



Carlo Minotti, Stephan Egli:: Paul Scherrer Institut

PSI Site report

SciCat Meeting Hamburg, July 4th 2022

Overview of Talk

- Current status of SciCat usage and toolset at PSI
- Gap-Analysis
 - of potential general interest
 - ... and probably only relevant for PSI

Some usage statistics

- About 440k datasets and 180 M files, more than 1500 ownerGroups
- Usergroups are from SwissFEL, SLS, Neutron, Proton/Muon beamlines, Electron Microscopy etc..
- Archiving to CSCS (remote HPC site in Lugano) with archive/retrieve workflows, over 10 PB so far











Cloud based deployment

- Installation scripts to setup K8s cluster from scratch (i.e. no hardware, nothing, total costs 200 Euro/month) on Hetzner cloud and prepare infrastructure components (Mongo, Keycloak):

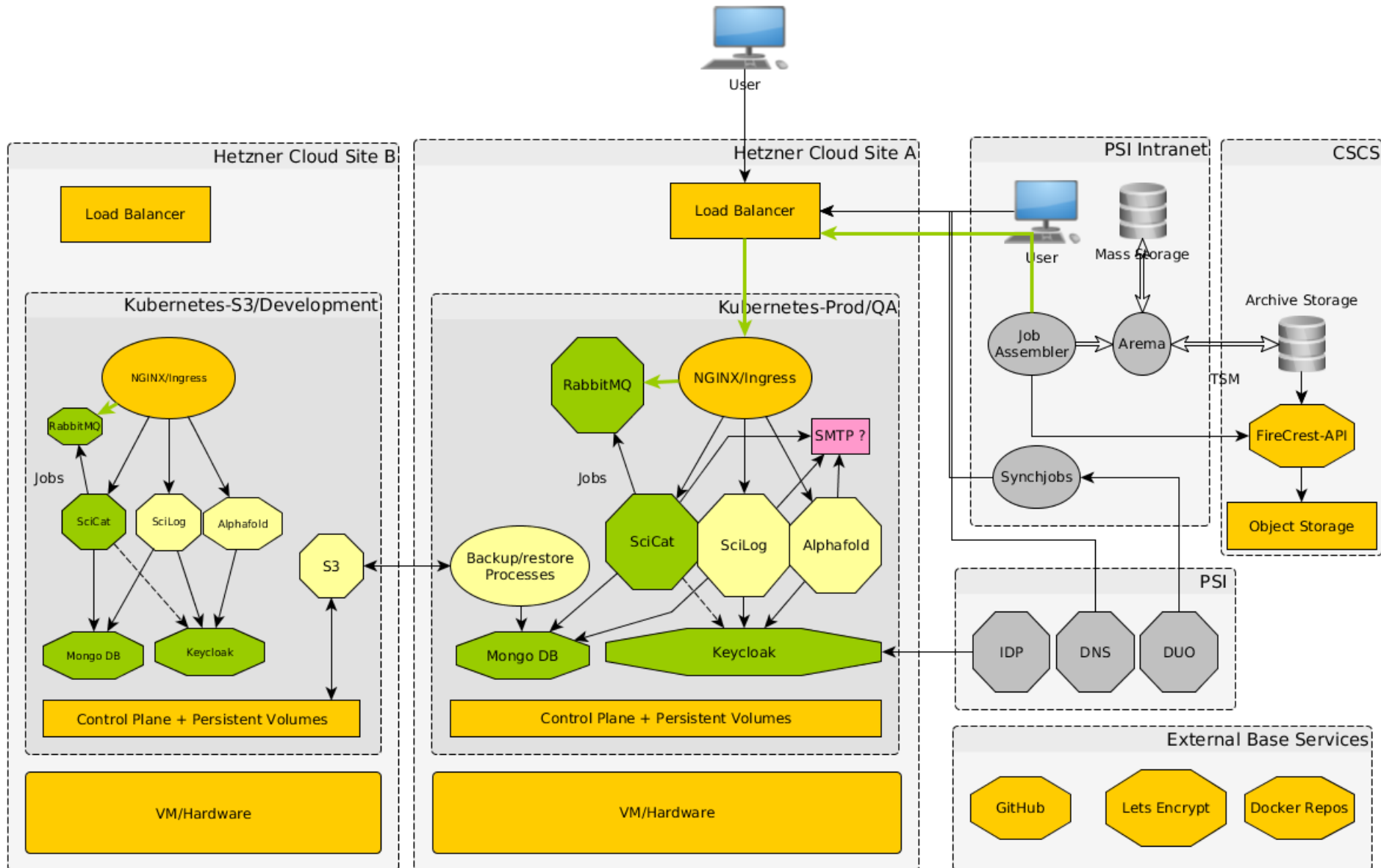
<https://github.com/paulscherrerinstitute/cloudsetup/>

