MANAGING UNSTRUCTURED METADATA AT ESS

Gareth Murphy, European Spallation Source

Helmholtz-Zentrum Berlin 2018-03-19
WHAT IS METADATA?

- a set of data that describes and gives information about other data.
- Can classify into separate types
  - administrative
  - structural
  - descriptive
  - scientific
... is often notoriously incomplete. Additional quantities and assumptions necessary to interpret the data may initially only be recorded on scraps of paper, hard-coded into analysis software or only exist in the experimenter's head.

- more extensive
- less predictable - "unknown unknowns"
ESS Neutron Instruments:

Project scope - 15 instruments + test beamline

ESS Lead Partners for instrument construction

37 ESS In-Kind Partners also collaborate on sample environment, data management systems etc.
METADATA AT ESS

- ESS metadata is complex
- Not always predictable what is important
- Most scientific data is unFAIR
- Findable, Accessible, Interoperable, Reproducible
CURRENTLY METADATA CAN BE STORED

• In filename
  (run1_vanadium)
• In Excel files on HDD or Dropbox - not accessible
• Not at all
USING A DATA CATALOGUE

- One source of information
- All data can be found on one website
- Manages permissions, publication
- User reads a publication, can get data from catalogue
WHY NOT USE EXISTING TOOLS?

- performance and flexibility issues
- SQL database technology aging
- NoSQL offers more opportunities for unstructured data
- SQL is highly structured in tables with rows and columns
- MongoDB, a NoSQL DB, uses documents organised in collections.
SCICAT

- [github.com/ScicatProject](github.com/ScicatProject)
- Manage the meta data of raw and derived data which is taken at experiment facilities
- Administrative: data management lifecycle, ownership, file
- Scientific: describing the sample, beamline and experiment parameters relevant for the users data analysis
DATASET, DATAFILE, DATABLOCKS

- A dataset includes all metadata related to a set of files.
- Has an owner, ORCid, creation time, science metadata etc, and includes datafile references.
- A datafile has path, size, permissions.
- Datablocks are storage media for archiving, one datafile can be stored across on or more datablocks.
RAW DATASET VS DERIVED DATASET

• RawDataset - experimental data directly from beamline

• Derived Dataset - has extra fields indicating origin of generated data.
Each dataset stores scientific metadata as an array.

Users will be able to add their own metadata fields.
SCICAT

- Enables management of the lifecycle of the data from creation, data analysis and eventual deletion
- Data can be linked to proposals and samples
- Data can be linked to publications (DOI, PID)
- Data can be migrated to and from longterm storage on tape
SCICAT

- Reproducibility - Helps keeping track of data provenance (i.e. the steps leading to the final results)
- Allows checking scientific integrity (checksum of data)
- Findability - Allows find data based on the metadata (your own data and other peoples public data)
- In the long term: help to automate standardized analysis workflow
SCICAT - COLLECTION OF MICROSERVICES

- Web frontend (catanie - angular based)
- API service backend (catamel- Automatically generated using IBM's loopback.io
- Database MongoDB
- Message/job queuing system (currently RabbitMQ -> migrating to Kafka)
- Flow-based editor Node-RED
KUBERNETES DEPLOYMENT

- Can test kubernetes deployment using minikube, without installing a full cluster
- Try it yourself!
- Working minikube config at http://www.github.com/ScicatProject/localdeploy
- RBAC disabled by default - can re-enable
KUBERNETES
MONGODB

- NoSQL storage of metadata, login, jobs
- Database requires persistent storage
- Currently we store on k8s nodes, not a longterm solution
CATANIE

- Angular website
- Javascript generated static html
- Data served by catamel
- Viewable on PC, phone etc
NODE-RED

- Translate metadata from Kafka stream XML to catamel format (JSON)
- Can be used to add in extra data cleaning or processing
CATAMEL

- Metadata server
- Loopback generated API
- Models defined in JSON
- Also provides connectivity to authentication server(s)
### RawDataset

<table>
<thead>
<tr>
<th>METHOD</th>
<th>URL</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PATCH</td>
<td>/RawDatasets</td>
<td>Patch an existing model instance or insert a new one into the data source.</td>
</tr>
<tr>
<td>GET</td>
<td>/RawDatasets</td>
<td>Find all instances of the model matched by filter from the data source.</td>
</tr>
<tr>
<td>PUT</td>
<td>/RawDatasets</td>
<td>Replace an existing model instance or insert a new one into the data source.</td>
</tr>
<tr>
<td>POST</td>
<td>/RawDatasets</td>
<td>Create a new instance of the model and persist it into the data source.</td>
</tr>
<tr>
<td>PATCH</td>
<td>/RawDatasets/[id]</td>
<td>Patch attributes for a model instance and persist it into the data source.</td>
</tr>
<tr>
<td>GET</td>
<td>/RawDatasets/[id]</td>
<td>Find a model instance by [id] from the data source.</td>
</tr>
<tr>
<td>HEAD</td>
<td>/RawDatasets/[id]</td>
<td>Check whether a model instance exists in the data source.</td>
</tr>
<tr>
<td>PUT</td>
<td>/RawDatasets/[id]</td>
<td>Replace attributes for a model instance and persist it into the data source.</td>
</tr>
<tr>
<td>DELETE</td>
<td>/RawDatasets/[id]</td>
<td>Delete a model instance by [id] from the data source.</td>
</tr>
<tr>
<td>GET</td>
<td>/RawDatasets/[id]/datablocks</td>
<td>Queries datablocks of the RawDataset</td>
</tr>
<tr>
<td>POST</td>
<td>/RawDatasets/[id]/datablocks</td>
<td>Creates a new instance in datablocks of this model.</td>
</tr>
<tr>
<td>DELETE</td>
<td>/RawDatasets/[id]/datablocks</td>
<td>Deletes all datablocks of this model.</td>
</tr>
<tr>
<td>GET</td>
<td>/RawDatasets/[id]/datablocks/[fr]</td>
<td>Find a related item by id for datablocks.</td>
</tr>
<tr>
<td>PUT</td>
<td>/RawDatasets/[id]/datablocks/[fr]</td>
<td>Update a related item by id for datablocks.</td>
</tr>
<tr>
<td>DELETE</td>
<td>/RawDatasets/[id]/datablocks/[fr]</td>
<td>Delete a related item by id for datablocks.</td>
</tr>
</tbody>
</table>
DEPLOYMENT OF SCICAT

- still outstanding:
- persistent storage
- file viewer
- data download
CONCLUSION

- SciCat on Kubernetes will be able to provide metadata services for ESS users needs
- MongoDB backend can handle unstructured metadata
- Users will be able to add their own metadata