Hydra Metadata WG

HydraConnect 2015

https://wiki.duraspace.org/display/hydra/Hydra+Metadata+Working+Group

Organizing for Action: Subgroups

- 38+ Participants!

- Technical Subgroup - Aaron Coburn & Nick Ruest, leads
- Rights Subgroup - Esmé Cowles, lead
- Applied Linked Data Subgroup - Steven Anderson, lead
- Descriptive Subgroup- Carolyn Hansen, lead
    - MODS subsubgroup - Steven Anderson, lead
- Structural Subgroup - Julie Hardesty, lead

Base Technical Metadata Recommendation

- Core file metadata from EBUCore:
  - Creation date (ebucore:dateCreated)
  - Modification date (ebucore:dateModified)
  - File name (ebucore:filename)
  - File size (ebucore:fileSize)
  - MIME type (ebucore:hasMimeType)

- Checksums from PREMIS
  - premis:hasMessageDigest
  - premishash:md5, premishash:sha256, etc.

Base Technical Metadata Recommendation

- Format identification
  - Pronom format (pronom:puid)
  - Byte order (sweetjpl:byteOrder)
  - Format (dc:format, PCDM File Type Vocab)
    - Image, Video, Dataset, Software, Presentation, Spreadsheet, etc.

- Additional descriptions
  - Label/description (rdfs:label)
  - File use (rdf:type, PCDM File Use Vocab)
    - Original File, Service File, Thumbnail, Extracted Text, Transcript, etc.

Rights Metadata Recommendation

- Rights statement (edm:rights)
  - URI to identify rights status
  - Creative Commons, RightsStatements.org, Europeana, etc.

- Rights holder (dcterms:rightsHolder)
  - URI to identify rights holder
  - LCNAF, VIAF, etc.

License/Embargo

- Rights override (pcdmrts:rightsOverride)
  - Rights status URI or embargo, etc.

- Rights override expiration (pcdmrts:rightsOverrideExpiration)
  - Date in standard format

Rights Decision Documentation

- Rights description (dc:rights)
  - Rights notes, special rights terms, etc.

- Copyright claimant (marcrel:cpc)
  - Former rights owner

- Copyright status (premis:hasCopyrightStatus)
  - PREMIS rights status URI
- Copyright jurisdiction (premis:hasCopyrightJurisdiction)
  - Country/jurisdiction code from ISO 3166 or other vocab

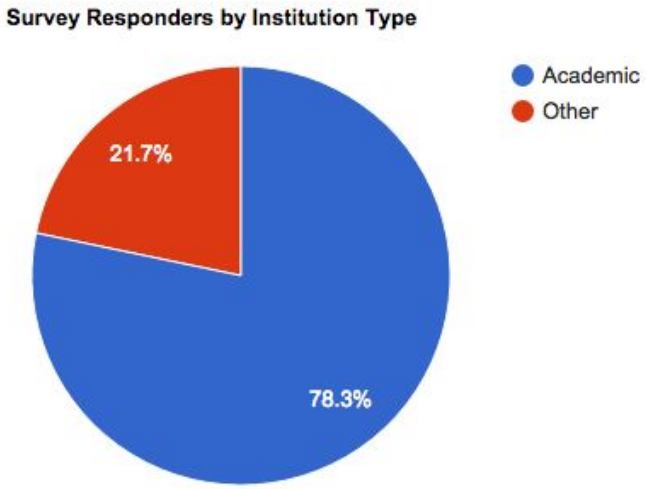
Descriptive Metadata Survey

- Environmental scan of Hydra Community; collecting user stories regarding descriptive metadata

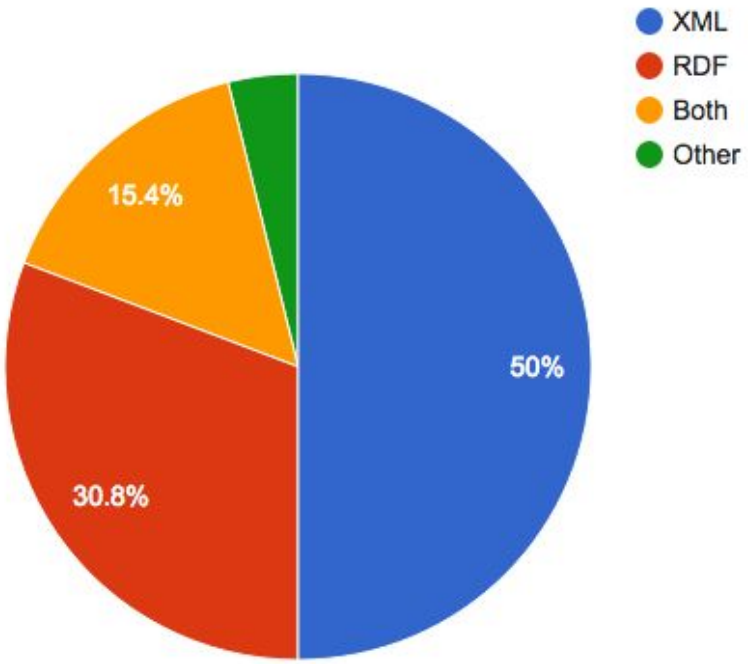
- In order to create...
  1. Best practices for handling blank nodes and nested attributes
  2. Best practices for using multiple schemas/vocabularies
  3. Base descriptive metadata application profile with mapping to Share Data Model, DC, and DPLA

