Finding the Right Rabbit to Pull Out of the Hat: Data Management in CSIRO

Tricia Kelly[1], Gerry Ryder[2], James Dempsey[3], Cynthia Love[4]
[1]CSIRO Information Management & Technology, Rockhampton, QLD, Australia, tricia.kelly@csiro.au
[2]CSIRO Information Management & Technology, Adelaide, SA, Australia, gerry.ryder@csiro.au
[3]CSIRO Information Management & Technology, Canberra, ACT, Australia, james.dempsey@csiro.au
[4]CSIRO Information Management & Technology, Melbourne, VIC, Australia, cynthia.love@csiro.au

INTRODUCTION

CSIRO is one of the world's largest and most diverse research agencies with staff located literally from one end of Australia to the other as well as internationally. As both a creator and a consumer of research data, CSIRO faces considerable data management challenges. To this end, the development of the CSIRO Data Management Service (DMS) Repository is a pivotal step in the right direction for managing CSIRO-generated data, third-party data and establishing vital links with research community portals such as Research Data Australia and the Atlas of Living Australia. From metadata mapping to collector and conversion tools, this presentation will discuss the experiences of the CSIRO Information Management & Technology (IM&T) team in applying new services and technologies to address the challenges of discovering, exchanging and re-using research data.

METADATA FRAMEWORK

The broad scope of research undertaken in CSIRO results in a considerable range of data types including collections of wildlife specimens, soil samples, interview transcripts and surveys, sensor output, small and large instrument outputs such as microscopes, telescopes and marine research vessels. Managing this diversity in data types is one challenge; an associated challenge is managing the specific metadata schema applicable to the various research disciplines. the metadata framework being developed by the Data Management Services team had to be rigorous enough to support the breadth of requirements identified below:

- To enable sharing of data with community portals supporting relevant search parameters, multiple discipline specific metadata schema will be supported (ANZLIC and Resource Metadata for the Virtual Observatory initially but with others such as Darwin Core and MIAME on the near horizon)
- To enable CSIRO metadata to be harvested by the Australian Research Data Commons (ARDC), all supported metadata schema will be mapped to RIF-CS
- To facilitate cross-schema and system searching, a core metadata set based on Dublin Core will be identified and mapped from all supported schema
- To enable sophisticated reporting and improved internal discovery, a number of CSIRO extensions (CSIRO Common record) will be incorporated in the metadata profiles. For example, project codes and Research Division
- To explore and exploit options for interrogating and re-using metadata captured in other CSIRO business systems adhering to the old maxim of "create once, use many times"
- To develop a standard approach to documenting, mapping and transforming metadata based on extensible, re-usable tools.

The documentation describing the metadata framework is a key input for the development team to automate conversion of the discipline specific metadata schema to other supported schema such as RIF-CS and Dublin Core. The conversion and collector tools are a key enabler in the research data discovery, exchange and re-use process.

CONVERSION AND COLLECTOR TOOLS

The development work for the conversion and collector tools has been undertaken using a flexible, incremental approach based on the Agile Development methodology which supports change and evolving requirements – a common situation in the fluid environment of research data management. The focus is on enabling automated conversion of discipline specific metadata schema to RIF-CS and Dublin Core. The conversion strategy was applied in the first instance to converting the CSIRO profile of the ANZLIC metadata schema to RIF-CS. The metadata mapping spreadsheet developed as part of the metadata framework was converted to an XSL stylesheet based heavily on the existing ANZLIC_RIF-CS stylesheet in GeoNetworks. The html callout capability of XML library (SAXON) was used to allow party processing via a servlet. A step-by-step conversion process was then developed which will be outlined in more detail in this

presentation. The result of this development work was the Dataset Metadata Collector - a custom Java application that reads metadata on datasets from a variety of sources, produces RIF-CS records, mints Persistent Identifiers where necessary and adds (or updates) them to the CSIRO Repository. From there, metadata records can be exposed for harvesting by community portals and aggregators such as the ARDC. The interaction of the Dataset Metadata Collector with the internal and external data 'ecosystem' is depicted in Figure 1.

