ISO/IEC JTC 1/SC 32/WG 2

Convenorship: ANSI (United States)

Document type: Contributions
Title: WG2 N2071 Expert Contribution on WG2 N2063 Metadata for Datasets
Status: Document for discussion of metadata for datasets
Date of document: 2015-05-18
Source: Ray Gates
Expected action: INFO

Email of secretary:
Committee URL: http://isotc.iso.org/livelink/livelink/open/jtc1sc32wg2
ISO/IEC JTC 1/SC 32/WG 2

Metadata

<table>
<thead>
<tr>
<th>DOCUMENT TYPE</th>
<th>Expert contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE</td>
<td>Input to discussion of Metadata for Datasets</td>
</tr>
<tr>
<td>SOURCE</td>
<td>Ray Gates, Canadian Expert</td>
</tr>
<tr>
<td>PROJECT NUMBER</td>
<td></td>
</tr>
<tr>
<td>STATUS</td>
<td>For discussion at WG 2 meeting in Glasgow, UK, May 18-22, 2015.</td>
</tr>
<tr>
<td>REFERENCES</td>
<td></td>
</tr>
<tr>
<td>ACTION ID.</td>
<td></td>
</tr>
<tr>
<td>REQUESTED ACTION</td>
<td>Comment</td>
</tr>
<tr>
<td>DUE DATE</td>
<td></td>
</tr>
<tr>
<td>Number of Pages</td>
<td>3</td>
</tr>
<tr>
<td>LANGUAGE USED</td>
<td>English</td>
</tr>
<tr>
<td>DISTRIBUTION</td>
<td>WG2 Convener and P &amp; L Members</td>
</tr>
</tbody>
</table>
Input to the discussion of Metadata for Datasets

Placement within the WG2 families of standards
Since this standard is about registering metadata, it would seem to belong in the 11179 family of standards, presumably as a new part.

Relationship to 11179-3
If we agree to include this standard in the 11179 family (and probably even if we don't), it would seem natural to take advantage of the common facilities provided by 11179-3 for Identification, Designation, Definition and Registration. This document should make some statements about the applicability to each of the objects described in Figure 1.

Scope of this standard
Clause 1.1 Scope Inclusions limits the scope of this standard to:
- data sets published or curated by a single agent
Why this limitation?

Describing a family of related datasets
The current WD seems to conflate metadata about a particular dataset instance, and metadata that could apply to a family of related datasets that may be considered to be of a single type.

For example, accrual_periodicity only makes sense in the context of the family of datasets, while issued_date applies to a particular dataset. Some keywords would likely apply to the dataset family as well.

WG2 should consider whether it makes sense to explicitly show the dataset family in Figure 1.

In addition, the information model associated with a family of datasets may change over time. Do we need to say anything about versioning beyond the little already specified for Administered Items in 11179-3.

Role of the Data Catalog
Clause 5.4.3 defines data catalogue as: a curated collection of metadata about one or more data sets.
How does a data catalogue relate to a metadata registry?
What are the pros and cons of treating these concepts separately versus combining them?
Should we support a standalone catalogue for organizations that do not want a full-blown metadata registry?
Should we support data catalogue functionality within an MDR for those organizations which already have one?
The following are some possible uses of a data catalogue / dataset registry:
1. An organization that publishes datasets maintains the catalogue/registry so that potential consumers can find them.
2. An organization that consumes datasets from multiple sources, maintains a catalogue/registry of all the datasets they have acquired, so users can find them.
3. An organization maintaining a 'data lake' (e.g. base on Hadoop) requires that all datasets be registered as they are ingested, as part of the data governance process over the data lake.

If we accept that a dataset may need to be catalogued/registered by both publishers and consumers, then we need to relax the constraint shown in Figure 1 that a dataset be described by at most one data catalog. Even if the data catalog were to be reserved for publishers only, a
publicly available dataset could potentially be listed in more than one catalog, or a corporate conglomerate could list a proprietary dataset in more than one catalog.

**Data Formats**

A note to the definition of 'data set' states:

> NOTE A data set is usually presented in a tabular form and stored and distributed in tables in word processed documents, spread sheets or databases.

In this world of big data and IoT we need to support many more formats than those listed above, including: AVRO, JSON, RDF and XML.

**Keywords**

It may be beneficial to separate out certain categories of keyword, such as: Subject Domain, Supported Standard, or at least to provide a way for keywords to be categorized.

**Quality Assessment**

The Data_Set_Quality_Assessment_Type includes a category for "relevance" to users' needs. There needs to be some associations with the corresponding user classes to specify which users the relevance applies to, since relevance may vary by user. Other aspects of quality may vary by user as well, since data suitable for on purpose may not be suitable for another.

**Completeness of metadata**

11179-6 specifies registration statuses that apply to the lifecycle of the metadata. 11179-3 also allows administrative statuses defined by the registration authority. Presumably these can be applied to

When using a data set registry to register the ingestion of data sets into a data lake, we may need a third way to capture statuses. For example:

1. Upon receipt of the dataset, we may just need to know: identifier, title, description, issue date and origin.
2. Before we use the data for a specific purpose, we will probably want more information, including: information model and quality assessment.