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

SCIENTIFIC METADATA

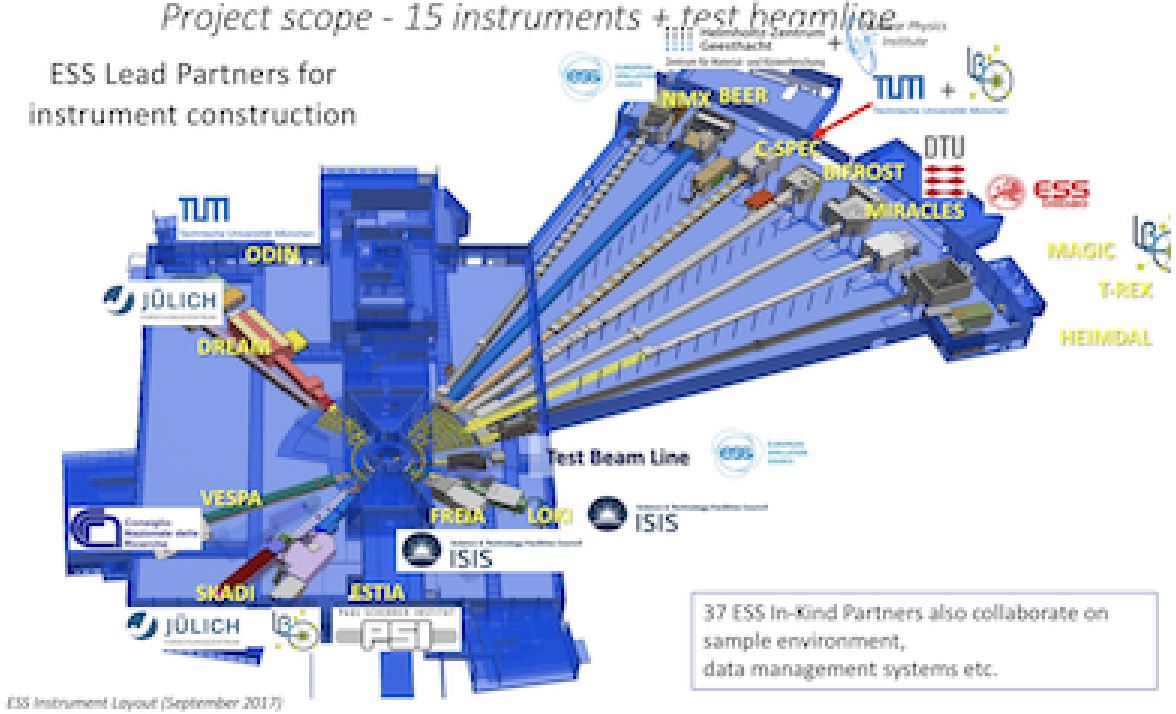
… is often notoriously incomplete. Additional quantities and assumptions necessary to interpret the data may initially only be recorded on scraps of paper, hard-coded into analysis software or only exist in the experimenter's head.

- more extensive
- less predictable - "unknown unknowns"

ESS Neutron Instruments:

Project scope - 15 instruments + test beamline

ESS Lead Partners for instrument construction



ESS Instrument Layout (September 2017)

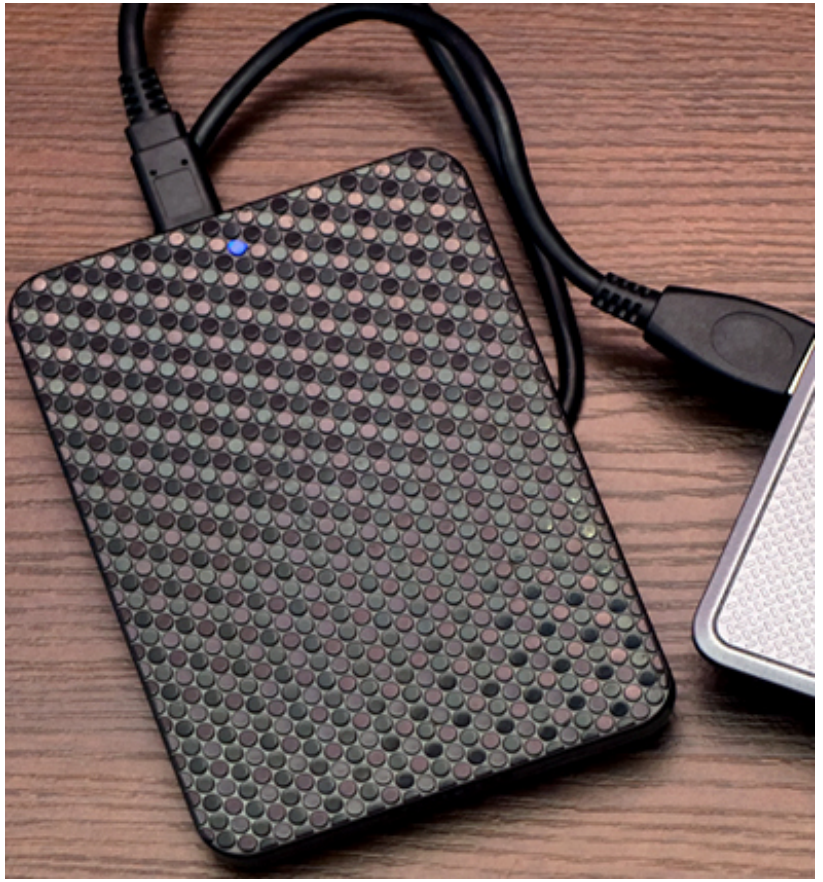
37 ESS In-Kind Partners also collaborate on sample environment, data management systems etc.