Survey at a Glance

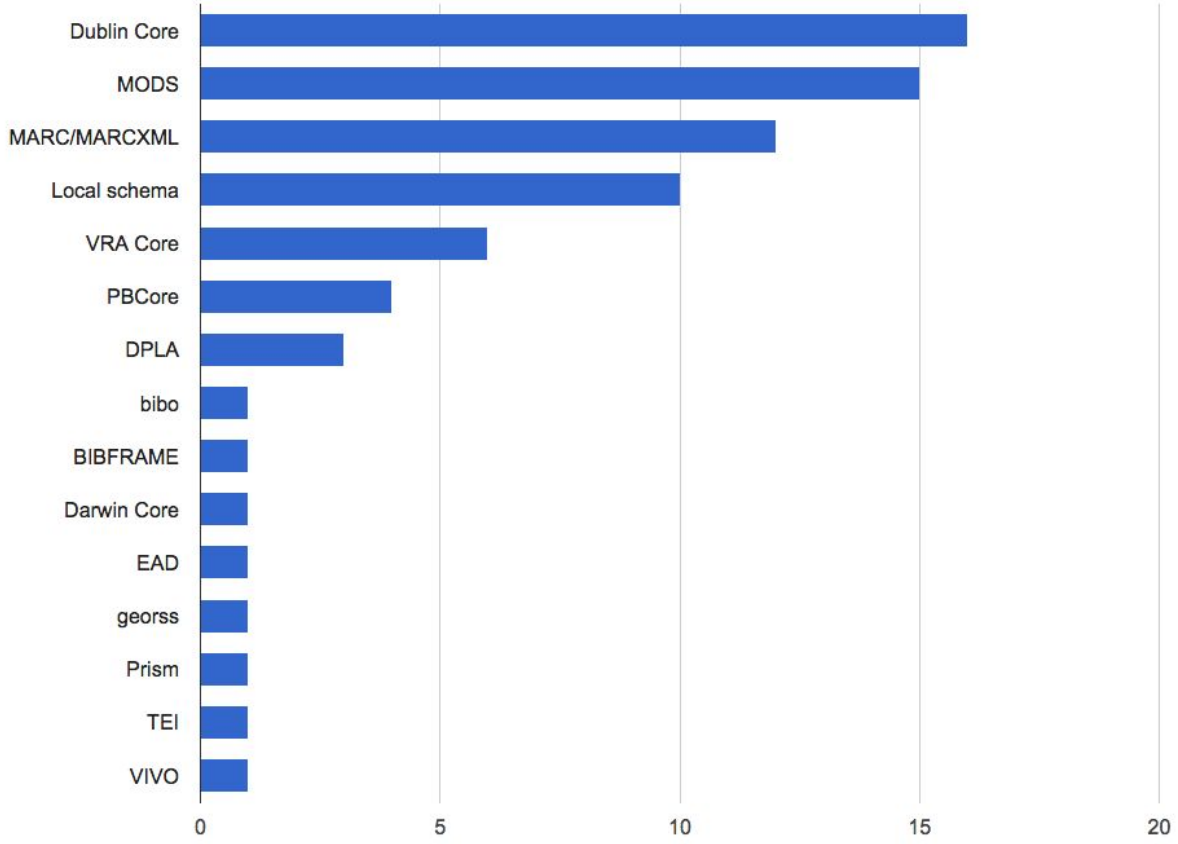
- 25 responses from 23 institutions
- Responders from US, Canada, and Denmark
- Mostly academic institutions



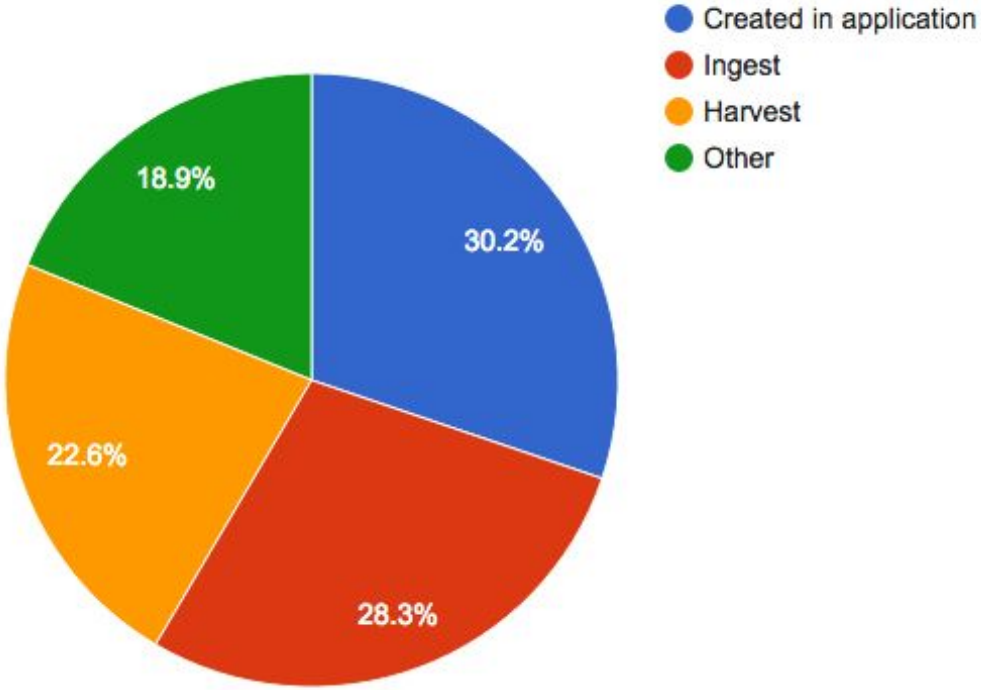
What encoding standards are you using?