- GitHub CI : action driven installation of SciCat related microservices :

<https://github.com/paulscherrerinstitute/scicat-ci>

 minotti_c and minottic Set runNumber enabled option to true	✓ afd71b0 19 hours ago	🕒 89 commits
 .github/workflows	Add oaipmh sub Set runNumber enabled option to true	last month
 backend @ 4ac55ef	Fix condition of email policy concat on retrieve	2 months ago
 frontend @ 407e35f	Add frontend submodule	4 months ago
 generic_service_chart	Add base64 validity check	last month
 helm_configs	Set runNumber enabled option to true	19 hours ago
 oaipmh @ 24a3a49	Add oaipmh submodule and deploy	last month
 pan-ontologies-api @ f2b555d	Add pan-ontologies-api deploy	4 months ago
 search-api @ 7b2bb69	Add search-api deployment	4 months ago
 gitmodules	Add oaipmh submodule and deploy	last month

Cloud architecture and connections to PSI



Keycloak as IDP Gateway

- Keycloak integration via OIDC.
- PSI IDP is connected via SAML and provides authentication and **authorization** (i.e. ownerGroup values) information

The collage consists of three overlapping screenshots:

- Top Screenshot:** A browser window showing the URL `https://idp.psi.ch/idp/profile/SAML2/POST/SSO?execution=e1s2`. The page header features the PSI logo and the text "PAUL SCHERRER INSTITUT".
- Middle Screenshot:** A browser window showing the Keycloak admin console at `https://kc.psi.ch/auth/admin/master/console/#/realm:`. The breadcrumb navigation is "Identity Providers > PSI Identity Provider > Mappers". The "Mappers" tab is selected for the "PSI Identity Provider". A search bar and a list of mappers (Emailmapper, first name importer, Group Membership, lastname importer, ogroup) are visible.
- Bottom Screenshot:** A "PSI Login" form. It includes a "service now" logo, a "PSI Login" title, and a login form with fields for "Username" (containing "egli") and "Password" (masked with dots). There is a "Login" button and a "Remember me" checkbox. Below the form, there is a checkbox for "Enable auto-login" and a link for "Auto-Login".

Retrieving data from tape



Datasets /

Search

Clear

Text Search

Location

Group

Type

Keywords

Start Date – End Date



+ Add Condition

+ Create Dataset

My Data

All Public Data

All

Archivable

Retrievable

Work In Progress

System Error

User Error

	Name	Run No.	Source Folder	Size	Start Time	Type	Run No.
<input type="checkbox"/>	Archive/TestDataset		...estDataset	14 MB	2022-05-11 Wed 15:02	raw	
<input type="checkbox"/>	add_using_ui		.../nfs	0 B	2022-04-12 Tue 09:28	derived	
<input checked="" type="checkbox"/>	30042021-testingest/normal		...est/normal	101 MB	2020-02-12	base	
<input checked="" type="checkbox"/>	30042021-testingest/normal		...est/normal			base	
<input type="checkbox"/>	S11850-20865_ID46-full		...2000-12999			raw	
<input type="checkbox"/>	S11850-20865_ID46-full		...2000-12999			raw	
<input type="checkbox"/>	S11850-20865_ID46-full		...2000-12999			raw	
<input type="checkbox"/>	S11850-20865_ID46-full		...2000-12999	21 GB	2019-09-27 Fri 13:10	raw	
					2019-09-26		

Really retrieve?

Optionally select destinat...

