Fig. 1

1. Gatewaying Metadata

Markus Demleitner
msdemlei@ari.uni-heidelberg.de

A metadata gateway does two things:

- Hide discipline metadata harvesting details (harvest from where? using which protocols?)
- Translate between different metadata schemas

This talk: How I'm running a gateway from the Virtual Observatory (VO) to B2FIND.

(cf. Fig. 1)

2. The Virtual Observatory

On one side of the gateway is the VO:

- ∼ 50 data centres
- ∼ 30′000 data collections/metadata records
- Largely service-based: Through standard APIs, you get access to a few 10^8 data sets (spectra, images...) and a few 10^{10} table rows
- Within the VO, most discovery is rather directly for API endpoints...
- ...and uses "full searchable registries"

An example for "API endpoint": There is a protocol for searching for spectra, SSAP. So, when people look for, say, infrared spectra, they will go to a registry and say: "Give me the URLs of SSAP services saying they have infrared data".

Since all these services and be queried uniformly (that's the "API" part), a client can then go to each URL in turn and ask a thing like "do you have spectra at position such-and-such?". Since all services reply uniformly, the client can then present a set of all matching datasets in the VO.

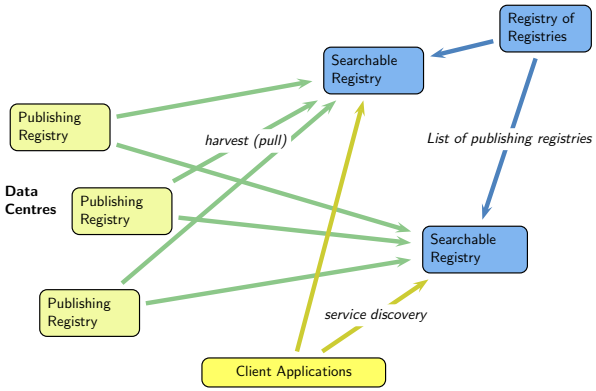

Fig. 2

3. Full Searchable Registries?

(cf. Fig. 2)

This is how metadata is managed in the VO: Data Centres put up their bibliographic records for harvesting in publishing registries. These are harvested using OAI-PMH by searchable registries, which know where to go because there's a registry of the searchable registries (which is the only central component here). When a client looks for services in the way just discussed, it queries a searchable registry using a VO-specific protocol names RegTAP.

The searchable registries that claim to have everything in the VO (there are three of those globally right now) are called "full searchable registries".

4. The VO's Schema

VO Registries communicate using OAI-PMH, and hence they know about oai_dc. But that really isn't enough for meaningful handling of digital artefacts.

The VO's native metadata format is ivo_vor, roughly comparable to oai_datacite plus:

- Per-protocol API endpoint definitions ("capability")
- Relational metadata ("tableset")
- Coverage in space, time, and spectrum
- (and a few minor additions)

Fig. 3

5. B2FIND

B2FIND is a cross-disciplinary research data search engine. Metadata schema: Roughly, DataCite plus extensions.

That is, there's no notion of "spectral service" or "table access service".

Nor should there: Non-astronomy users have no clients that would care for those.

Anyway: We will have to translate when going from the VO to B2FIND.

(cf. Fig. 3)

6. VO in B2FIND: Fanning out?

Since B2FIND speaks OAI-PMH, it could harvest the publishing registries.

However:

- B2FIND would have to translate the metadata (which requires some domain expertise that B2FIND probably does not want to have for potentially dozens of domains)...
- ...or all publishing registries would need to know about B2FIND,
- B2FIND would have to worry about failing publishing registries,
- and it would have to know about the IVOA Registry of Registries.

7. VO in B2FIND: Use a Gateway

Instead: have one particular searchable registry figure out the translation.

- B2FIND isolated from domain details
- B2FIND isolated from domain operations
- Consistent translation for the entire domain.

(cf. Fig. 4)

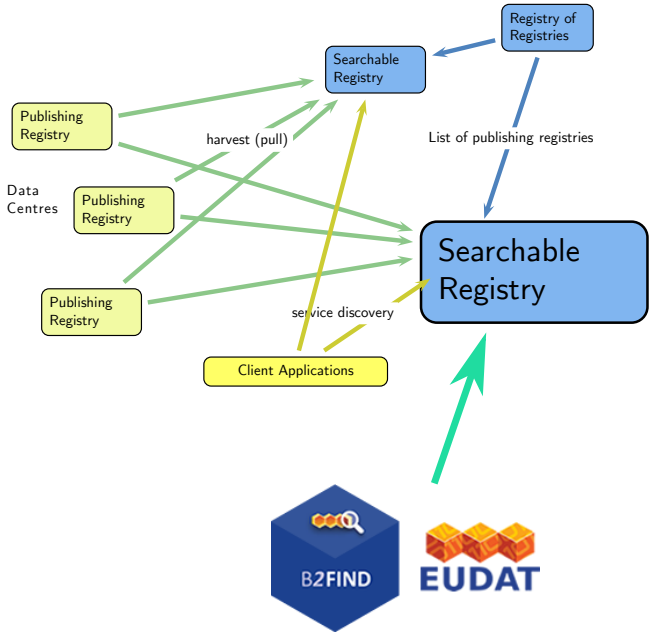

Fig. 4

8. Examples for Schema Mapping

The actual rules are encoded in a piece of XSLT[1]; here are just a few examples to give you an idea what sort of thing one has to deal with when bridging domains in data publishing.

- content/description → descriptions/description
- altIdentifier → identifier *if it's a DOI*
- content/referenceURL → relatedIdentifier of type URL
- curation/contact/name and email → *formatted* contact
- content/subject → keyword *plus mapping to top-level terms*
- coverage/temporal → temporalCoverages/startDate and endDate, *converting from MJD to ISO* – the spectral and spatial coverages don't map because of missing concepts or reference frames.
- capability/interface → relatedIdentifier of type URL *if it's a browser interface* (as opposed to something that can only be used with a protocol client)

You can probably see why it is a good idea to concentrate conventions like those in one place so they're at least the same for all records within the domain and can be quickly changed if found lacking.

1 `http://svn.ari.uni-heidelberg.de/svn/gavo/hdinputs/rr/res`

9. Rotten Metadata

Rule #1: Unused metadata is quite likely broken.

Example: In the current VO, few people look at the referenceURL (this is where people will find human-readable documentation, and few humans read documentation).

For... reasons (i.e., hack alert!), we were using referenceURL as the identifier.

Of course: 22 of 4975 test records ended up without a proper identifier because nobody bothered to check the referenceURLs (beyond XSD xs:anyURI).

At least most data providers fixed their records when we manually followed up. Which is another reason for why the gateway should sit within the domain: I knew most of the people involved and they knew me, which I claim helped a lot when I asked them to touch their records.

10. Conclusions

- When linking metadata collections, designate a gateway *within* the discipline: The translation rules and curation work are then concentrated at one place with more domain knowledge than a generic repository could hope to have.
- The less a piece of metadata (or, really, anything a standard requires) is actually used, the more validation you will need on it to keep it usable.

Thanks!