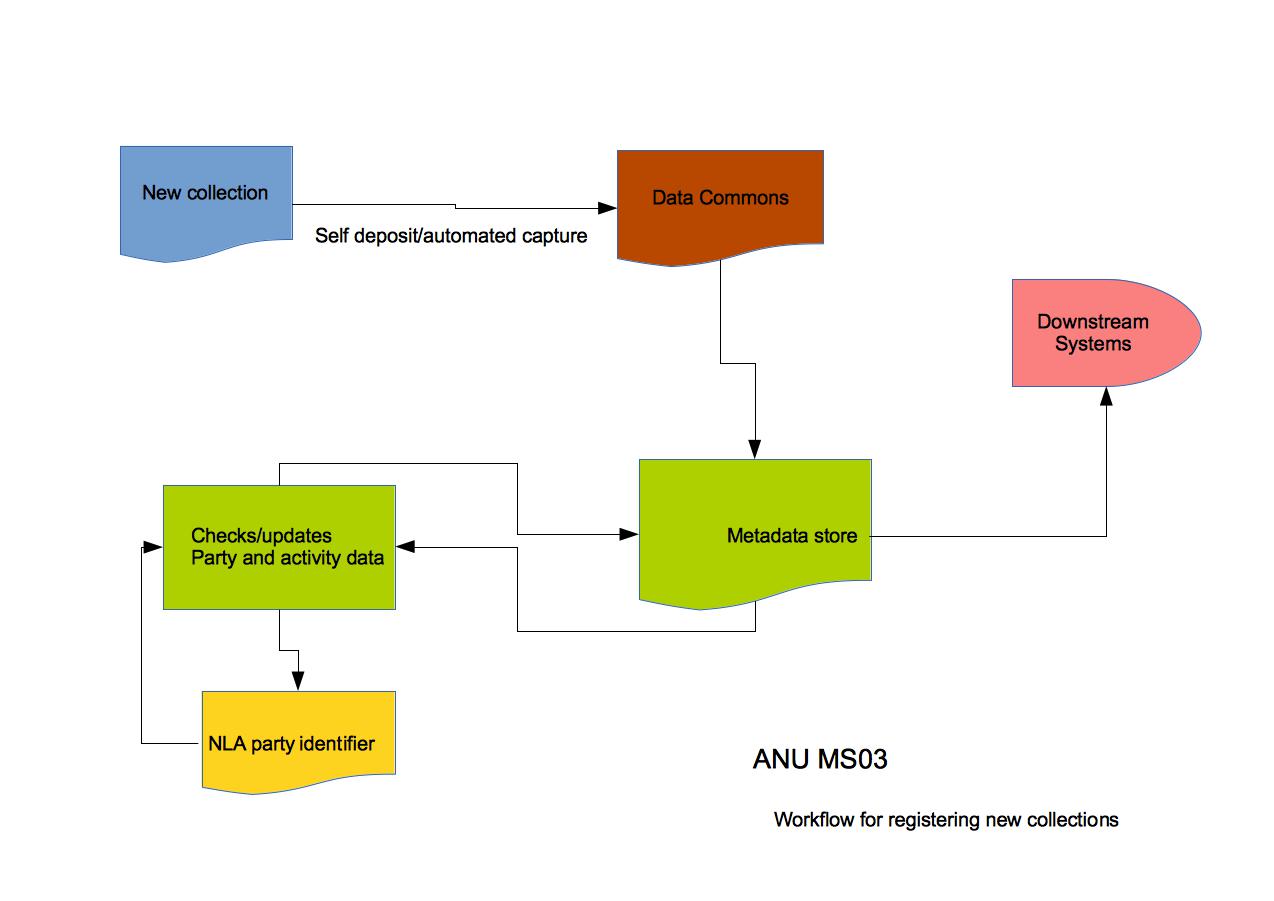
With an activity based view on can say that

A project has inputs (Funding, Researchers who do the work) and outputs (Datasets and Publications)

This is however, not the only way that datasets can be deposited.

Particularly in the Arts and Humanities, researchers can generate a data collection independent of any formal research activity, which they then organize and annotate using a variety of tools.

Equally, researchers in all disciplines can have legacy datasets that they wish to deposit, or collections such as long-term survey datasets that are automatically deposited. In such cases the following workflow would apply:



In this scenario researchers are invited as part of the dataset ingest process to associate party and activity data with the data set. Where there is no formal activity associated with the data there is no requirement to do so.