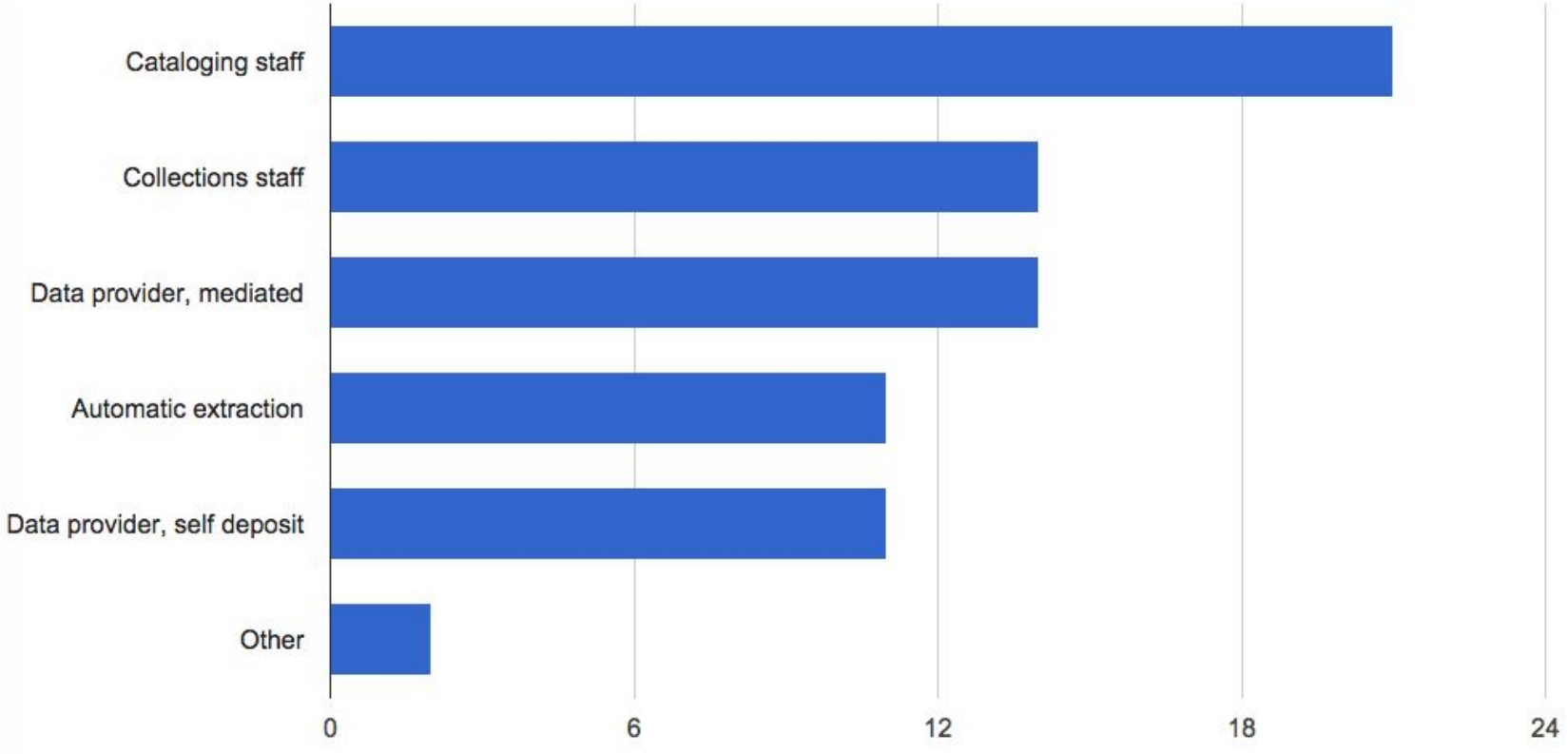
What descriptive metadata schemas are you using?



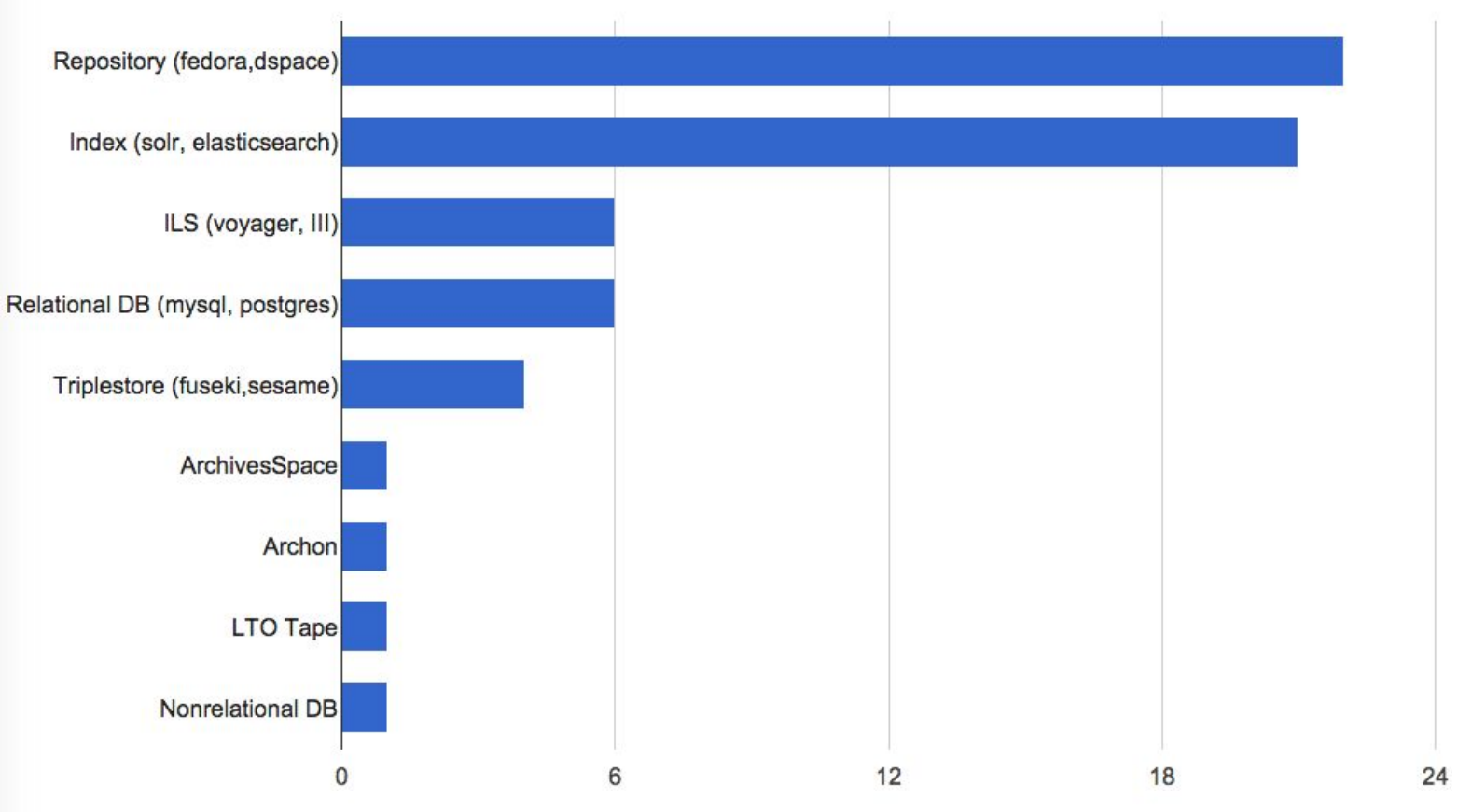
Where is metadata coming from?