Ok

No Thanks

- For ease of deployment (standalone binaries) implemented in GO
 - datasetIngestor: Purpose: support the automation of processes, does necessary API calls in the background
 - datasetGetProposal
 - datasetArchiver
 - ...
 - Example syntax:
<https://scicatproject.github.io/documentation/Ingestor/ingestManual.html#sec-3-2-6>
- Qt-GUI «wrapper» application around the CLI tools to ingest and retrieve
 - <https://scicatproject.github.io/documentation/Ingestor/ingestManual.html#sec-4> -> Screenshot
- Web based Metadata Editor with Template support, helps to create a specific dataset document (metadata.json) -> Screenshot
- So far PSI specific, not (yet) open source, but we are happy to share.

Qt GUI wrapper around CLI tools

SciCat Archiver

Username: wakonig_k
Server: QA

Select the pgroup

p17970

Proposal

PI first name	Andreas
PI last name	Menzel
PI email	andreas.menzel@psi.ch
Proposal number	20.500.11935/20190797
Title	Fourier Ptychographic X-ray Computed Tomography
Abstract	X-ray Fourier ptychography uses multiple acquisitions, each comprising complementary frequency content of the sample. This is achieved by moving the objective lens such that the different section of the sample's spectrum is covered for each measurement. Having demonstrated the concept and capabilities of X-ray Fourier ptychography in recent experiments, with this experiment we aim to explore its applicability for tomography.

Retrieve data

Archive data

Export

```

[ ] ▼ object {11}
[ ]     principalInvestigator : albrecht.gessler@psi.ch
[ ]     creationLocation : /PSI/EMF/JEOL2200FS
[ ]     dataFormat : TIFF+LZW Image Stack
[ ]     sourceFolder : /data/project/bio/gessler/myimages
[ ]     datasetName : myimages
[ ]     owner : Wilhelm Tell
[ ]     ownerEmail : wilhelm.tell@psi.ch
[ ]     type : raw
[ ]     description : EM micrographs of amygdalin
[ ]     ownerGroup : a-12345
[ ] ▼ scientificMetadata {3}
[ ]     ▼ sample {3}
[ ]         name : Amygdalin beta-glucosidase 1
[ ]         uniprot : P29259
[ ]         species : Apple
[ ]     ▼ dataCollection {1}
[ ]         date : 2018-08-01
[ ]     ▼ microscopeParameters {3}
[ ]         ▼ pixel size {2}
[ ]             v : 0.885
[ ]             u : Å
[ ]         ▼ voltage {2}
[ ]             v : 200
[ ]             u : kV
[ ]         ▼ dosePerFrame {2}
[ ]             v : 1.277
[ ]             u : e/Å²

```

«Gap analysis» (of potential general interest)

What we have to do in any case :

- Testing the new backend and bring it into production
 - How to keep track of changes in loopback/nestjs during migration period ?
- Implement auto-publishing workflows after embargo period
 - How is the mapping between datasets and publishedData defined: per proposal ? User steerable ?

«Gap analysis» (of potential general interest)

What we would like to have in addition

- Improved dataset table: more user steerable options to steer what is shown and what can be filtered/facet searched. Search by PID
- Easy creation of usage statistic plots (from Jupyter notebooks or GUI or some external tool ?)
- Schema support (Linus proposals) for more control of what data is ingested
- Better fulltext search (e.g. arbitrary substring support, scoring)
- Landing page display: integrate into SciCat (Published data display)
 - Portal like homepage with search, e.g. like <https://opendata.cern.ch/> ?
- One Helm chart for full Scicat installation (including optional Mongo DB)
 - We have helm charts for each microservice separately (backend, frontend, search...)
- Make documentation up-to-date
- Query performance for very large databases if filename based search is needed