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

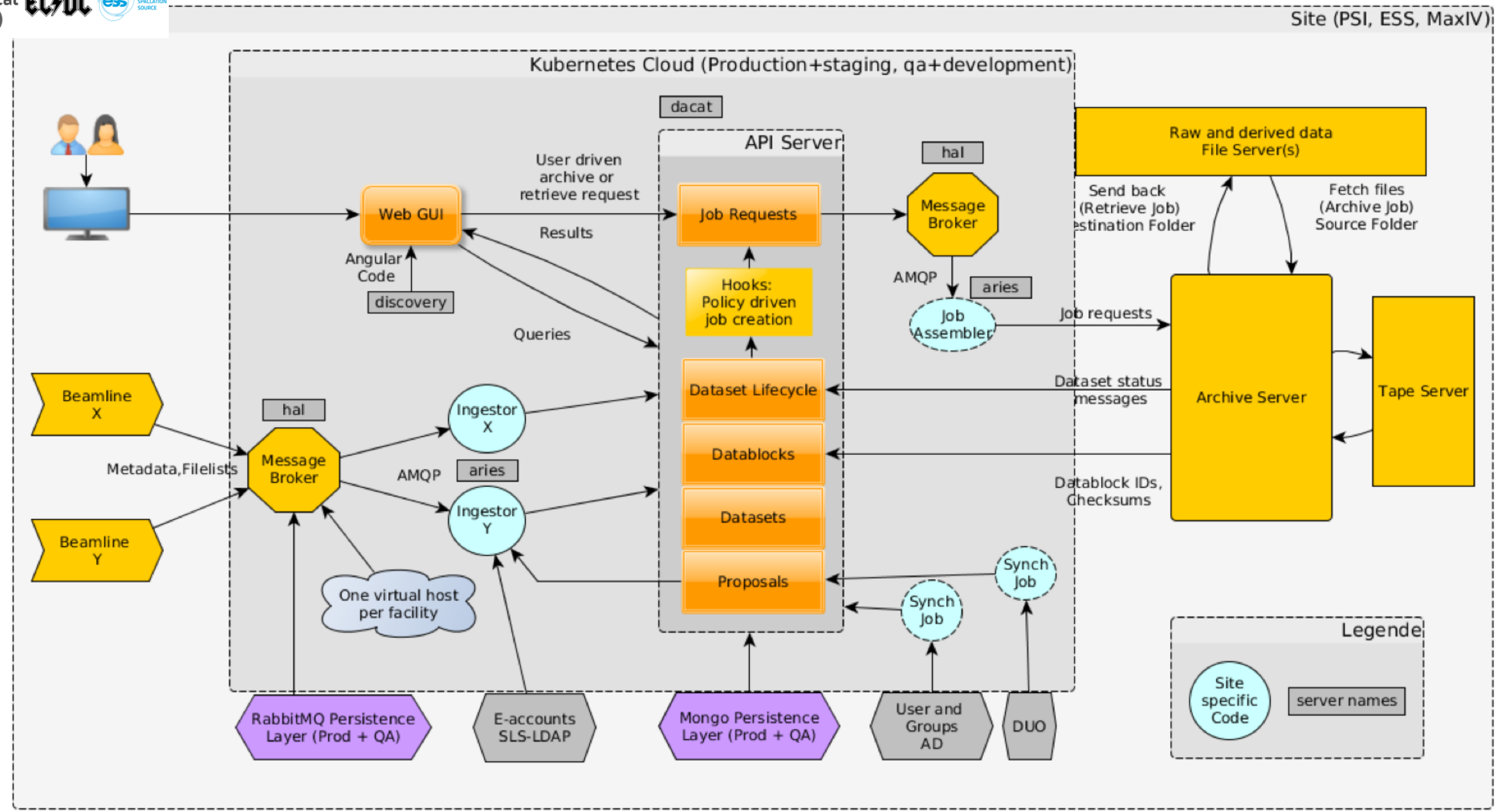


USING A DATA CATALOGUE

- 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.


Site (PSI, ESS, MaxIV)
Kubernetes Cloud (Production+staging, qa+development)
dacat
API Server
Job Requests
Hooks:
Policy driven
job creation
Dataset Lifecycle
Datablocks
Datasets
Proposals
Web GUI
User driven
archive or
retrieve request
Results
Queries
Angular
Code
discovery
hal
Message
Broker
AMQP
aries
Job
Assembler
Raw and derived data
File Server(s)
Send back
(Retrieve Job)
estination Folder
Fetch files
(Archive Job)
Source Folder
Job requests
Archive Server
Tape Server
Dataset status
messages
Datablock IDs,
Checksums
Beamline
X
Beamline
Y
Metadata,Filelists
hal
Message
Broker
Ingestor
X
Ingestor
Y
AMQP
aries
One virtual host
per facility
Synch
Job
Synch
Job
RabbitMQ Persistence
Layer (Prod + QA)
E-accounts
SLS-LDAP
Mongo Persistence
Layer (Prod + QA)
User and
Groups
AD
DUO
Legende
Site
specific
Code
server names