## Strategic reporting of information

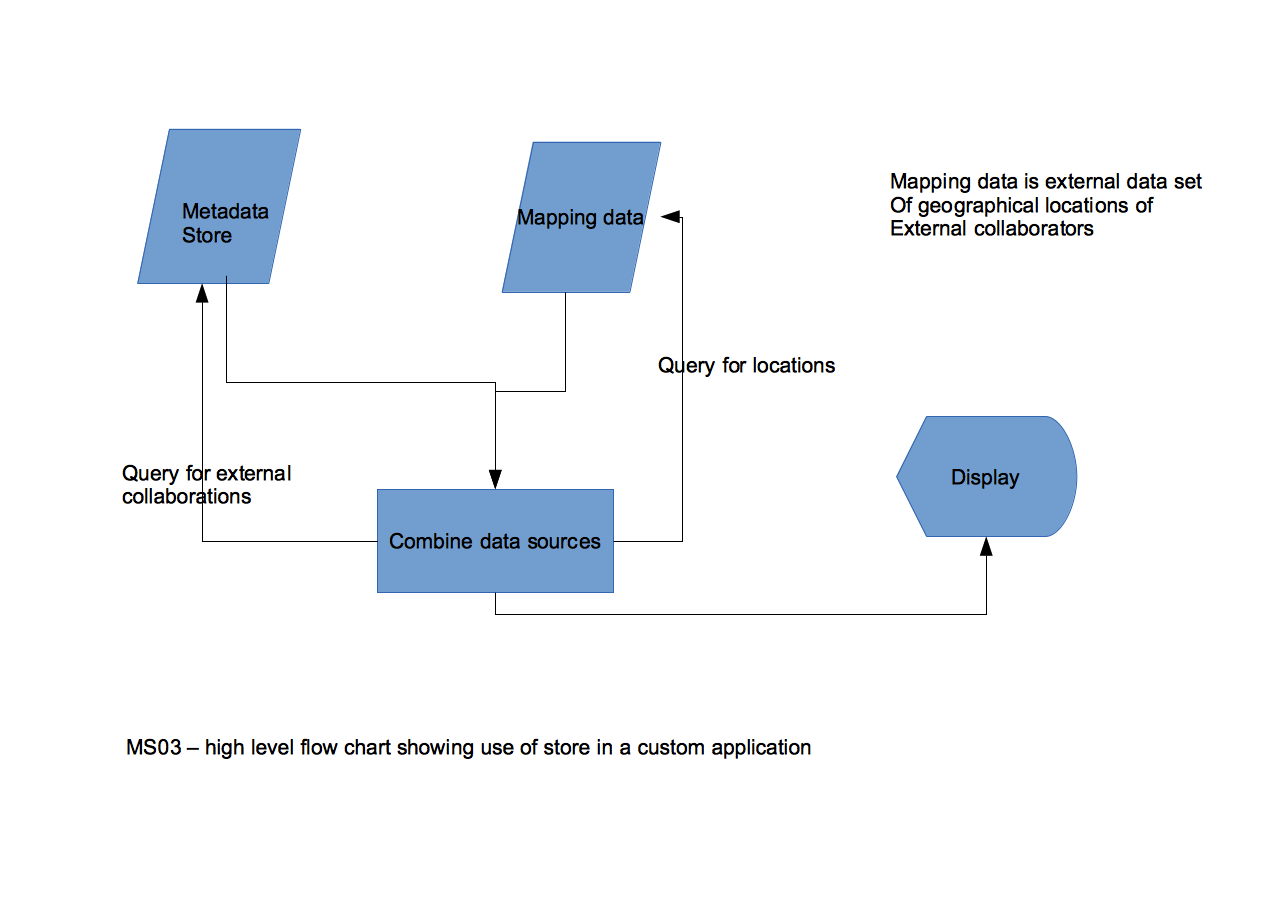
Jasper reports, a well known open source reporting tool is used to obtain a range of information to display updates changes and the like and the output from Jasper reports can be readily imported into a range of products for further analysis by means of a CSV export.