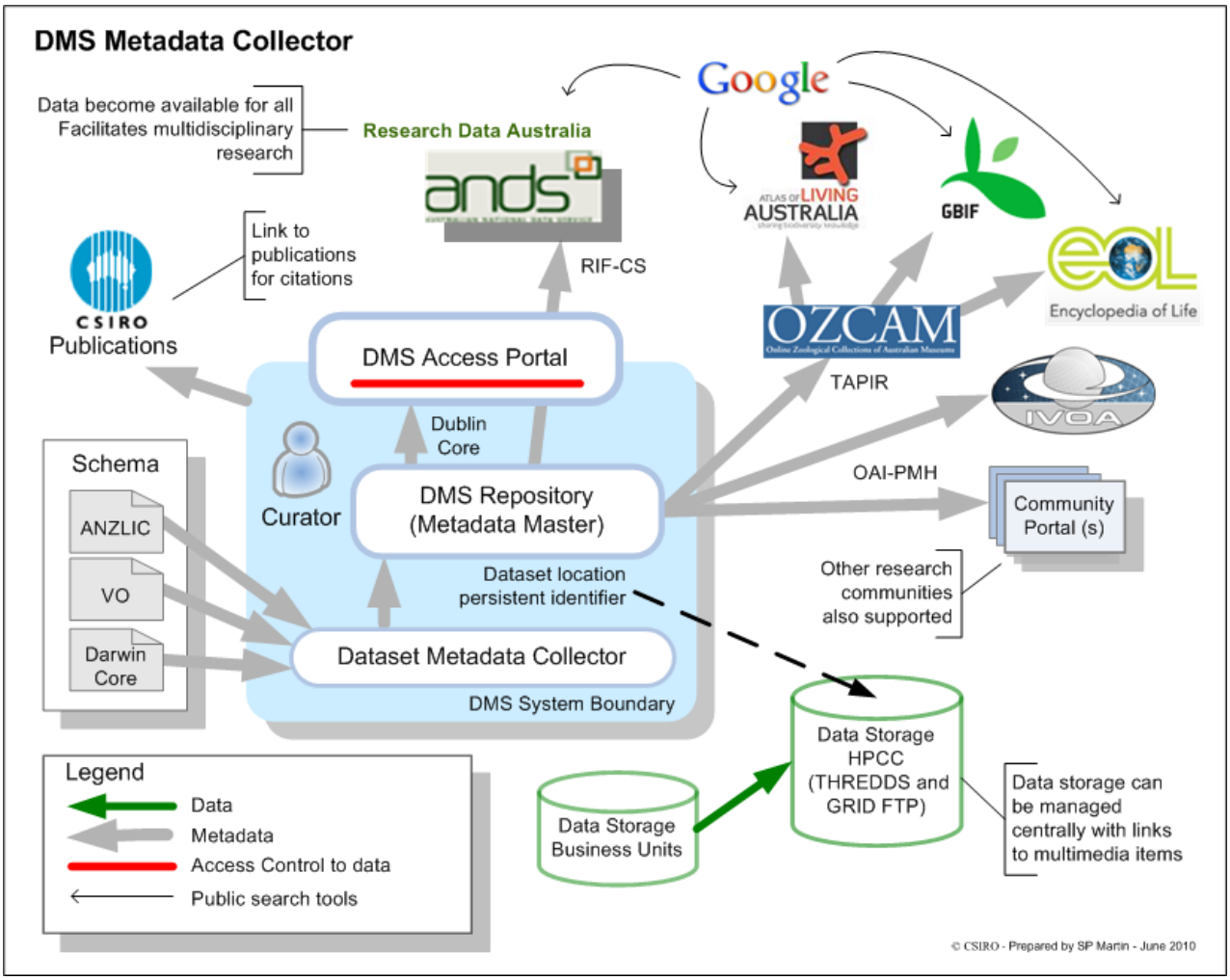

Figure 1: Data Management Service (DMS) Metadata Collector

CONCLUSION

The breadth of research in CSIRO presents both challenges and opportunities in research data management. This presentation will outline the experiences of the CSIRO IM&T team in applying new services and technologies to address these challenges. One of the key lessons from the experiences so far is that effective management of research data is all about partnerships – within CSIRO and external to the organization. It is the magical combination of these partnerships that makes it possible to pull the right rabbit out of a hat full to the brim with research data.

ACKNOWLEDGEMENTS

Members of the CSIRO IM&T Data Management Services Team; Members of the CSIRO IM&T Software Services Team; and Susan Martin, CSIRO IM&T, for supplying the diagram used in Figure 1.