SCICAT

- 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

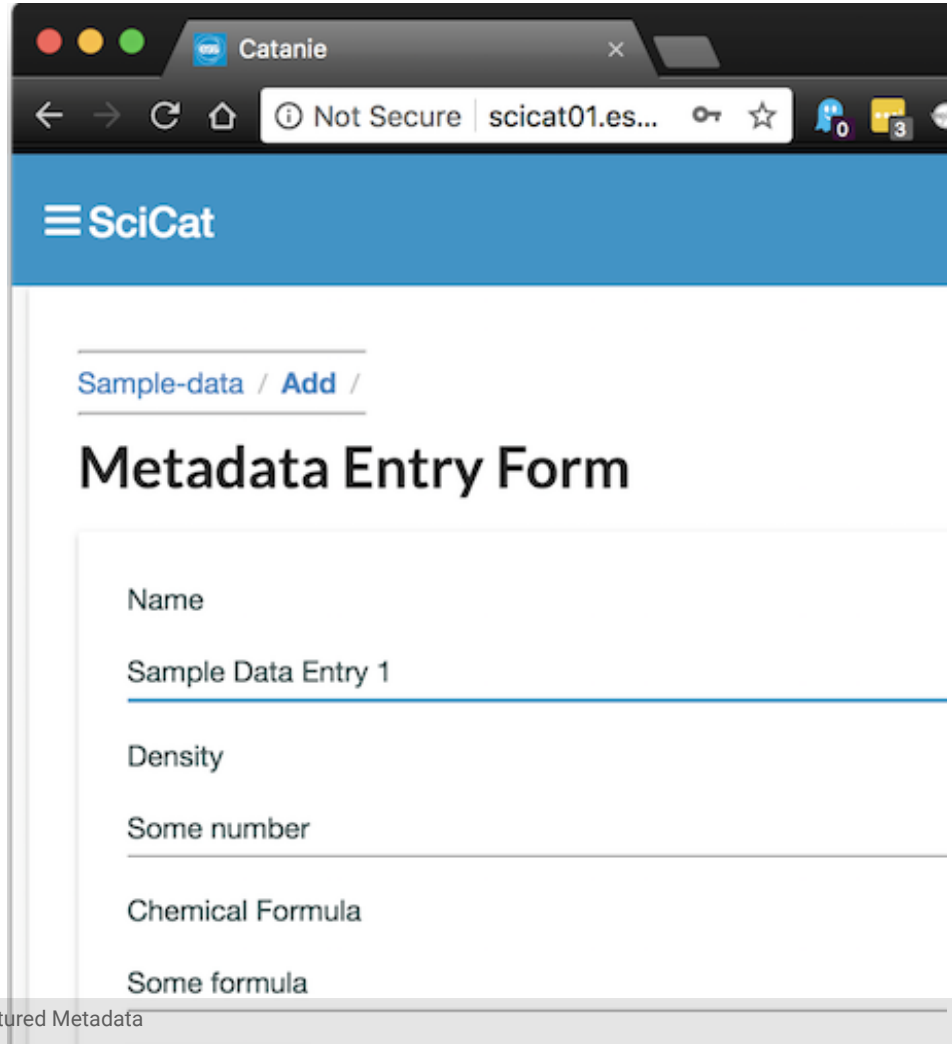
- A dataset includes all metadata related to a set of files
- Has an owner, ORCid, creation time, science metadata etc, and includes datafile references
- A datafile has path, size, permissions
- Datablocks are storage media for archiving, one datafile can be stored across on or more datablocks

RAW DATASET VS DERIVED DATASET

- RawDataset - experimental data directly from beamline
- Derived Dataset - has extra fields indicating origin of generated data.

SCIENTIFIC METADATA

- Each dataset stores scientific metadata as an array.
- Users will be able to add their own metadata fields.



SCICAT

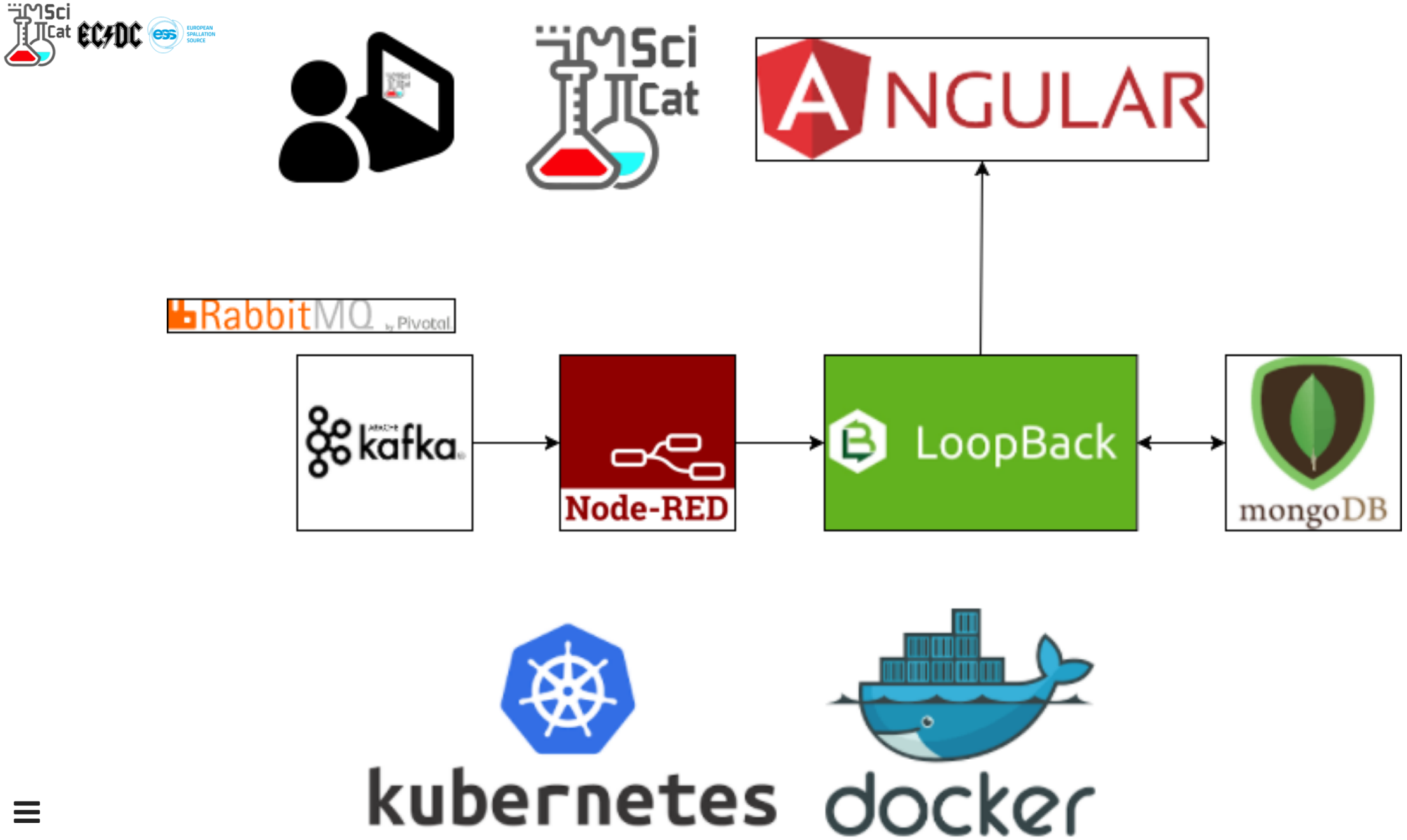
- Enables management of the lifecycle of the data from creation , data analysis and eventual deletion
- Data can be linked to proposals and samples
- Data can be linked to publications (DOI, PID)
- Data can be migrated to and from longterm storage on tape

