A system for extracting and updating common organisational metadata into a central authoritative source

A description of the methods and technologies developed as part of a CSIRO Information Management & Technology project funded by the Australian National Data Service

Daniel Miller, Kavinga Hubert, James Dempsey, Sean Tregeagle, Sathish Sathya-Moorthy, Brendan Sullivan
3 October 2012
Citation

License
Creative Commons Attribution-Noncommercial-Share Alike Australia 3.0. You are free to copy, distribute, display, and perform the work; and to make derivative works; under the following conditions: Attribution - You must give the original author credit; Non-Commercial - You may not use this work for commercial purposes; Share Alike - If you alter, transform, or build upon this work, you may distribute the resulting work only under a licence identical to this one. See the following URL for further details: http://creativecommons.org/licenses/by-nc-sa/3.0/au/

Important disclaimer
CSIRO advises that the information contained in this publication comprises general statements. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional and technical advice. To the extent permitted by law, CSIRO (including its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.
Acknowledgments

This work has been produced as part of a project that is supported by the Australian National Data Service (ANDS). ANDS is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy Program and the Education Investment Fund (EIF) Super Science Initiative.
1 Introduction

1.1 Purpose

This document describes a system for extracting and updating common organisational metadata into a central authoritative source. It provides a description of the methods and technologies developed as part of a CSIRO Information Management & Technology (IM&T) project funded by the Australian National Data Service (ANDS).

1.2 Background

Historically, CSIRO researchers have been inputting the same organisational metadata fields into different enterprise systems manually.

The intent of this project will be the enhancement of the existing Org-Data-Central system to centrally manage metadata, and provide links so that metadata can be re-used and associated with data stored in repositories or support tools.

The target outcomes of the system include:

- Increased use of authoritative sources for standard metadata;
- Increased quality of metadata records;
- Increased use of metadata across CSIRO’s data;
- Decreased effort to ‘tag’ information sources with metadata.

1.3 Target Audience

The target audience for this document includes, but is not limited to the following:

- Members of CSIRO IM&T responsible for the operations and maintenance of the ODC system;
- Business areas within CSIRO that gain benefit from the ODC system;
- Members and partners of ANDS;
- Organisations that wish to develop and operate their own system that enables collection of common organisational metadata into a central authoritative source.
2 Organisational Metadata

The primary purpose of Org Data Central is to extract organisational metadata from SAP and make it available for easy use to the rest of CSIRO's applications. This is currently focused around hierarchy data describing CSIRO's matrix organisational structure. This structure is divided into two sides:

**Inputs/Resources** (aka left side) - This is the structure of business units and research groups, including line management. This is represented in ODC as Business Units, Research Programs, Research Groups and Cost Centres.

**Outputs** (aka right side) - This is the structure of portfolios in which research projects are managed. This is represented in ODC as Groups, Portfolios, Themes, Streams and Projects.

![Organisational Structure Diagram](image-url)
3 System Context

- SAP: Authoritative system for output hierarchies
  SAP provides authoritative output hierarchy information to the ODC system.

- ODC system: Organisational metadata system
  Output and input hierarchies are maintained within this system. Output hierarchies are read nightly from SAP and stored in a database. They can also be refreshed on demand by the ODC admin users. Input hierarchies were initially loaded from SAP but are currently manually maintained. External systems are able to consume this hierarchy information via the SOAP interface or the newly built REST interface. ODC users are able to search and browse this hierarchy information. They are also able to download these extracts in excel format.

- DAP (and ANDS via DAP), ePublish: Systems consuming metadata from ODC
  There are number of systems consuming metadata from the ODC system. These systems can either use the existing SOAP interface or the newly built REST interface. Currently, most clients also do this on a nightly basis, storing the structures locally for reuse during the day.

- Active directory: Provide staff information
  SAP provides authoritative staff information to the active directory system. It is then used by various systems as a lookup for staff information.
4 Architectural Decomposition

![ODC Architecture Overview](image)

**Figure 3: ODC Architecture Overview**
4.1 Architecture Components

**User Interface** - A web interface to allow users to browse the data, and to allow privileged users to administer the system. While the system is built mainly using JSP technology, some aspects have been built using ZK technology. In the future this will also allow CSIRO staff to enter further information into the system.

**SOAP API** - An interface to allow other CSIRO applications to retrieve the organisational data via the Simple Object Access Protocol. This requires a reasonable amount of coding to interact with and is commonly used by systems written in Java, C#, C etc.

**REST API** - A proposed interface to allow other CSIRO applications to retrieve the organisational data via RESTful web services. This is relatively easy to use and is commonly used by systems written in JavaScript, Python, Java etc.

**Business Rules** - The encoding of the rules for organisational data.

**Data Access** - Definitions of data access objects and database access services to allow the storage and retrieval of organisational data. The Hibernate library is used along with Spring to provide a high level of independence from database vendors. Despite this we will continue to use Oracle database for the project.

**Security** - Interface to CSIRO’s active directory instance (aka NEXUS) used for authentication of users and extract of user information such as name, email etc.

**SAP Interface** - Read-only interface to CSIRO’s SAP implementation. This is the primary source of output hierarchy data. ODC calls SAP ABAP functions using the SAP Java Connector (JCO).
5 Application Structure

The Org Data Central application is a Java web application. It is hosted on a Apache Tomcat web container proxied by the Apache 2 HTTPD server.

