Addition of “Provenance” to Related Information Type Vocabulary (CC-1179)

Suggested schema change

To enable encoding of additional provenance information that is located outside the ANDS Collections Registry and is described using a URI, this proposal recommends to add a new type “provenance” to the relatedInfo element type.

"provenance": any information about how the data was derived or produced; including what events (such as creation, assemble, annotate, and transform etc.) happened to data, which instruments or software were used, who were involved, when and where an event happened.

Problem this suggestion addresses

RIF-CS v1.6.0 has some capacity to encode provenance information for a collection object, for example, description:type=”lineage”, date:type=”dc. created”, as well as related service and party objects. However, data providers may have fine-grained provenance information that can’t be transformed into a RIF-CS record, or may not own thus not be able to describe all resources in a provenance chain. Some data providers may even take a further step of setting-up a provenance access and query service[1] and generating a provenance report on the fly. The proposed addition of type= “provenance” to the relatedInfo suggested vocabulary is intended to meet these requirements.

Identified by

ANDS staff

RIF-CS schema components affected

relatedInfo Type vocabulary

An entry in a Collection record may appear as:

Use case 1: data provider makes available a provenance access and query service or provenance report generate service (e.g Galaxy IReport), which will generate a whole trail of provenance information around a data collection:

```
<relatedInfo type="provenance">
        <title>CSIRO Provenance Access and Query Service</title>
        <identifier type="uri">http://data.csiro.au/service/provenance/function/search?target=http://researchdata.ands.org.au/object/1234<
/identifier>   #a current RIF-CS record with a PID 1234

<note>This provenance query results in a trail of provenance information of the collection.</note>
</relatedInfo>
```

Note: This element is repeatable, if a RIF-CS record is a collection that has more than one data item, while data provider’s provenance is data item based.

Use case 2: data provider has captured and saved provenance information into a text file, e.g. an un-processed log file or a report generated from R code:

```
<relatedInfo type="provenance">
        <title>Steps taken to produce this collection  </title>
        <identifier type="uri">https://rawgit.com/yihui/knitr-examples/master/003-minimal.html</identifier>  #knitr generated report from R markdown
</relatedInfo>
```

Optional:
Use case 3: data provider has captured provenance information in a metadata record external to Research Data Australia:

```
<relatedInfo type="provenance">
        <title>More provenance information in this  landing page/metadata record</title>
        <identifier type="uri">http://www.ga.gov.au/metadata-gateway/metadata/record/83161/</identifier>

<note>The URL points to a landing page with more provenance information.</note>
</relatedInfo>
```

Data providers are highly recommended to follow an approach as described in use case 1 and 2. Use case 3 describes a scenario that sets low entry point for data providers who have provenance information but haven’t implemented policy, procedure and system to manage provenance information. In some cases, the URI in the use case 3 may be the same as in rifcs:location when location URI points to a landing page (source metadata managed by data providers), however it still worths to point out in the relatedInfo:type=provenance that the target page has provenance information not included in the RIF-CS record in RDA.

Impact on content providers

This addition provides more options to data providers; there will be no impact on those organisations that choose not to use it. For those who do wish to use the option, they will need to decide: does the page a provenance URI points to have provenance information additional to that already presented in the current RIF-CS record? If the answer is yes, they will then need to implement the option in their own systems, or accommodate it in their harvest of records to RDA. It is the responsibility of data providers to decide what and how to present provenance information on the linked page.

Pros

More and more ANDS providers are considering data provenance issues, especially those involved in data-intensive computing such as virtual labs. Some of them have taken the step of capturing and storing provenance information in log files or in a database, and describing provenance information from free-text to in PROV-O[2] (provenance ontology) which may or may not be encoded directly in a metadata record. The proposed "provenance" option will offer these data providers the option to link a RIF-CS record to richer provenance information than that provided in a RIF-CS record.

Cons

If a data provider chooses to adopt this vocabulary option, the provider will need to implement it in their own system, if capturing metadata in native RIF-CS, and/or make a change to their harvest to RIF-CS.

Technical options

- The new relatedInfo type "provenance" will need to be added to the ORCA registry database table.
- The vocabs.xml file will need to be amended to add the new relatedInfo type and definition.
- The vocabularies.html will need to be regenerated to reflect the addition of the new relatedInfo type.
- Changes will be required to the Content Providers Guide and Metadata Content Requirement document.

Note:

RIF-CS description types include an attribute "lineage". A clear guidance should give to data providers on when description:type=lineage and relatedInfo: type=provenance should be used. The description:type=lineage is used for collection records to articulate lineage information in RIF-CS records for those repositories capturing such detail in their native schema. For example, ISO19115-2 lineage statement element can be naturally mapped to RIF-CS description:type=lineage. While relatedInfo:type=provenance is mainly to show what other resources are involved in creating or use a collection and how these resources are related to the collection.

---

[1] http://www.w3.org/TR/2013/NOTE-prov-aq-20130430/#resource-represented-as-rdf
[2] http://www.w3.org/TR/prov-o/