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 meta data (your own data and other peoples public data)
- In the long term:help to automate standardized analysis workflow

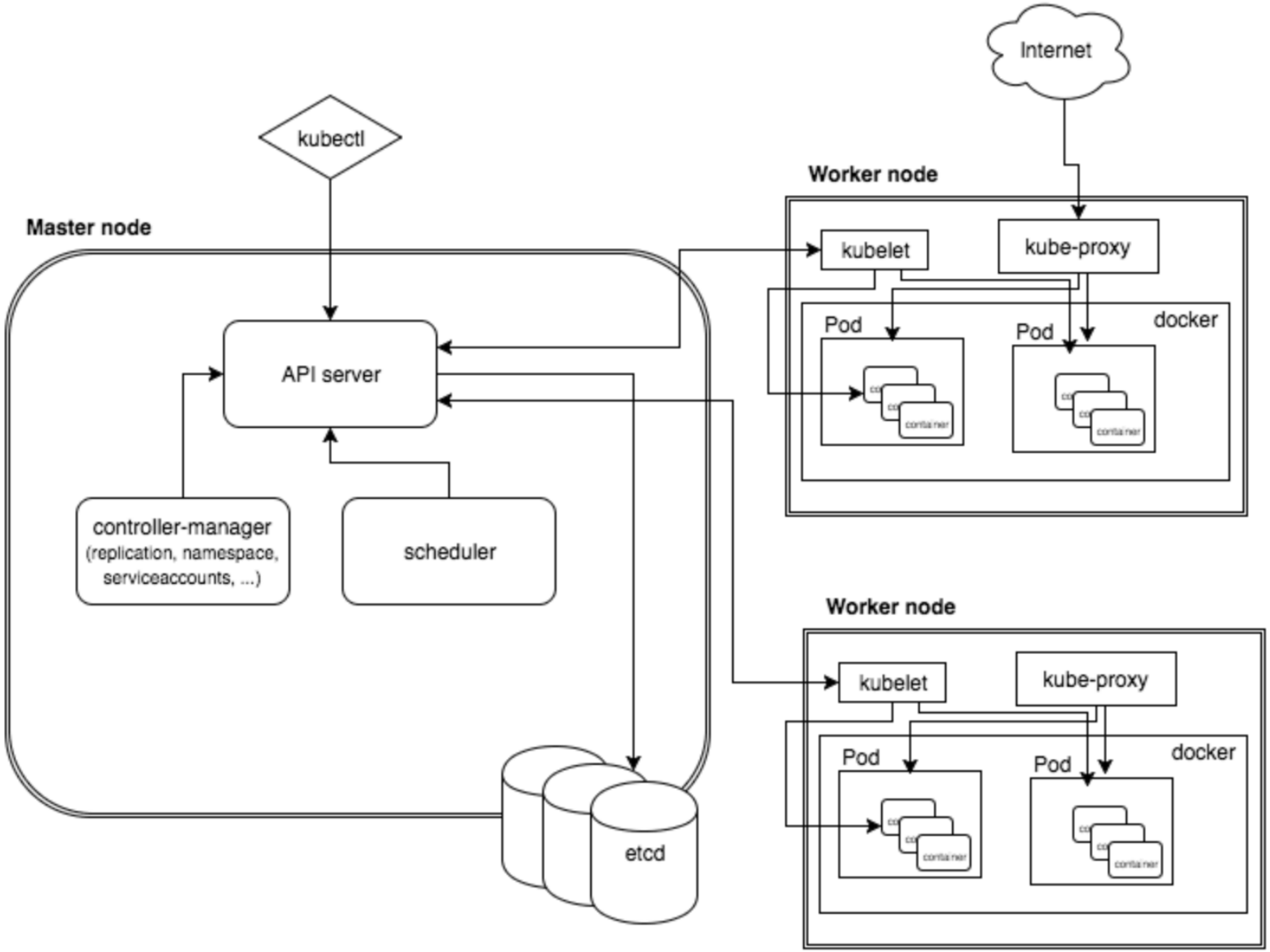
SCICAT - COLLECTION OF MICROSERVICES

- 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

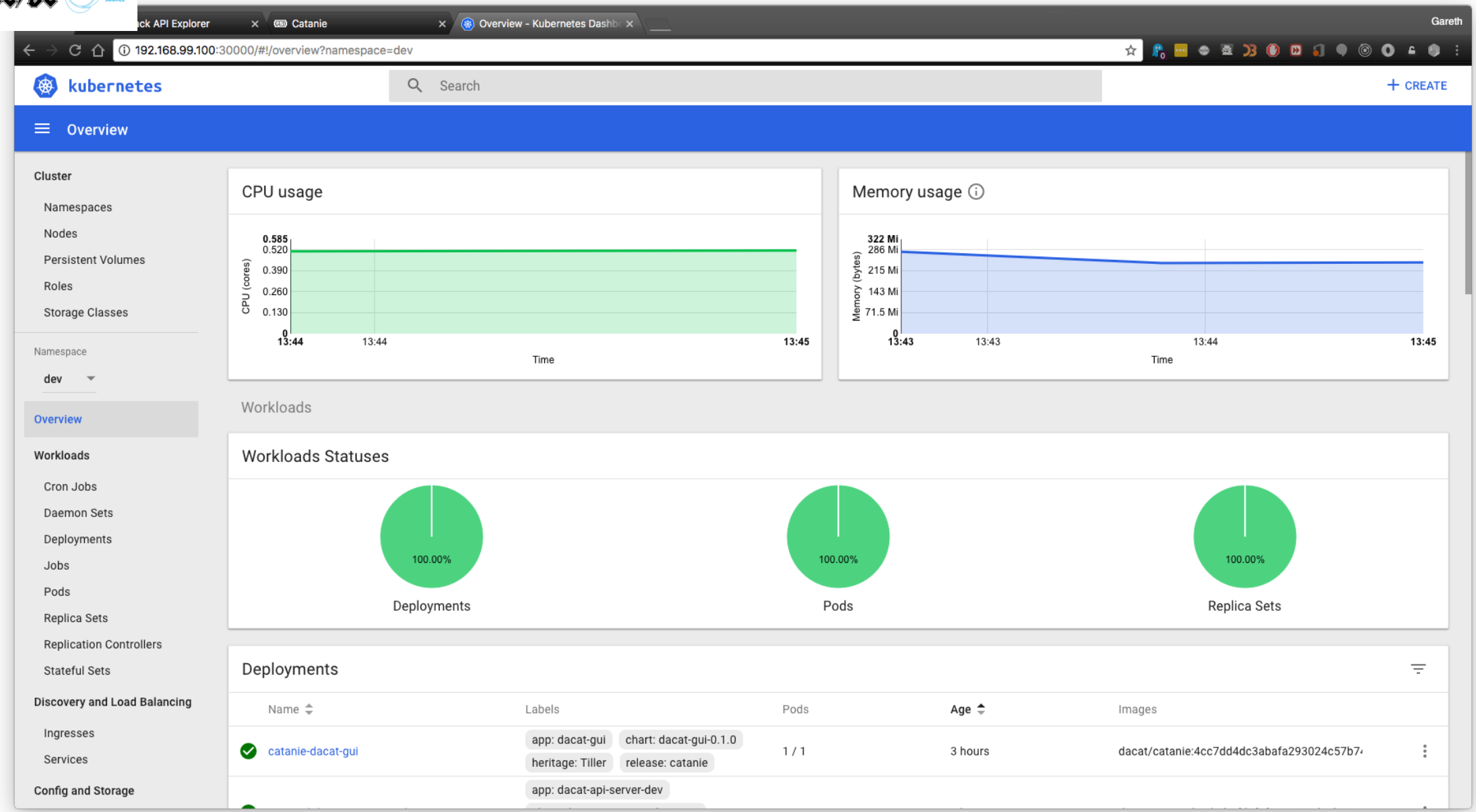

SciCat
EC/DC
ESS EUROPEAN SPALLATION SOURCE
SciCat
ANGULAR
RabbitMQ by Pivotal
kafka
Node-RED
LoopBack
mongoDB
kubernetes
docker

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


Internet
kubectl
Worker node
Master node
kubelet
kube-proxy
Pod
Pod
docker
API server
controller-manager
(replication, namespace,
serviceaccounts, ...)
scheduler
Worker node
kubelet
kube-proxy
Pod
Pod
docker
etcd

