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

Soren Scott ､., ◠ ､., ◠ ､., ◠ ､., ◠､., ◠ ､., ◠ ､., ◠ ､., ◠ ､., ◠ ￣(,,◉ ∧ ◉,,)￣◠ ､., ◠ ､., ◠ ､., ◠ ､., ◠ ､., ◠ ､., ◠ ､., ◠ ､., ◠ ､., ◠ ､., ◠ ､., ◠ ､., ◠ ､., ◠ ､., ◠ ... ､., ◠

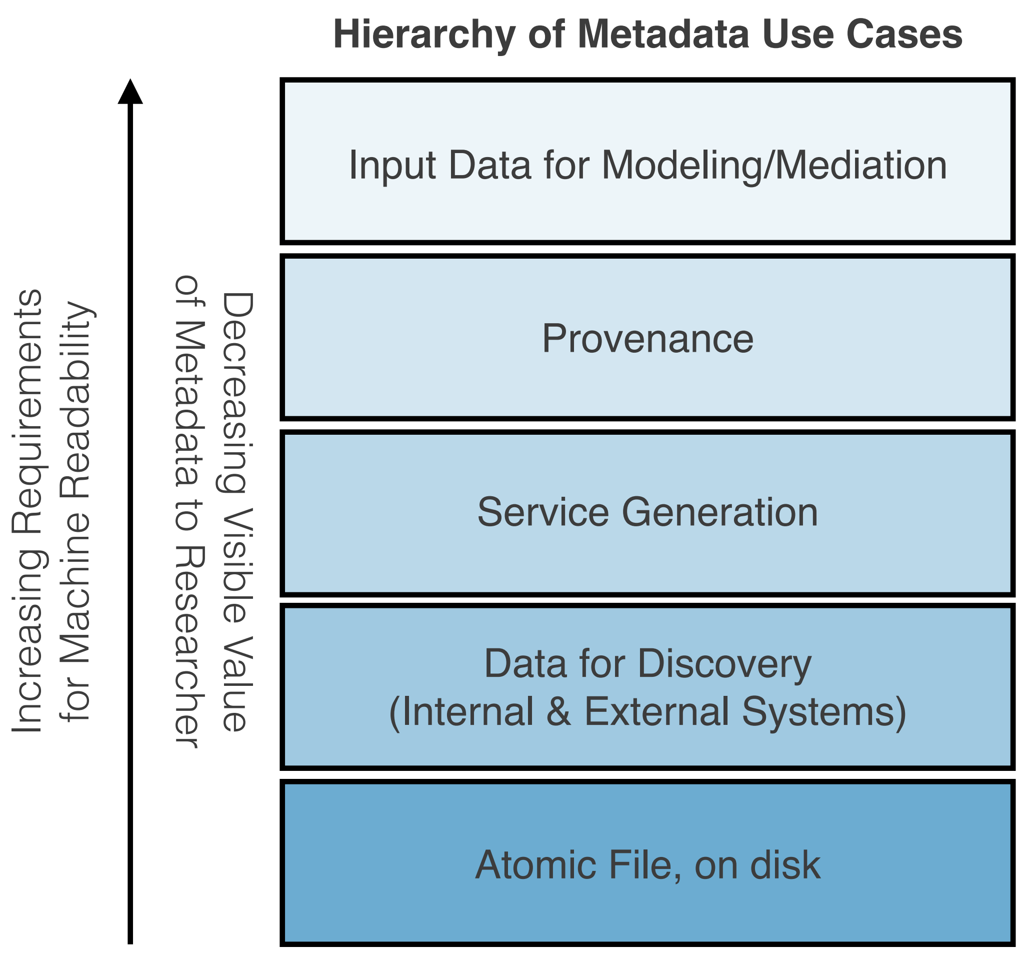
The value of these communities is to make a certain kind of space, one to step back and have a think about what we're doing and why. So I'll ask: is our approach to this documentation meeting our needs? The researchers? The educators?

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.

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
Increasing Requirements for Machine Readability
Decreasing Visible Value of Metadata to Researcher

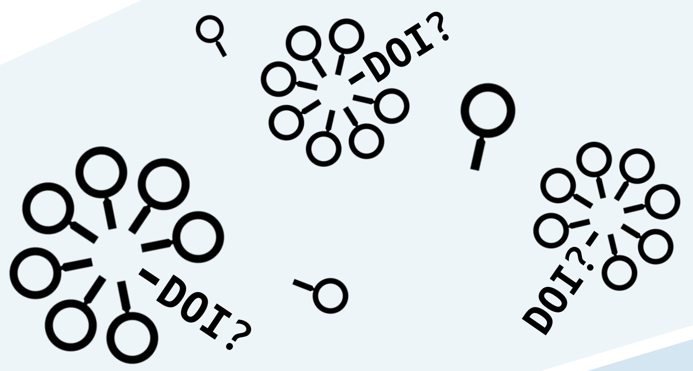
Some paths forward:

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.

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.

We find a handful of Identifier elements that can be linked to a dataset and that are characterized as DOIs.

Similarly, for URL characterization, the use of well-defined strings such as those defined by the Cat-Interop project is limited to services based on a particular open source geospatial data server.



