med.data.edu.au

COLLECTION-LEVEL METADATA STANDARDS AND METADATA GATHERING PROCESS

Oct 2015

1. Minimum, core set of collection-level metadata

The med.data.edu.au project team have developed a minimum set of core RIF-CS[1] elements that are required for any datasets to be stored on med.data.edu.au. All nodes of the project have agreed upon the content of this minimum.

The minimum collection-level metadata set is an instantiation of RIF-CS, in which the core fields are made mandatory. This ensures that the collection-level metadata collected for med.data.edu.au are conformant with the RIF-CS standard and can be passed directly to/from Research Data Australia (RDA)[2], a national registry of datasets that is maintained by the Australian National Data Service (ANDS)[3].

The core set of mandatory metadata elements is shown in the table below in orange. Fields in yellow denote recommended metadata elements (i.e. if relevant). The minimum has been favourably reviewed by ANDS.

Section	Field	Occurs	Notes
Collection Descriptors	Name	1	A collection must have one name.
	Description	1-n	A collection must have at least one description of type 'full' or 'brief'.
	Technical description	0-n	A description of what sort of software or other tools are required to use the data, and where to download that software. Recommended.
	File formats	0-n	If the collection contains data files, the file types should be listed. Recommended.
Collection Coverage	Subject	1-n	A collection must have at least one subject in ANZSRC Field of Research (FoR) form. Additional subjects of other specified vocabularies or in free-text format are accepted in addition.
	Date Range (temporal coverage)	0-n	Date elements (specifying beginning and end dates) refer to the temporal coverage of the collection, and is recommended where relevant. For example if a collection contains interview data, the temporal coverage element would refer to the date/s of recording.
	Location (spatial coverage)	0-n	Spatial coverage elements is recommended if a collection contains spatial data. For example a public health study in which the subjects' locations/addresses are relevant.
Access conditions to	Access rights	1	One of either open, conditional or restricted.
	License	1	A collection must either have a license, or otherwise

[1] http://www.ands.org.au/resource/rif-cs.html
[2] https://researchdata.ands.org.au/
[3] http://www.ands.org.au/

data in the collection			explicitly state 'no license'.
	Rights Statement	0-1	A free text description of the access restrictions on the data, e.g. 'To access data, you must submit an ethics application to ...' or 'contact the custodian to negotiate access to this collection'. Recommended.
	Data location	1	Must have an electronic location, that points as close to the data on med.data as possible. This can point to a restricted location (i.e. a location that isn't accessible without authentication) if the collection is not open access. This could also be a contact email address of the Data Custodian who should be contacted to negotiate access.
Citation information for the collection	Citation	1	A collection should specify how others should cite it. Recommended.
Parties related to the Collection	Owner	1-n	A collection must have a nominated owner. Can be an institution rather than an individual.
	Collector	0-n	A collection is aggregated by the 'Collector'. A collection should have at least one collector and may have more than one.
	Manager	1-n	The Collection is maintained and made accessible by the Manager (Manager is a RIF-CS element type and includes the role of Data Custodian). A collection may have any number of nominated managers. This can also be an administrative position rather than a party or an individual.
Other Items related to the Collection	Related collections	0-n	Collection:Collection relation = isPartOf or hasPart.
	Related Activities	0-n	Information on the grant that has funded the research, which is related to the collection. Recommended if funded by ARC or NHMRC (where activity records exist in RDA for these grants). Collection:Activity relation = isOutputOf
	Related Publications	0-n	Publications from research that was undertaken using data in the collection. Recommended.
	Related Services	1-n	All data collections stored on a node of med.data.edu.au will link to a published RIF-CS service record for med.data.edu.au in RDA. Collection:Service relation = isAvailableThrough It is envisaged this link will be added in reverse from a published Service record in RDA for med.data.edu.au by the med.data.edu.au team (in which case the Service:Collection relation = makesAvailable).

Obligatory RIF-CS fields for all med.data.edu.au collections are:

- Name
- Description
- Subject
- Access Rights
- License
- Data location

Additionally, the following linked party records are required for each collection

- Owner
- Manager (i.e. Data Custodian)

2. Operational design for metadata assembly and display

Collections stored on the 5 nodes participating in the med.data.edu.au project originate either from institutions that have mature metadata management capabilities (i.e. metadata stores and associated processes) or from institutions that do not.

A visual representation of the architecture that has been implemented to support the creation and display of RIF-CS descriptions on the med.data.edu.au website (for any datasets stored on med.data.edu.au infrastructure), as well as in RDA, for data custodians based at either type of organisation is shown in Figure 1 below.