KUBERNETES

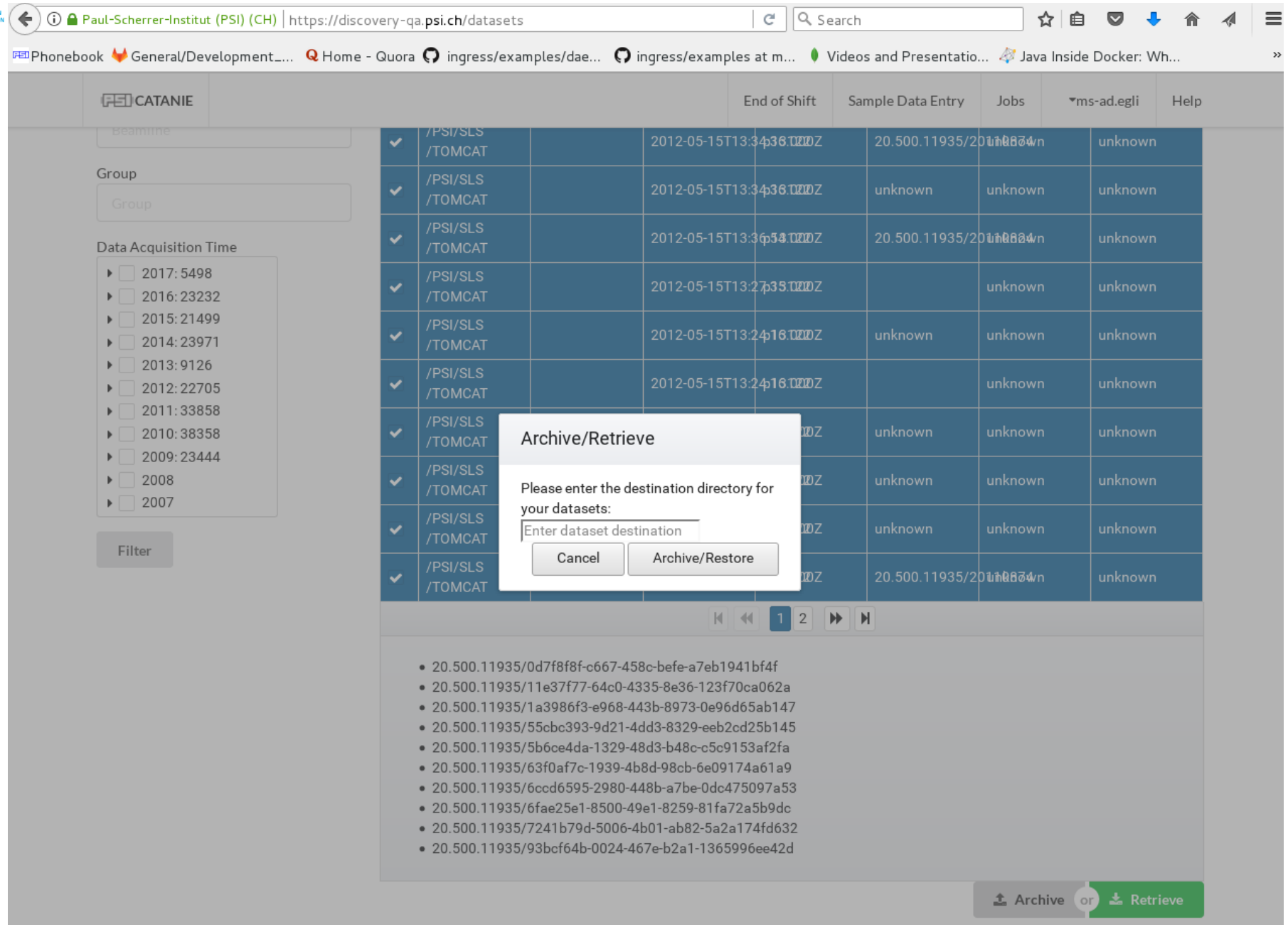

ck API Explorer
Catanie
Overview - Kubernetes Dashb
Gareth
192.168.99.100:30000/#!/overview?namespace=dev
kubernetes
Search
+ CREATE
Overview
Cluster
Namespaces
Nodes
Persistent Volumes
Roles
Storage Classes
Namespace
dev
Overview
Workloads
Cron Jobs
Daemon Sets
Deployments
Jobs
Pods
Replica Sets
Replication Controllers
Stateful Sets
Discovery and Load Balancing
Ingresses
Services
Config and Storage
CPU usage
0.585
0.520
0.390
0.260
0.130
0
CPU (cores)
13:44
13:44
13:45
Time
Memory usage
322 Mi
286 Mi
215 Mi
143 Mi
71.5 Mi
0
Memory (bytes)
13:43
13:43
13:44
13:45
Time
Workloads
Workloads Statuses
100.00%
Deployments
100.00%
Pods
100.00%
Replica Sets
Deployments
Name
Labels
Pods
Age
Images
catanie-dacat-gui
app: dacat-gui
chart: dacat-gui-0.1.0
heritage: Tiller
release: catanie
1 / 1
3 hours
dacat/catanie:4cc7dd4dc3abafa293024c57b7
app: dacat-api-server-dev