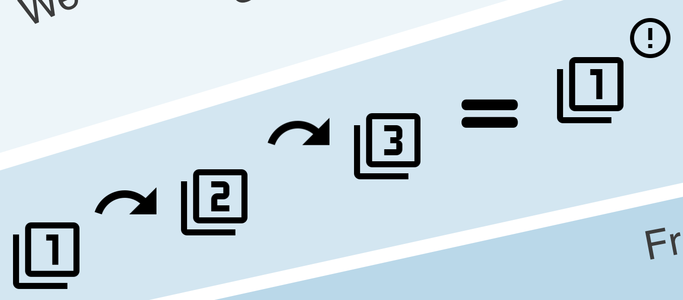
<(_ _)>

Anecdotally, we see a trend towards a single lineage statement, moving away from the more explicit set of processing steps.

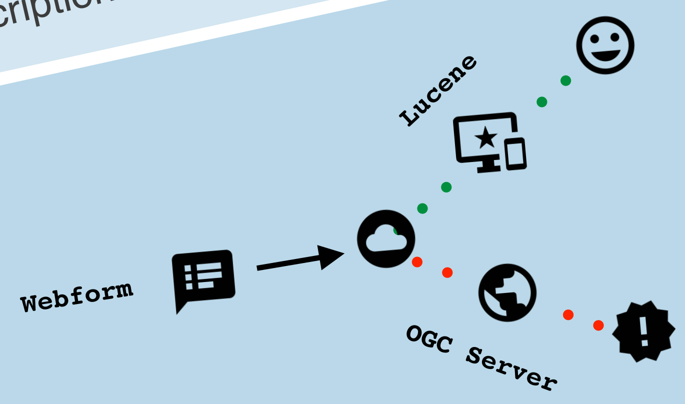
So, as work continues to align the ISO lineage with PROV, we move away from those lineage structures.

Regardless of structure, the ratio of tokens/element is slowly dropping — the descriptions are getting shorter.

(ﾉ_ _)ﾉ



From a data publication system based on the common Lucene-based search front-end and an OGC server for data access, we find that, often, the metadata collected for the discovery is not transferred to the data access service descriptions.



(⊙_⊙)

1992 was a great year for metadata.

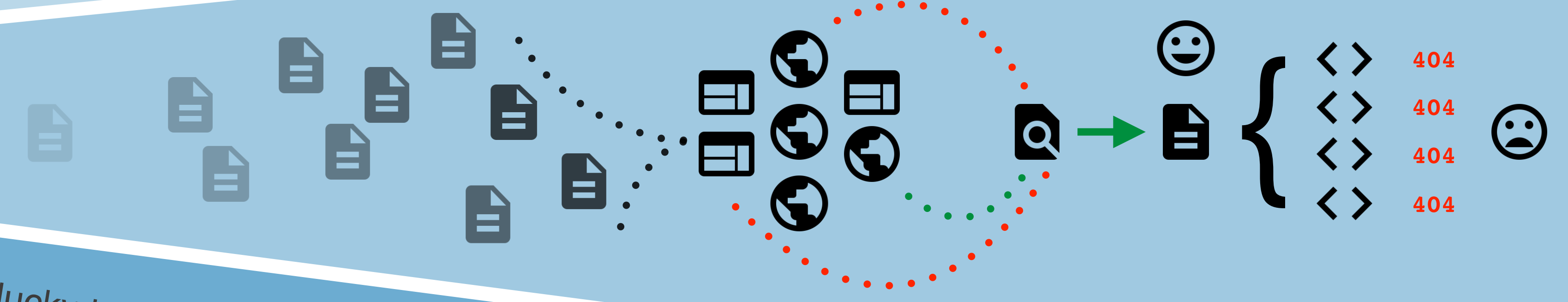
Token counts are really pretty good for 2040.

(¬_¬) So good job, future selves!

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.

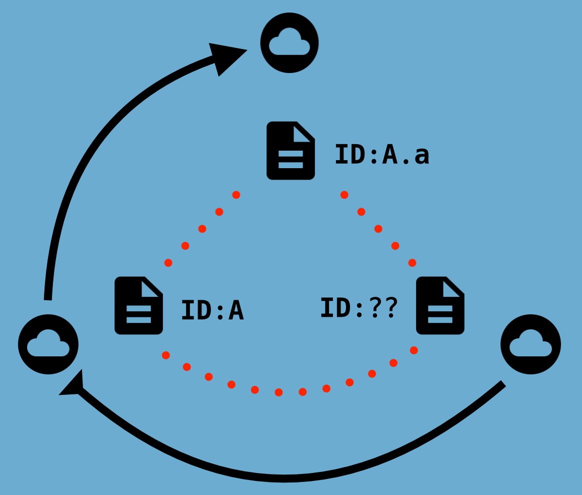


We're lucky to find a recognizable identifier beyond the URL the document came from.

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 needles (the identifiers) back into little haystacks (CURIEs or some other representation) and our algorithms confuse the haystacks for the needles.

And the namespacing...



Is it valid? Can it be validated?

What does that say about these documents?

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?

(･_|

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

￣(,,◉ ∧ ◉,,)￣