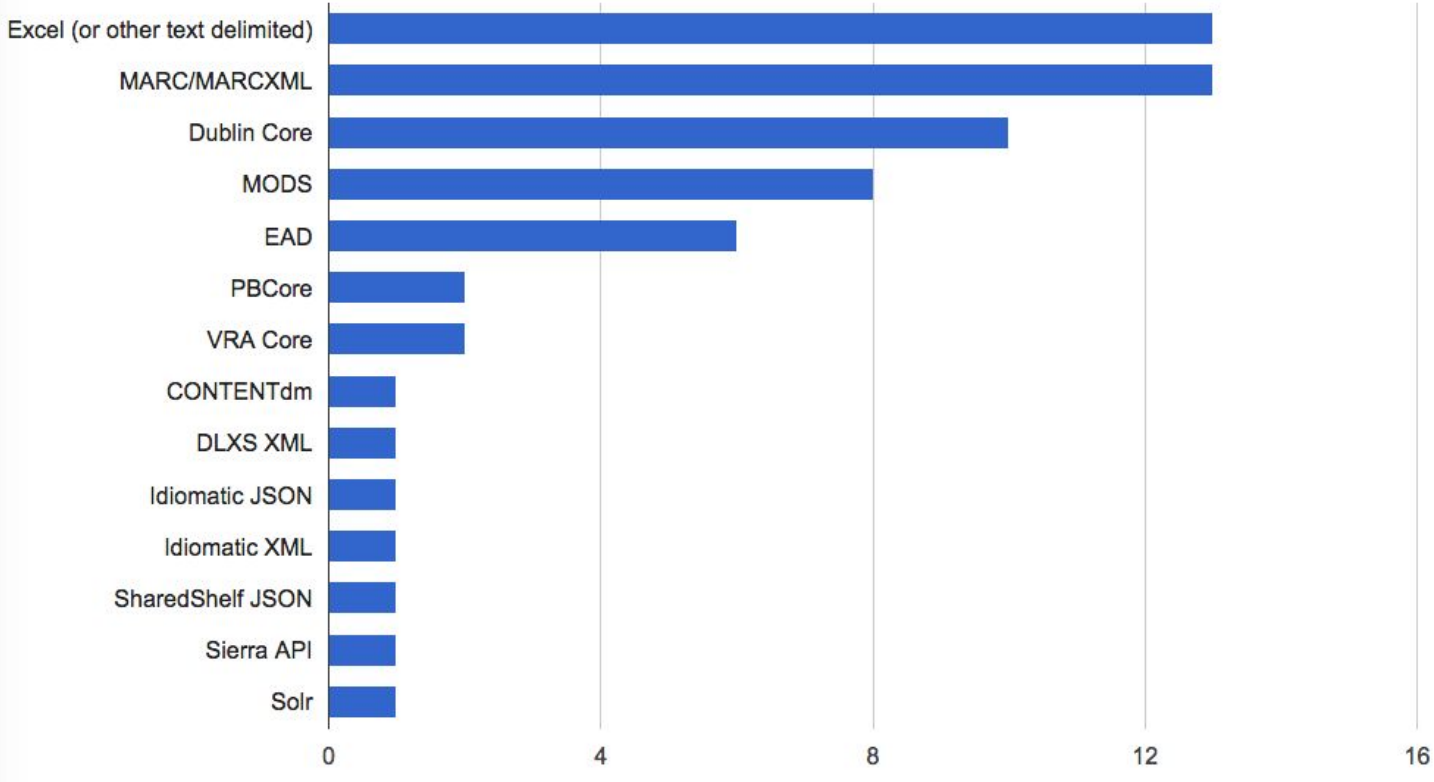
Who's creating metadata?



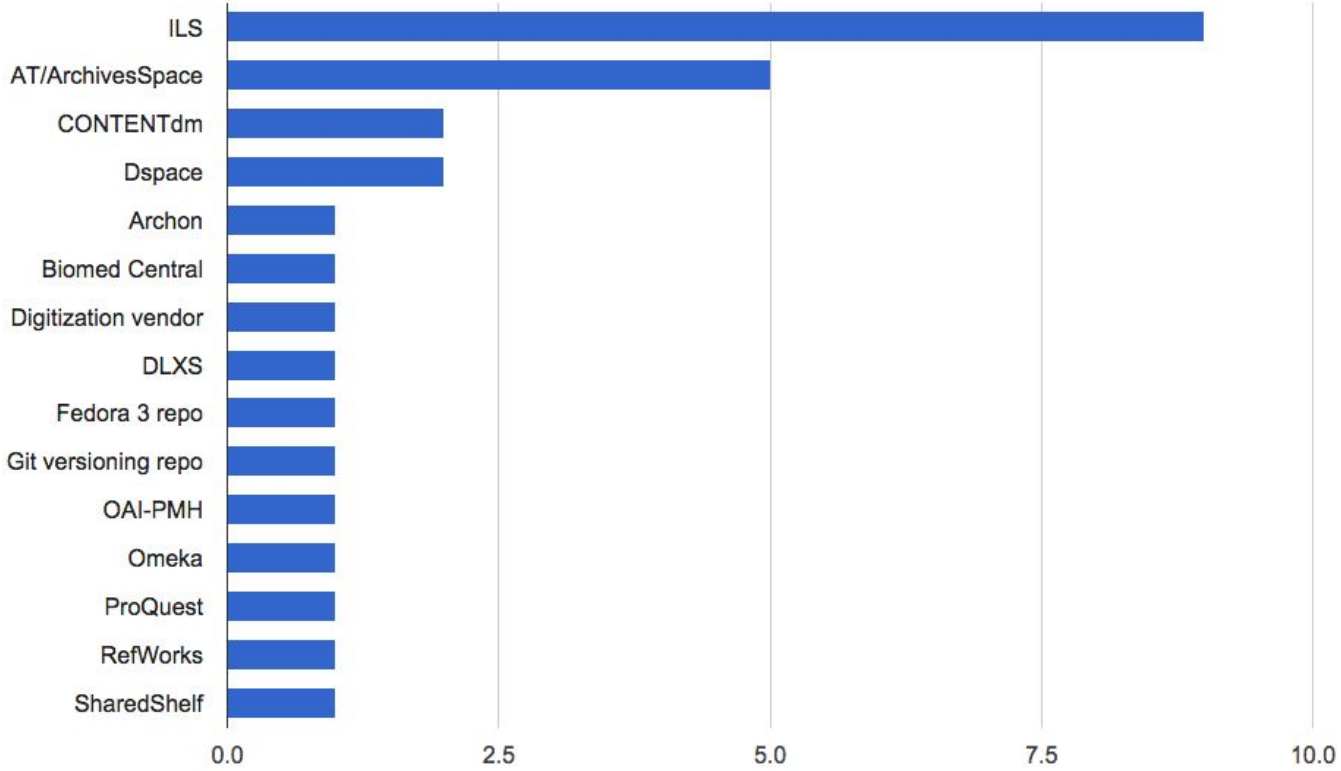
Where are you persisting the descriptive metadata?