MONGODB

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

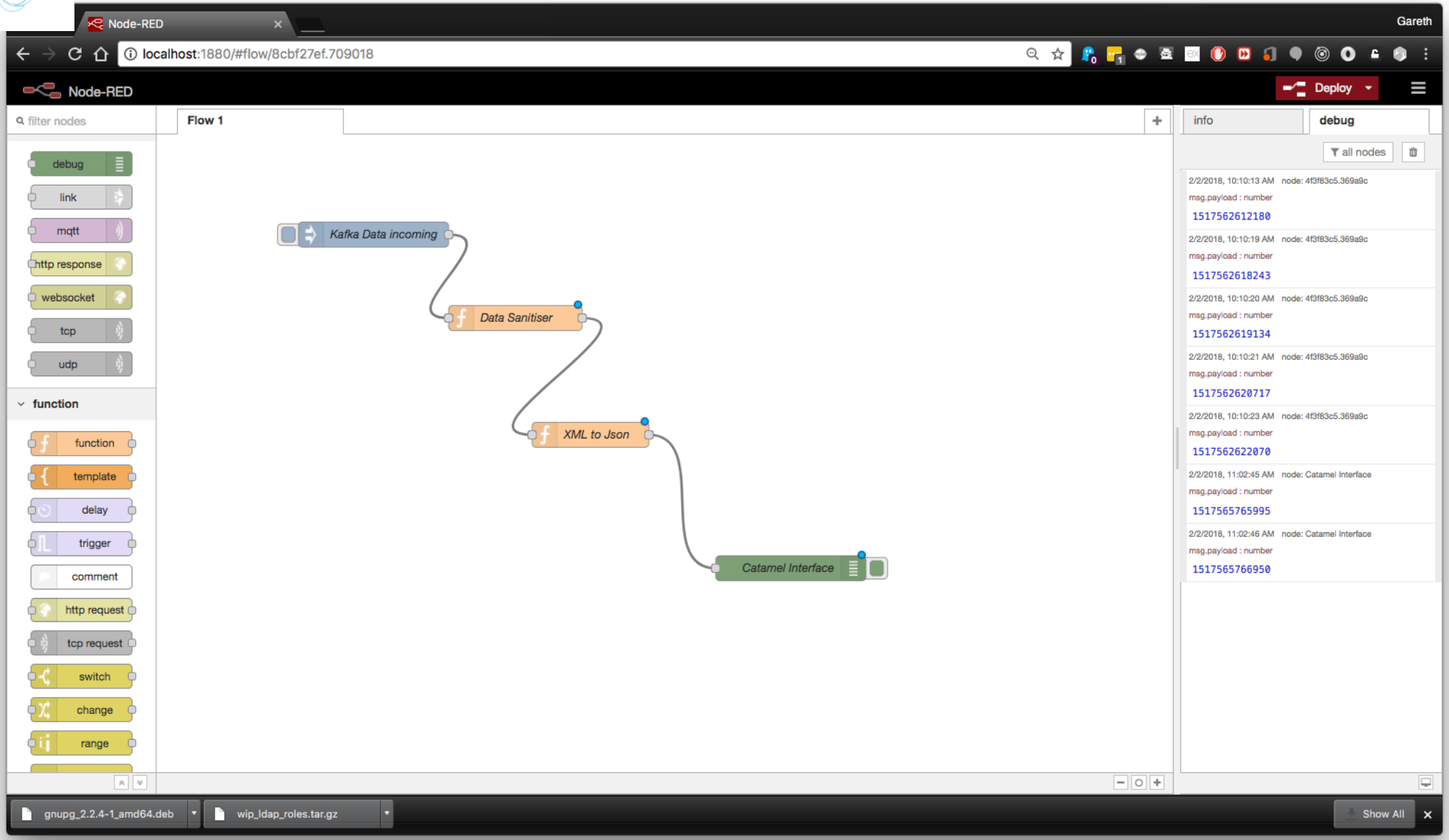
CATANIE

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


Paul-Scherrer-Institut (PSI) (CH) | https://discovery-qa.psi.ch/datasets
Search
Phonebook
General/Development_...
Home - Quora
ingress/examples/dae...
ingress/examples at m...
Videos and Presentatio...
Java Inside Docker: Wh...
CATANIE
End of Shift
Sample Data Entry
Jobs
ms-ad.egli
Help
Group
Data Acquisition Time
2017: 5498
2016: 23232
2015: 21499
2014: 23971
2013: 9126
2012: 22705
2011: 33858
2010: 38358
2009: 23444
2008
2007
Filter
/PSI/SLS /TOMCAT
unknown
Archive/Retrieve
Please enter the destination directory for your datasets:
Enter dataset destination
Cancel
Archive/Restore
1
2
20.500.11935/0d7f8f8f-c667-458c-befe-a7eb1941bf4f
20.500.11935/11e37f77-64c0-4335-8e36-123f70ca062a
20.500.11935/1a3986f3-e968-443b-8973-0e96d65ab147
20.500.11935/55cbc393-9d21-4dd3-8329-eeb2cd25b145
20.500.11935/5b6ce4da-1329-48d3-b48c-c5c9153af2fa
20.500.11935/63f0af7c-1939-4b8d-98cb-6e09174a61a9
20.500.11935/6ccd6595-2980-448b-a7be-0dc475097a53
20.500.11935/6fae25e1-8500-49e1-8259-81fa72a5b9dc
20.500.11935/7241b79d-5006-4b01-ab82-5a2a174fd632
20.500.11935/93bcf64b-0024-467e-b2a1-1365996ee42d
Archive
or
Retrieve

