Evaluation of data catalogue systems

ELI ALPS-PaNOSC WP3 Catalogue Integration Best Practices Meeting 05/19/2021

Outline

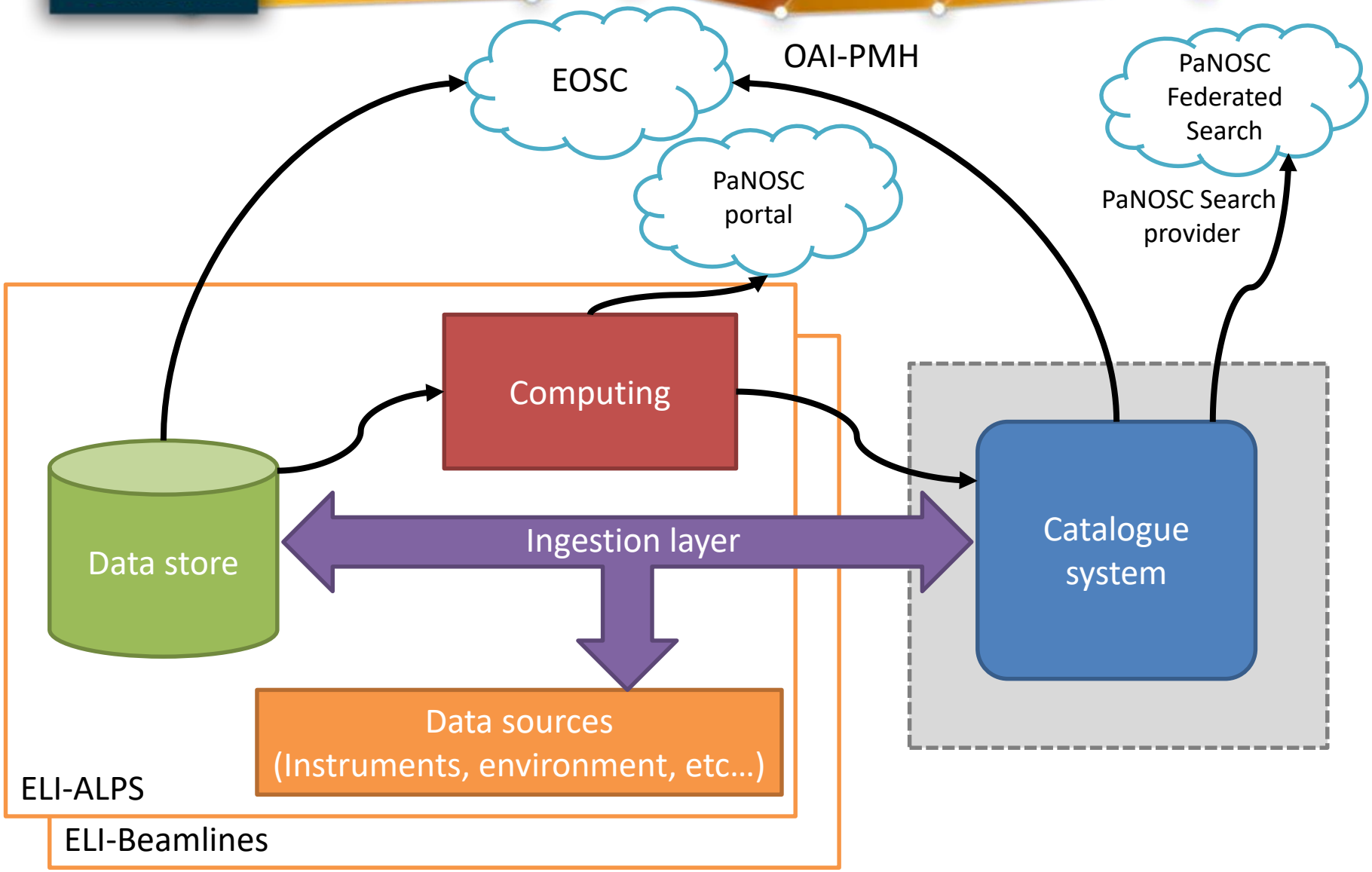
- ELI-ERIC
- Planned data management architecture
- Evaluated catalogue systems
- The main aspects for the evaluation
- Aspects in details
- Summary

- **Distributed research infrastructure**
  - ELI-ALPS, ELI-BL, ELI-NP will join in the future
- **New facilities**
- **Strategies**
  - Shared / common
    - policies
    - services
    - solutions
    - development infrastructure

KPIs - Before PaNOSC - 2018

	ILL	ESRF	CERIC	XFEL	ELI	ESS
Data / yr	200 TB	8 PB	1 PB	3PB	< 1 PB	0
Data Policy	2011	2016	2014 (3/8)	2017	In progress	2017
Metadata catalogue	Local	Icat	Local	myMdC	No	SciCat
Metadata definitions	Nexus	Nexus	custom	myMdC	?	Nexus
DOI	2012	2018	No	2018	No	2018
Open Data	Yes	Yes	No	Yes	No	No
Data Services	Pilot	In progress	Remote	In progress	?	In progress
Common data API	No	No	No	No	No	No
User training	No	No	No	No	No	No

