Wednesday, October 19 • 2:15pm - 3:30pm

Bluejeans: https://bluejeans.com/839962867

Slides: http://www.slideshare.net/azaroth42/iiif-discovery-of-resources

Two main aspects to discovery in IIIF: how to find resources? Then how to get them into an environment to use them (drag and drop, opening viewer, etc.) - both are equally important

Success = user finds resources. Clear spec/best practices for how to build thing to make discovery work

Scope - content resources (Image API) (images, A/V later) and -description resources (manifests, annotations for example) (Presentation API - although it's only for humans to read, not machines) = we can rule out images bc we need the descriptive and annotations bc we have Search API. So the real scope is - need to discover manifests and collections of manifests

Constraints =
- Presi API doesn't include semantic descriptions, can be referenced with seeAlso
- Resource centric view of the web (sitemaps), not a service oriented view (sparql, or OAI PMH)

Solution: Content provider has client that submit link to central registry, should be some API to get it in. Should be a browse API for looking for links. Crawler populate search index. User uses search index. Get content back. Woo hoo

Europeana - when provider gives them iiif links - load them into...what? There's now a place in EDM to put links to manifests. Europeana already has a pipeline - they can provide access to resources when providers tell them they exist. Providers have to update Europeana.

OAI PMH has validator gatekeeping function - can't load a manifest without it validating…

Sitemaps - provide link to top level collection or sitemap, validate, crawl. Might be a way to say "just so you nkow, here's how to get to my collection."

Central registry - of what?
- Top level collections?
  - Could be sitemap type thing. Is this not enough?
- All collections?
- All manifests?
  - Is this overkill?

And how to do this? Need to distinguish audiences.

Metaphor - Go to library to find a book - you can search the catalogue. Or you can go to the shelves.

Could have different registries for different content types/domain specific registries? Is that a business that we don't want to get into? If the registry is only top level collection, it's more work...use seeAlso to see what kind of resources you're finding

Sitemaps
- implemented by google and web community (advantage)
- what would google do with a sitemap that has manifests? We don't know...is this a downside of sitemaps compared to collections? Google wouldn't know what to do with collections anyway. At the moment Rob doesn't see an advantage to using sitemaps over collections. Maybe eventually google would recognize that there are tons of manifests and start doing something with them...right now google is looking for pages, not manifests
- resource sync has extensions for adding metadata to sitemaps, we could use that to be more specific within sitemap
- is it better to do collections or sitemaps? Nuno - talking to providers in Europeana - some providers already provide sitemaps, others don't use sitemaps but they can do collections easily. Sitemaps containing info about timestamps to know when something has been updated to know when to harvest?
Simeon - do we want to support both? Presenting manifests in sitemaps
Glen - sitemap for newspapers is public site. But manifests are at different URL. sitemap is something that they have to generate and keep up to date
Tom - red herring around sitemaps. What will Google understand?
Video extension to Sitemaps. Simeon: Google-specific this one.
Google throws namespaces away.
Where to put the complexity?
Question about Collections: Collections are quite curated entities. Multiple different kinds of collections. Rob: Living specs so we can revise them.

Possibility of publishing conflicting information about manifests if both aggregator and home institution both publish Collection with same materials. How to organize agreements.

"By submitting you agree that…"

Need something with timestamps for more efficient crawling.

Only aggregating links to images--not aggregating the actual image resources.

Central Registry

- API or web form. (Is Github a good model? Doubt.)
- How should crawler get to collections? Collection of Collections?

Crawlers would start with Collections (or Sitemaps). We need implementations and what crawling means as far as politeness. How do we provide notifications? Notifications could be an optimization that helps.

Community groups could determine best practices for the kind of metadata so that could help (see below)

The semantic data could be indexed while the Presentation data used to control presentation which eases.

SeeAlso could link to what you're already doing say Schema.org.

We could task the Community Groups with set of profiles important for your domain for how to. Show some examples for each domain.

How can we better promote what's on the other side of a seeAlso reference? See issue https://github.com/IIIF/iiif.io/issues/841 in Discovery on document profiles.

Display is covered so may want a lowest common denominator for how the community publishes metadata. Other comment that this might not be quite enough since (title, date, name).

We can publish more than one profile of metadata.

Slide: "Search Engines: Transfer" and Next Steps?

Where can we get some of our initial information?

Should IIIF-C host the registry?

Discuss!

Wikidata: Could we use it for semantic metadata? Could Wikipedia use images from other institutions?

Question about understanding use case for only allowing some aggregators to have access to certain Collections.

From Chip in slack: is there a reason we are not looking at ipfs as a discovery solution? When there is more to look at we ought to take a look.

Tom Crane examples: http://bit.ly/seeAlso
One example is the Schema.org that already powers SEO.
Could include several other seeAlso links. Just repurpose the data that you already have.

Tom Cramer: Question about difference between Sitemaps and Collections. [Anyone understand the question?]

For larger collections Sitemaps (and ResourceSync) can be of benefit.

What is next? We continue the discussion. Create a technical interest group to advise on this. Still could use better universe than just the several that are there now.

Maybe we need to have a "discovery-a-thon" day or something remote where we focus on this topic for a whole day or something...just a thought