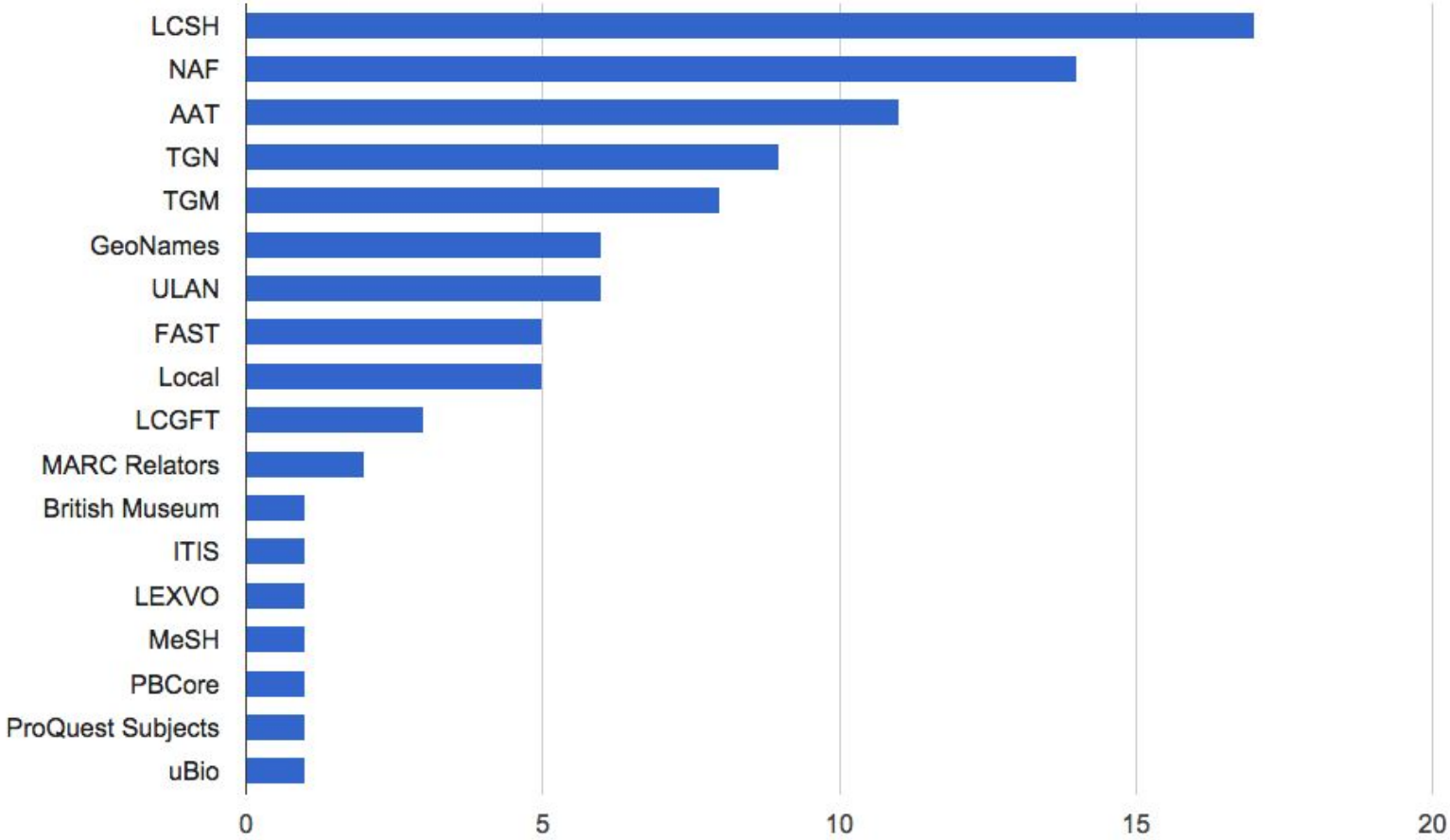
For harvested/ingested records, what record formats are being brought in?