Planned data management architecture



Evaluated catalogue systems

- **iCat**
- **SciCat**
- **Invenio RDM**

- **Other data catalogue systems:**
  - IBM Cloud Pak for Data
  - Oracle Cloud Infrastructure Data Catalog
  - Google Data Catalog
  - Apache Atlas
  - Omero

Main aspects for evaluation

- **Data model**
- **Authentication and authorization**
- **Ingestion**
- **Search**
- **GUI**
- **Deployment and operation**
- **Documentation**

Data model

Sub-aspect \ System	iCat	SciCat	Invenio RDM
Schema	Based on the Core scientific data model	An inhouse developed schema	Extension of the DataCite schema
Support for configurable schema	Some element can have “parameters”	It supports parameters	Yes, with tutorial (it is recommended to use built-in schema)
Investigation life-cycle support	The model has Investigation and Publication, but the Proposal is not a part of the schema.	The model has support for Proposal and Published data and has a DatasetLifecycle element, which contains the executed operations on a Dataset.	Investigation life-cycle could be supported through the resource type property and the referenced identifiers.
Interoperability with other schemas	Dublin Core, DataCite (via OAI-PMH, configuration is facility dependent)	Dublin Core, DataCite (via OAI-PMH)	Dublin Core, DataCite (derived from DataCite)

Authentication and authorization

Sub-aspect \ System	iCat	SciCat	Invenio RDM
Authentication	Simple, LDAP, OpenID Connect	Simple, SAML, LDAP, OAuth/OpenID Connect	Simple, SAML, LDAP, OAuth/OpenID Connect
Authorization	Rule based access contorl	Role based access control	Role based access control
User and group management	Yes	Yes	Yes

Ingestion

Sub-aspect \ System	iCat	SciCat	Invenio RDM
API	RESTful API (Java, Python packages)	RESTful API (openAPI/Swagger)	RESTful API (Python, CLI)
HDF5 and NeXus support	Yes	Yes	Yes
Documentation	Python tutorial	Tutorial (PSI, ESS)	Tutorial for Python and CLI

Search

Sub-aspect \ System	iCat	SciCat	Invenio RDM
Query system	JPQL based (Jakarta/Java Persistence Query Language)	Loopback query (filtering)	Elasticsearch Query DSL (Domain Specific Language)
OAI-PMH support	Yes	Yes	Yes (provider and harvester)
PaNOSC search provider	Yes	Yes	N/A

GUI

Sub-aspect \ System	iCat	SciCat	Invenio RDM
Web based GUI	Topcat	Catanie	Invenio framework
Technologies	AngularJS – EOL Dec. 2021 (a React based solution in development)	Angular	React
Customization	A cookbook	N/A	Tutorials

Deployment and operation

Sub-aspect \ System	iCat	SciCat	Invenio RDM
High availability and scalability	Yes*	Yes (Kubernetes)	Yes (Openshift, Kubernetes)
Containerization	It does not have official docker support.	Docker images	Docker images
Technologies and languages	MySQL, Oracle, MariaDB Glassfish, Payara, Java, Python(2!) EOL 2020	MongoDB,NodeJS,Rabbit MQ, Kafka,Loopback, Javascript	HAProxy, Nginx, Apache UWSGI,Gunicorn, NodeJS, Celery, PostgreSQL, Elasticsearch, RabbitMQ, Redis, Memcache Python
Operation model	Self service Multi-site support	Self service	Self service or commercial model: https://tind.io/ Multi-site support
Documentation	Installation process is not clear.	Docker based installation with CLI based on make	Installation is based on a simple python based CLI tool.

Documentation

Sub-aspect \ System	iCat	SciCat	Invenio RDM
General	The documentation is fragmented.	A well written documentation, the development section builds on the documentation of the Loopback framework	Detailed documentation for both the framework and the RDM application as well.
Tutorials	Developer page: TODOs and “Page not found” errors	Developer guide: Contains tutorials	Developer page: Plenty of tutorials and even more for the main framework
Online support	Google group, Issue tracking on Github	Issue tracking on Github	Discord channel, Issue tracking on Github, LTS, Commercial support

Summary

	ICat	SciCat	Invenio RDM
Pros	• used in PaN community • “PaNOSC ready” • „ERIC” support	• used in PaN community • “PaNOSC ready” • Compatible with modern solutions (containerization, kubernetes)	• Framework based • A customizable turn-key solution • Based on modern and exchangeable technologies/solutions • Good support • Commercial option
Cons	• EOL technologies (Angular JS, Python2) • Small developer community • Fragmented documentation	• Small developer community • „ERIC“ support is not clear	• Under development – LTS in July 2021 • “PaNOSC unready”

Thank you