What We Find on the Other Side of the Metadata Rabbit Hole

Soren Scott

The value of these communities is to make a certain kind of space, one to step back and have a think about what we’re doing and why. So I’ll ask: is our approach to this documentation meeting our needs? The researchers? The educators?

I am a developer. This is not a technical issue. It is a cultural one. We have to ask, still, if the technical solutions we build support the culture of open science or do they hinder it? At its most basic, if we don’t value this data, how can we expect anyone else to?

Metadata is a public good. It is an intentional act for another’s benefit and, as soon as it’s on a public-facing system, those external use cases are unavoidable. Just the simplest use case — a researcher downloading the file — requires an object containing enough information for understandability offline and enough to support easy re-discovery.

Metadata is a public good. It is an intentional act for another’s benefit and, as soon as it’s on a public-facing system, those external use cases are unavoidable. Just the simplest use case — a researcher downloading the file — requires an object containing enough information for understandability offline and enough to support easy re-discovery.

It is the pot calling the kettle black in that we all know metadata is pretty bad. But when you look at the ways in which it is bad, there are commonalities that speak to several possible paths forward.

We start asking the smaller questions as we work towards answering the much larger one — how do we get to a Google for Data? Or at least good metadata?

Hierarchy of Metadata Use Cases

- Input Data for Modeling/Mediation
- Provenance
- Service Generation
- Data for Discovery (Internal & External Systems)
- Atomic File, on disk

Decreasing Visible Value

1. Consider targeted documentation for desired community recommended practices and design patterns for implementing ISO-19115 and other standards.
2. Make republication recommended practices explicit, i.e., define practices for identifier handling, link rot concerns, and basic document integrity.
3. Let’s talk to the data publication framework developers to update those systems to support well-described service descriptions.
4. Continue the discussions around standards meeting the community needs and work to improve existing or revisit other solutions.

Some paths forward:

- Consider targeted documentation for desired community recommended practices and design patterns for implementing ISO-19115 and other standards.
- Make republication recommended practices explicit, i.e., define practices for identifier handling, link rot concerns, and basic document integrity.
- Let’s talk to the data publication framework developers to update those systems to support well-described service descriptions.
- Continue the discussions around standards meeting the community needs and work to improve existing or revisit other solutions.

Simply taking advantage of the structures that we commonly use more fully would address so many of the issues we face building client systems. Finally, we don’t need more geeks with solutions, we need geeks with empathy, willing to say “I don’t know. Let’s work on this together.” It is time to consider new approaches. Oh, and validators.

What does that say about these documents?

If one can extract an identifier, it is unlikely to be identified as a reference to an object, dataset or metadata.

It is a peculiar game — we toss our well-designed identifiers back into little haystacks (the identifiers) back into little haystacks.

And the namespaceing…

Is it valid? Can it be validated?

EO metadata and service descriptions validate at lower rates than other XML.

If you look at services related to EO data more directly, those are almost always valid. Metadata as second class objects?

1992 was a great year for metadata.

Token counts are really pretty good for 2040.

The rest of us struggle with metadata creation dates.

We struggle more with maintaining distribution links.

These degrade quickly, if they are present at all.

The catalog of issues in metadata and service descriptions: On the Need for Self-describing Data and Services on the Web. Scott, S., R. Duerr, S.J.S. Khalsa. [In progress]. 2016. (Ask me or Ruth Duerr for info) Any opinions, and all of the speculation, are entirely my own (and possibly a very confused marmot but that’s for another day). Creative Commons Attribution 3.0 United States. Winter ESIP 2016