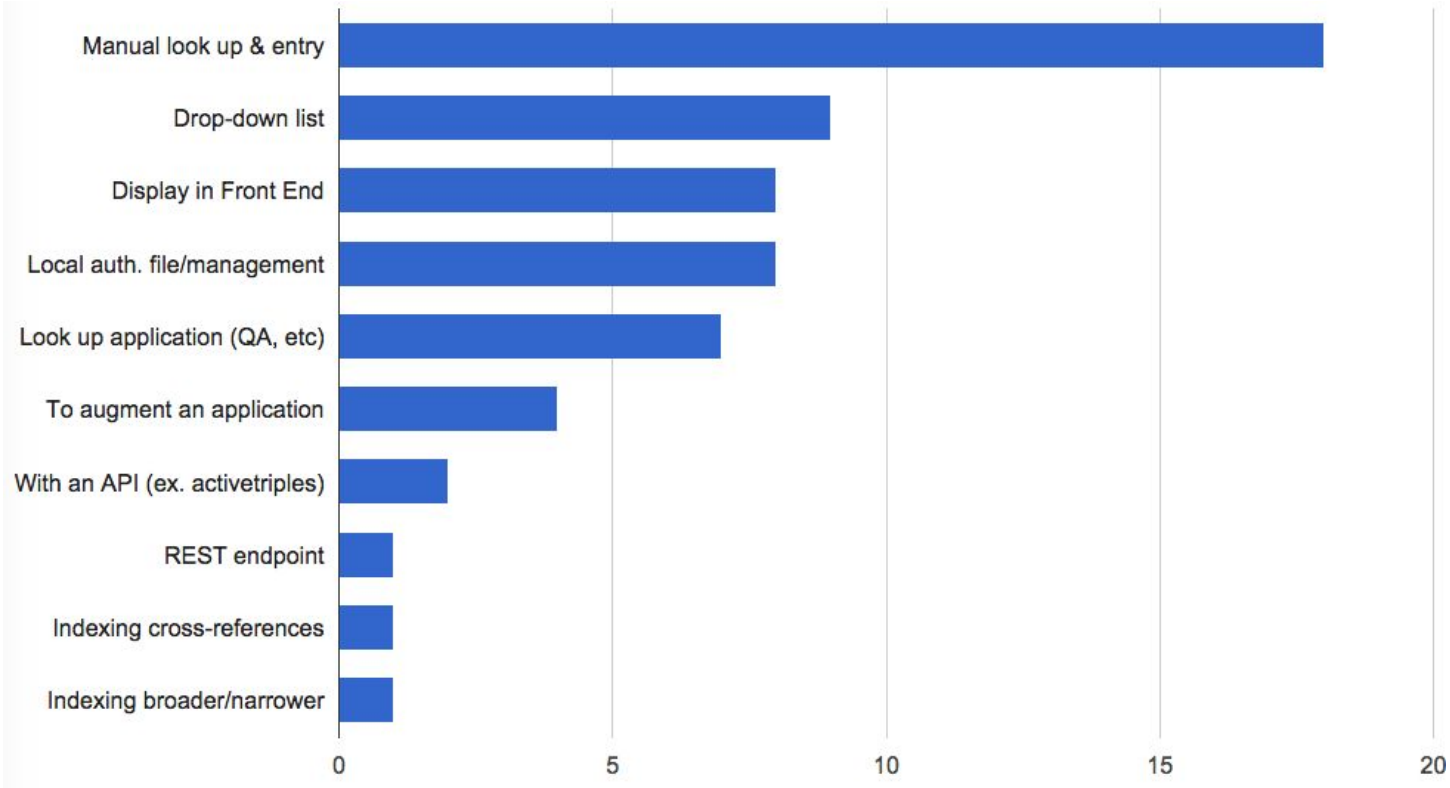
For harvested/ingested records, what system are these coming from?



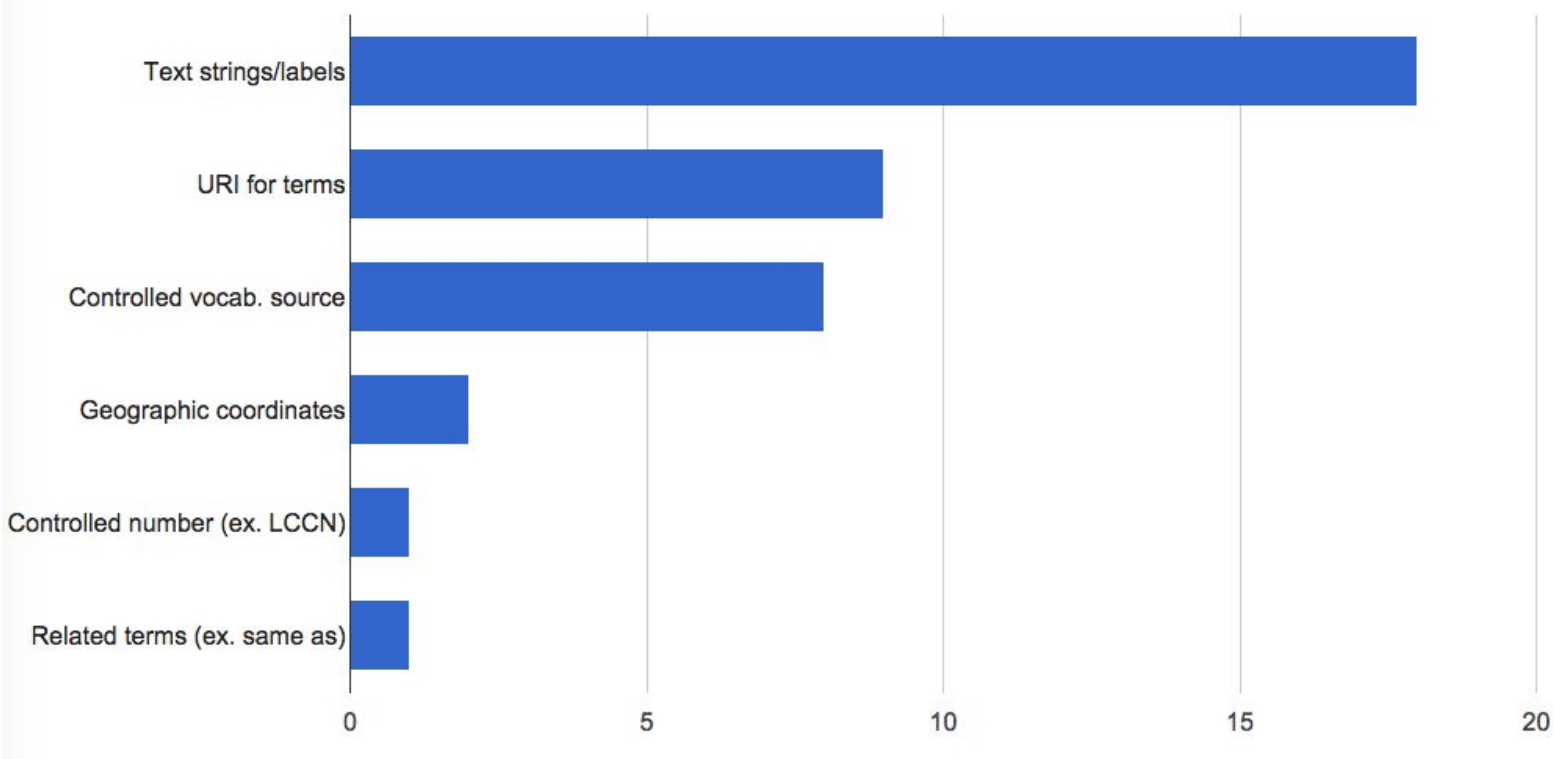
What controlled vocabularies are you using?