The reports are executed through jasper reports ([http://community.jaspersoft.com/project/jasperreports-library](http://community.jaspersoft.com/project/jasperreports-library" \t "_blank)).  I create/design the reports through iReport ([http://community.jaspersoft.com/project/ireport-designer](http://community.jaspersoft.com/project/ireport-designer" \t "_blank)) which uses the jasper reports engine to run the reports while design is in progress.

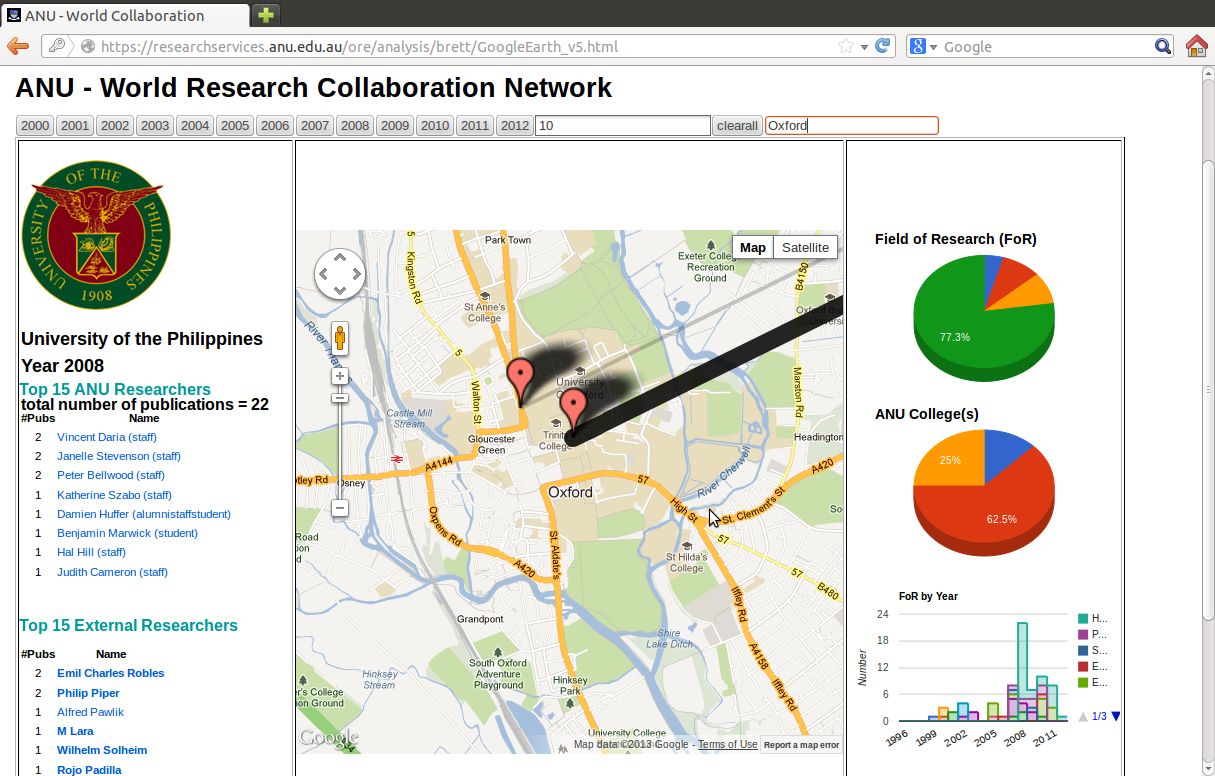
The solution is designed primarily to allow colleges and other users of the system query the metadata store and display data within their own application rather than relying on inbuilt reporting mechanisms.

To demonstrate the flexibility inherent in the separation of client facing capabilities from the store an application was developed independently of this project, and separately financed, to display the distribution of ANU collaborations around the world.

This application had the following basic architecture:



In it, the application queries the metadata store for information on activities where one of the grant holders is external to the ANU, combines this data with set of data specifying the geographic locations of each collaborating institution and plots the display using Google Maps.



This data is then plotted on Google maps to give a view of where in the world are the institutions that the ANU collaborates with. Drilling down allows a user of the application to go as far as the individual institution and to list the top 15 ANU researchers and top 15 collaborating external researchers on the basis of citation rates, and to identify the nature of this collaboration on the basis of Field of Research codes.

This provides valuable information on the nature of collaborations and the extent of collaborations with a particular institution in a visually attractive manner.