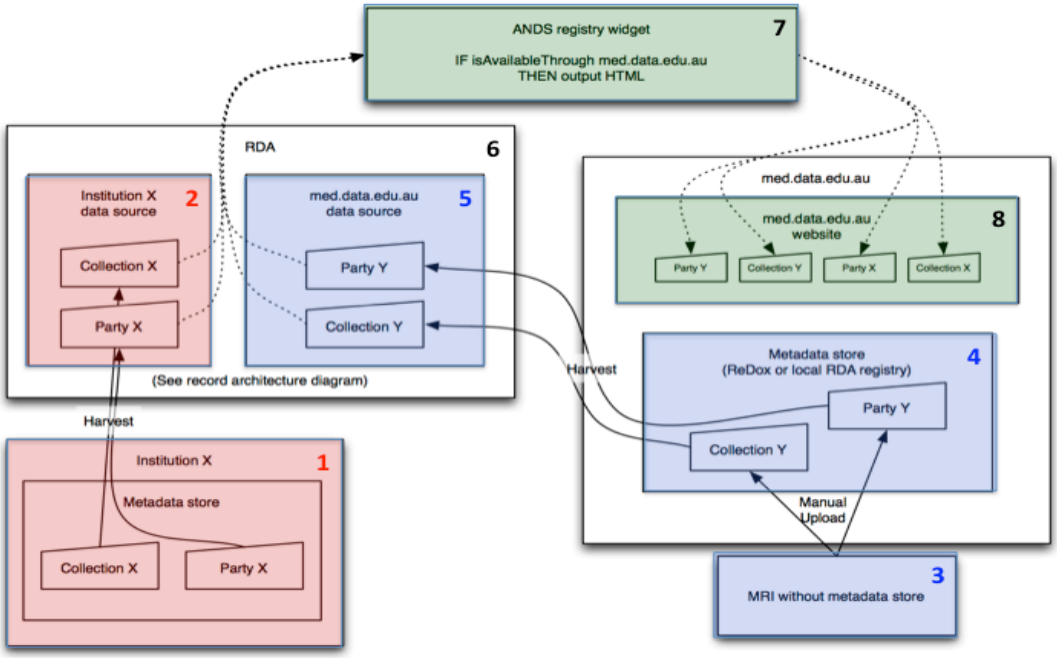

Figure 1. Model of process for creation of collection level metadata for data stored on med.data.edu.au, and display of metadata descriptions on the med.data.edu.au website and RDA.
(1) Data custodians based at institutions that have a metadata store, create RIF-CS collection and party descriptions using their local infrastructure. (2) These RIF-CS descriptions are exposed to and harvested by RDA. (3) Data custodians at institutions that do not have a metadata store (4) create RIF-CS collection and party descriptions using a metadata collection tool provided by med.data.edu.au. (5) RIF-CS collection and party descriptions created using the metadata repository provided by med.data.edu.au are exposed to and harvested by RDA. (6) Collection descriptions in RDA are exposed through (7) the ANDS registry widget[4], which is used to (8) display RDA content directly on the med.data.edu.au website (see http://med.data.edu.au/find-data/) for datasets stored on med.data.edu.au infrastructure.

Associating RIF-CS collection descriptions in RDA with the med.data.edu.au service

[4] http://developers.ands.org.au/widgets/registry_widget/

A RIF-CS service record (https://researchdata.ands.org.au/meddataeduau/632288) has been created for med.data.edu.au by project staff at Intersect. Collection descriptions within RDA (for data stored on med.data.edu.au) is associated with the med.data.edu.au project via the inclusion of an 'isAvailableThrough' link between the med.data.edu.au service record and each collection record as illustrated in Figure 2 below. These links are managed by med.data.edu.au staff. Alternatively, Data Custodians can add the 'makesAvailable' link from each collection to the med.data.edu.au service record. The RIF-CS location element in each collection record is used to denote that data is stored on med.data.edu (and at which node).

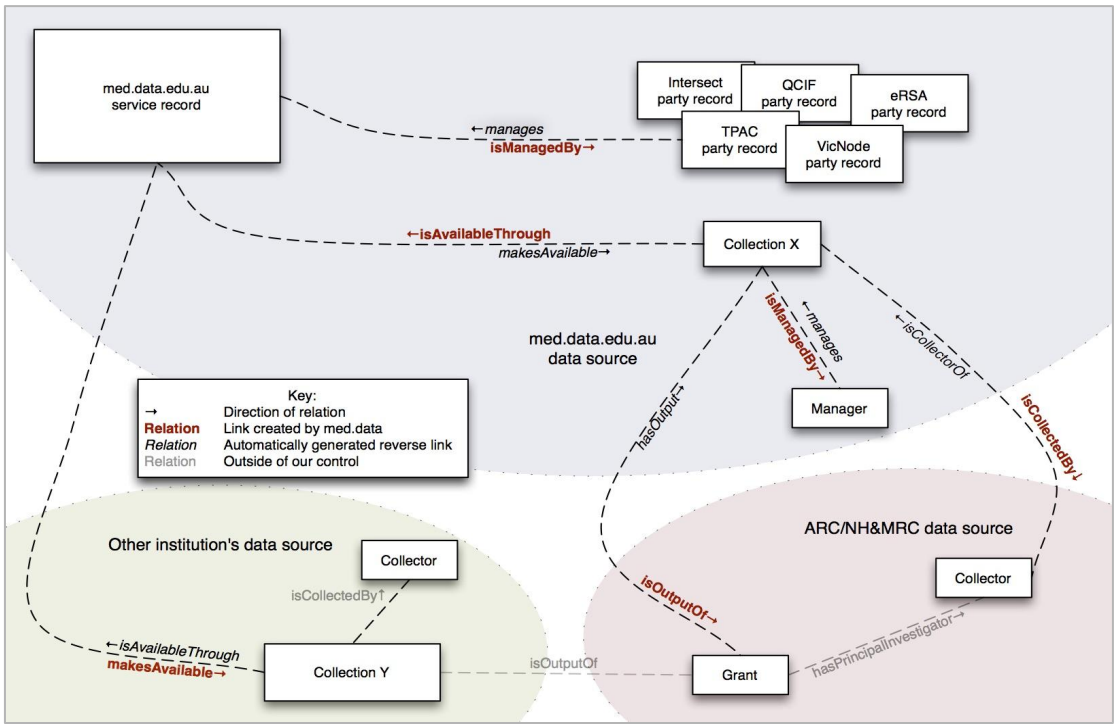

Figure 2. Model of relationship between metadata records in RDA for med.data.edu.au collections, parties and services.

This system design means that the med.data.edu.au infrastructure, consisting of a website, a metadata repository and a presence within RDA, will work effectively alongside any existing capabilities from within institutions, and will mean that universities will not have to modify existing workflows in order to accommodate the med.data.edu.au project.