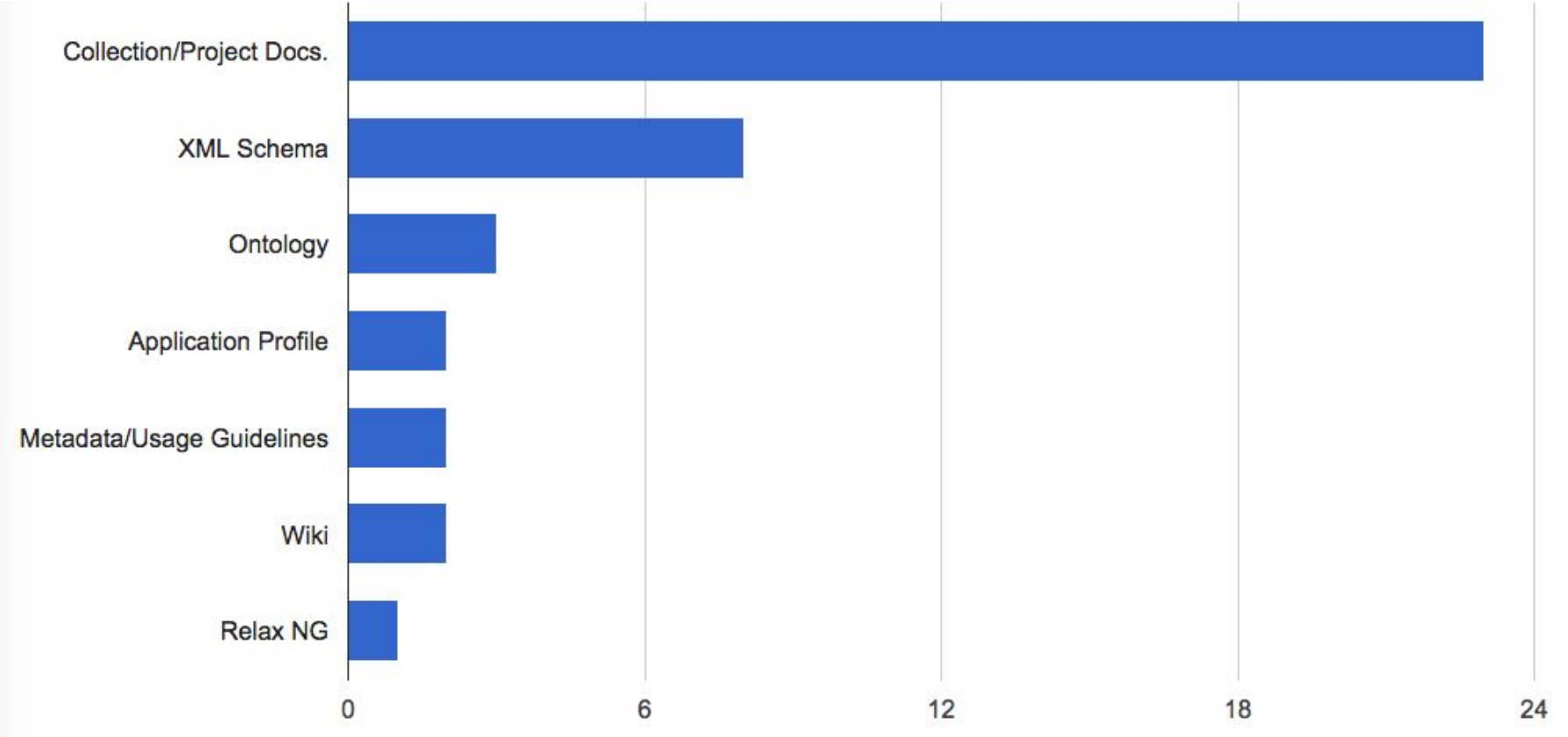
How are controlled vocabularies being used or managed?



For controlled vocabulary terms, what are you encoding?



What kind of documentation do you have?



Is there anything you'd like to do with your repository that you aren't currently doing?

- "Find the right vocabs/mappings to convert XML metadata into RDF"
- "Move to Fedora 4 and native RDF"
- "Make our metadata available via a SPARQL endpoint"
- "Link directory to existing controlled vocabularies instead of maintaining them locally"
- "Automated generation of MODS from other descriptive schemas"
- "Automated generation of PREMIS"
- "Move from storing strings to URIs"
- "Develop ontologies for each collection that are extensions of our core schema/ontology and publish as linked data"

Are there any roadblocks preventing you from doing new things with your repository?

- “Complexity of migrating MODS XML to RDF”
- “Educational blockers on learning how to apply existing tools”
- “No clear idea of how to do authority control or implement linked data”
- “The inflexibility of the Hydra data models make doing new things challenging, as diverging from Hydra data models inhibits sustainability over time”
- “The remaining questions/tooling around RDF in Hydra”
- “Converting PBCore XML to RDF”
- “Liabilities of mapping and nesting RDF”

Applied Linked Data (http://j.mp/hy-alds)

- Discussions of:
  - Broader and Narrower SKOS concepts such as a search for “sports” returning “football” or “soccer” subjects in the results.
  - Handling Alt Labels such “Beantown” as another name for “Boston” (https://goo.gl/Sks4gj).
  - How linked data type-ahead is absolutely great for demos but rarely seems to work in practice. Idea of a “Metadata Enrichment Interface” instead (https://goo.gl/OhwSuo).
  - Information from RDF Predicates most applications are not using that could be useful for tool tips (https://goo.gl/LgusuS).
  - **Linked Data Fragments**

Linked Data Fragments - The Problem

- Using RDF vocabularies for metadata creation and enhancement often relies on availability and performance of external services.
  - e.g. getting labels for subject headings to populate a search index
- These services, unfortunately, are not always reliable (e.g. SPARQL endpoints).
- Any implementation should be reusable across applications and ideally software independent.