The application is built on a number of important libraries:

- The Spring Framework underlies the application structure. It is used to manage dependencies between different parts of the code, to enable testing of each part of the code in isolation and to manage application security. [http://www.springsource.org/](http://www.springsource.org/)
- The ZK Web Framework is used to provide a rich web interface for the application. This allows us to produce a highly interactive user interface over the web without delving deeply into JavaScript and raw AJAX coding. [http://www.zkoss.org/](http://www.zkoss.org/)
- The Hibernate library is used for all interaction with the database. It allows the database code to be portable and easily secured. [http://www.hibernate.org/](http://www.hibernate.org/)
- The XFire library is used to provide the existing SOAP web services. This is likely to be replaced with the use of CXF during the current development phase. [http://xfire.codehaus.org/](http://xfire.codehaus.org/)
- The SAP Java Connector (SAP JCO) native library is used to interact with CSIRO’s SAP implementation.
6 Interfaces

6.1 Import interface

6.1.1 SAP Interface

Output hierarchy

‘Project’ and ‘Project Abstract’ functional modules are available in SAP to provide required output hierarchy information. This information is collated from the project definition and level 1 WBS levels in SAP. All active projects and groups, themes, portfolios, streams associated with active projects will be retrieved by these function modules.

SAP cannot be updated via this interface. Following diagram is a depiction of how ODC output hierarchy is populated from these two SAP function modules.

![Diagram of SAP Extraction of Output Hierarchy](image)

**Figure 4: SAP Extraction of Output Hierarchy**

Input Hierarchy

Input hierarchies were initially loaded from SAP but are currently manually maintained.
6.2 Export interface

6.2.1 SOAP interface

Using the SOAP interface, external systems can retrieve input/output hierarchy information from the ODC.

Summary: One method per each extract is available within this interface. Using the appropriate method, users can retrieve all records from the required extract.

Logging: Server logs are used to record start/complete/error information.

Permissions: No authentication required.

6.2.2 REST interface

Using the REST interface, external systems can retrieve input/output hierarchy information from the ODC.

Usage:

Summary: There are number of methods available in this interface. They can be categorised into two main types:

- Project requests
- All other extract requests

There are number of project requests:

- Retrieve all project records
- Retrieve project records based on a criteria (see diagram for the list of criteria)

There are number of other extract requests:

- Retrieve all supported hierarchies
- Retrieve all records of a selected extract (using the correct extract Type)
- Retrieve records based on a criteria (see diagram for the list of criteria)

Logging: Server logs are used to record start/complete/error information.

Permissions: No authentication required.
6.3 User interface

6.3.1 Search

Search functionality offers two types of search

- Basic search
- Boolean search

Both input and output hierarchies are searchable. Users can perform common search related tasks such as results filtering and sorting using the ODC user interface.

Basic search

Users can enter a single or multiple terms. By default, Search combines terms with AND, so that the more terms you enter, the more specific are the Search Results. Thus a search for CSIRO Energy will return all records that contain the words CSIRO AND Energy in any of the searchable fields.

E.g. Both of the following scenarios will return positive results in this search

- Both CSIRO AND Energy in one searchable field
- CSIRO in one fields AND Energy in another field

If the search term is enclosed within quotation marks, search will return all records containing the exact search term in any of the searchable fields.

Boolean search

To activate Boolean operators (i.e. AND, OR, and NOT to behave like operators and not as keyword), user is required to type them in capital letters.

Results are described in the following 5 examples:

- apples AND oranges
  returns results if both "apples" and "oranges" are found in a record; returns nothing if either apples or oranges have no matches in any of the records
- apples OR oranges
  returns results if either "apples" or "oranges", or both "apples" and "oranges" are found; returns no results only if "apples" and "oranges" both have no matches
- oranges NOT apples
  returns all results for which "apples" has no matches but has matches for oranges.
- apples AND oranges NOT lemons
  returns all results if both "apples" and "oranges" are found in a record but not "lemons";
  returns no results if either "apples" or "oranges" have not being found or if "apples" or "oranges" are found, by they have also word "lemons" in them.
- apples OR oranges NOT lemons
  returns results if either "apples" or "oranges", or both "apples" or "oranges" are found and they do not have word "lemons"; returns no results if "apples" and "oranges" are both with no matches or if they have matches but they have word "lemon" in them

Wildcard search

- "?" And "*"

Question mark '?' can be used to replace a single character. This will be useful where people are looking for a name with a spelling variation that involves only one change of letter. For example Sm?th, will return results for both Smith and Smyth.
• The asterisk '*' can be used to represent 0 or more characters. For example, where a 'Mc' surname may have been recorded as Mac or Mc, a wildcard can be used to find both variants - M*cDonald shall find both McDonald and MacDonald.
• Limitations: wildcard characters ('*' and '?') cannot be used as the first character in a query.

6.3.2 Browse

Users are able to view all records of a selected input or output hierarchy in a table format.

6.3.3 Download

Users are able to download extracts from both input and output hierarchies in excel format.

6.3.4 Hierarchy maintenance

Users are not able to update the output hierarchy refreshed from SAP. If any changes are required in these extracts, SAP must be updated. The input hierarchy was originally refreshed from SAP but is currently manually maintained by administrator users.
A system for extracting and updating common organisational metadata into a central authoritative source.