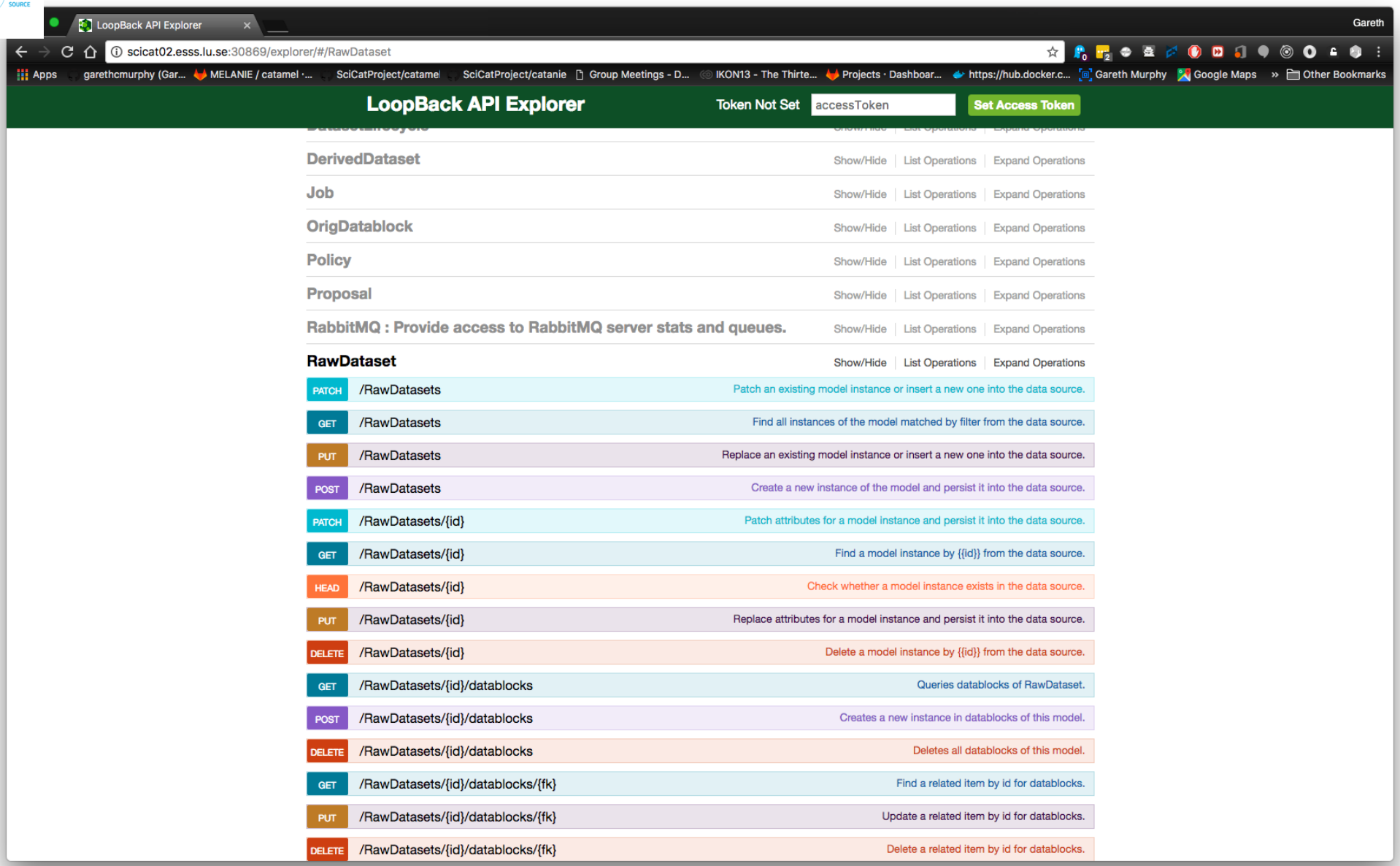
NODE-RED

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


Node-RED
Gareth
localhost:1880/#flow/8cbf27ef.709018
Node-RED
Deploy
filter nodes
Flow 1
info
debug
all nodes
debug
link
mqtt
http response
websocket
tcp
udp
function
function
template
delay
trigger
comment
http request
tcp request
switch
change
range
Kafka Data incoming
Data Sanitiser
XML to Json
Catamel Interface
2/2/2018, 10:10:13 AM node: 4f3f83c5.369a9c
msg.payload : number
1517562612180
2/2/2018, 10:10:19 AM node: 4f3f83c5.369a9c
msg.payload : number
1517562618243
2/2/2018, 10:10:20 AM node: 4f3f83c5.369a9c
msg.payload : number
1517562619134
2/2/2018, 10:10:21 AM node: 4f3f83c5.369a9c
msg.payload : number
1517562620717
2/2/2018, 10:10:23 AM node: 4f3f83c5.369a9c
msg.payload : number
1517562622070
2/2/2018, 11:02:45 AM node: Catamel Interface
msg.payload : number
1517565765995
2/2/2018, 11:02:46 AM node: Catamel Interface
msg.payload : number
1517565766950
gnupg_2.2.4-1_amd64.deb
wip_ldap_roles.tar.gz
Show All

CATAMEL

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


LoopBack API Explorer
Gareth
scicat02.esss.lu.se:30869/explorer/#/RawDataset
Apps
garethcmurphy (Gar...
MELANIE / catamel ·...
SciCatProject/catamel
SciCatProject/catanie
Group Meetings - D...
IKON13 - The Thirte...
Projects · Dashboar...
https://hub.docker.c...
Gareth Murphy
Google Maps
Other Bookmarks
LoopBack API Explorer
Token Not Set
accessToken
Set Access Token
DerivedDataset
Job
OrigDatablock
Policy
Proposal
RabbitMQ : Provide access to RabbitMQ server stats and queues.
RawDataset
Show/Hide
List Operations
Expand Operations
PATCH /RawDatasets Patch an existing model instance or insert a new one into the data source.
GET /RawDatasets Find all instances of the model matched by filter from the data source.
PUT /RawDatasets Replace an existing model instance or insert a new one into the data source.
POST /RawDatasets Create a new instance of the model and persist it into the data source.
PATCH /RawDatasets/{id} Patch attributes for a model instance and persist it into the data source.
GET /RawDatasets/{id} Find a model instance by {{id}} from the data source.
HEAD /RawDatasets/{id} Check whether a model instance exists in the data source.
PUT /RawDatasets/{id} Replace attributes for a model instance and persist it into the data source.
DELETE /RawDatasets/{id} Delete a model instance by {{id}} from the data source.
GET /RawDatasets/{id}/datablocks Queries datablocks of RawDataset.
POST /RawDatasets/{id}/datablocks Creates a new instance in datablocks of this model.
DELETE /RawDatasets/{id}/datablocks Deletes all datablocks of this model.
GET /RawDatasets/{id}/datablocks/{fk} Find a related item by id for datablocks.
PUT /RawDatasets/{id}/datablocks/{fk} Update a related item by id for datablocks.
DELETE /RawDatasets/{id}/datablocks/{fk} Delete a related item by id for datablocks.

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