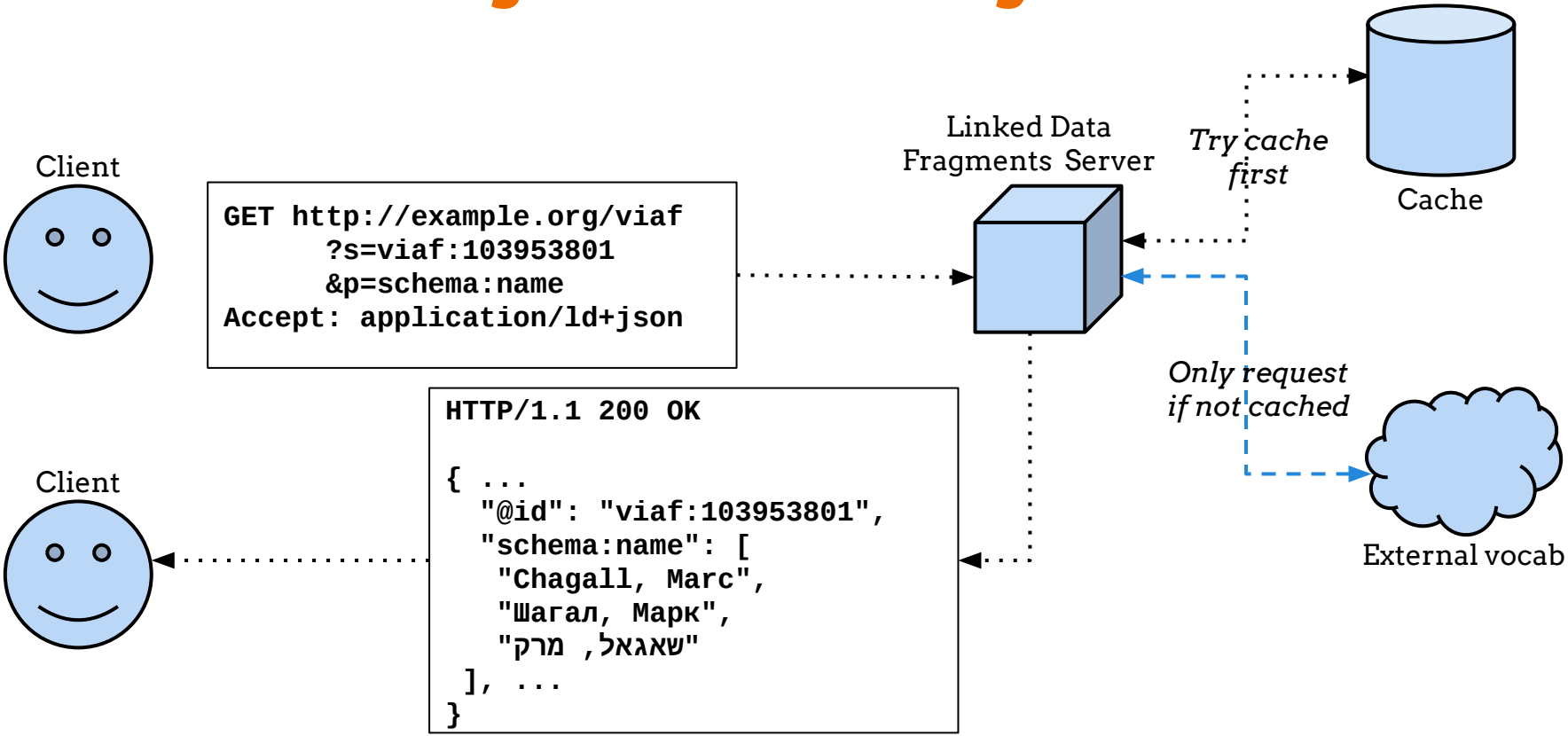
Linked Data Fragments - The Solution

- Implement a mechanism that allows you to use selector patterns to specify subsection of triples you care about, e.g. **skos:prefLabel** for LCSH.

- Provide a configurable cache layer for remote resources to speed up lookups on often-requested subjects.

- The most promising solution is Linked Data Fragments (http://linkeddatafragments.org/)
    - Simple standardized requests that are easily cacheable (less unique than SPARQL queries) with most query processing done client side

Linked Data Fragments - The Diagram



Linked Data Fragments - More Information

- Work In Progress: http://j.mp/hy-ldf
- Sample Config:

```
development:
  uri_endpoint: 'http://localhost:3000/{?subject}'
  uri_root: 'http://localhost:3000/#dataset'
  cache_backend:
    provider: 'marmotta'
    url: 'http://localhost:8983/marmotta'
    context: 'linked_data_fragments_dev'
```

- Sample URL request from above to return TTL:
  - http://localhost:3000/http://dbpedia.org/resource/Berlin?format=ttl

Applied Linked Data - Future Work

- Additional backend support beyond Marmotta for Linked Data Fragments.

- “Sidecar Indexer”
  - Polls Solr based on a frequency an institution has configured.
  - Gets relevant URI’s and corresponding labels for a record.
  - Uses Linked Data Fragments to check that the labels are current. If not, updates that field using a Solr Atomic Update.

MODS Subgroup (https://goo.gl/i1XiDq)

- MODS allows for a high level of specificity. Take, for example, a title that could be supplied or not and one of the following:
  - Single Primary title
  - Parallel title
  - Translated title
  - Uniform title
  - Alternative title

- That level of specificity doesn't currently seem to exist in most Fedora 4 systems while avoiding blank nodes. What is one to do?
  - Could simply create our own "one off" RDF equivalent that meets our institutional needs. But then we lose the benefit of a shared standard and we could choose predicates that don't stand the test of time.

MODS Subgroup - Approach

- Going through the MODS top level elements and creating mappings for each elements and the subelements / attributes it can contain.
  - A “simple” mapping for the primary piece of information being captured.
  - A “complex” mapping that keeps all **relevant** specificity.
    - For a relevancy example, we all agreed that separating “subtitle” from “title” had no use case that anyone could come up with. So our complex case concatenates those subelements.

- Once we have complemented our mapping document, we plan to do a community coding effort for:
  - Converting a MODS XML record into our agreed upon “unofficial” standard.
  - Converting from the “unofficial” standard back into MODS XML.

MODS Subgroup - Current Work

- Title Collaboration Document (w/ Simple and Complex Tabs): https://goo.gl/R3Rwgu
  - See discussion on a predicate proposed for (titleForSort): https://github.com/pulibrary/plum/pull/54

- The next element we are looking at is MODS:Name with institutions doing individual mappings attempt at: https://goo.gl/pulV7g

- If you are interested in more about this at the conference, we have an unconference session at 3:30 to 4:20 on Thursday! http://connect2015.curationexperts.com/sessions/49

Upcoming Activities

Time-based structural

Come to Breakout Sessions and Talk more Metadata!

Thank you!

Questions