Plans likely PSI specific

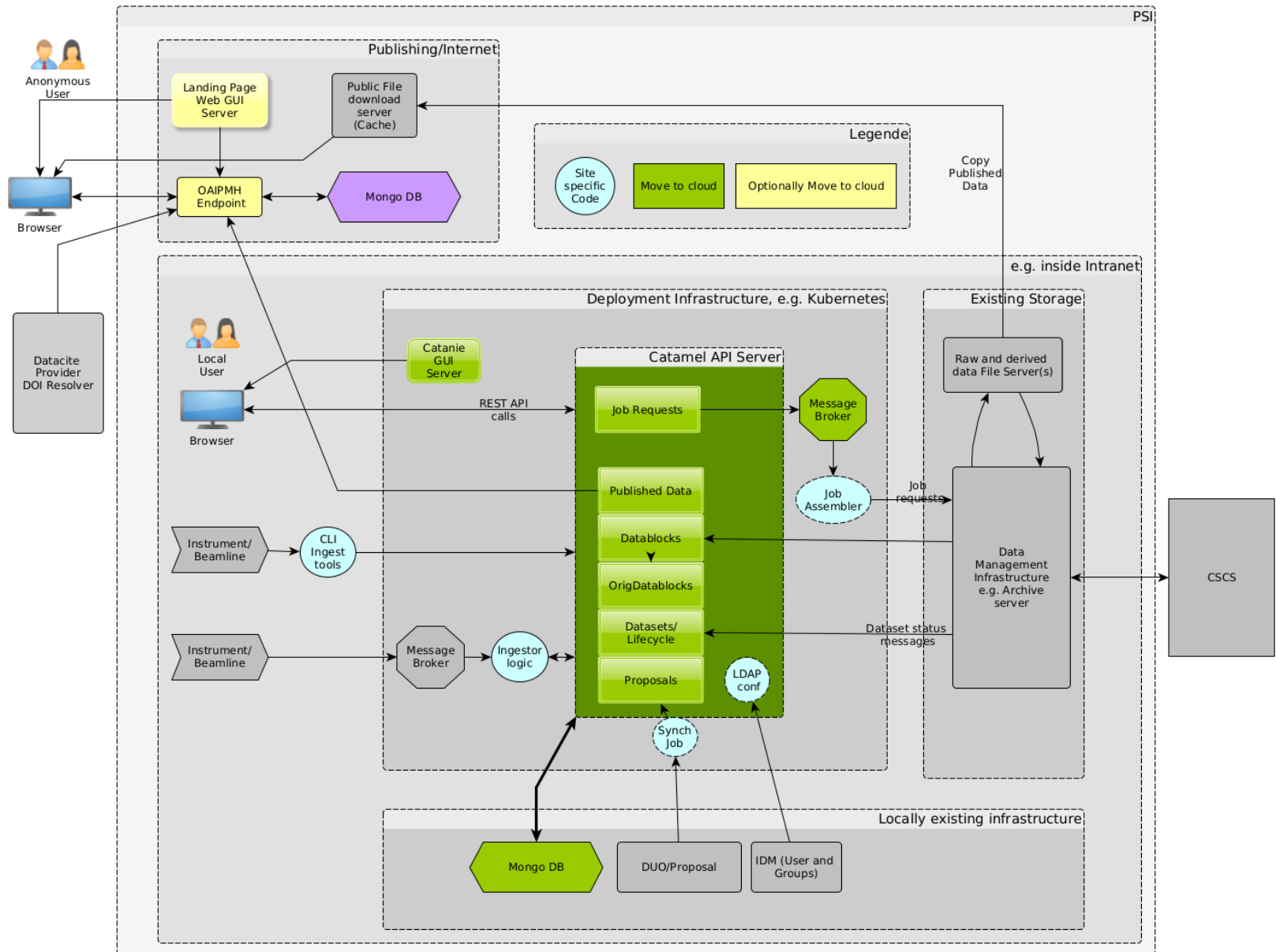
- Offer data retrieval directly from CSCS: modified Job triggered node red workflow
- Expands/Panosc related TODOs
 - Add techniques information to metadata, involve scientists,
 - add scoring engine
- Move OAIPMH and Landingpage server to the cloud
- Monitoring tools for our installation, Logging etc
- Extend **existing** data to add SI units information
- Use of SciCat for other customers within ETH domain: ETH-ORD program



Additional Material



SciCat status within PSI before cloud migration



Namespaces in K8S cluster



Alle Namespaces ▾



Suchen

Cluster > Namespaces

Namespace

Name	Labels	Phase
scilog-qa	kubernetes.io/metadata.name: scilog-qa	Active
zammad	kubernetes.io/metadata.name: zammad	Active
ingress-nginx	kubernetes.io/metadata.name: ingress-nginx name: ingress-nginx	Active
scicat-production	kubernetes.io/metadata.name: scicat-production	Active
scicat-qa	kubernetes.io/metadata.name: scicat-qa	Active
rabbitmq	kubernetes.io/metadata.name: rabbitmq name: rabbitmq	Active
rabbitmq-qa	kubernetes.io/metadata.name: rabbitmq-qa name: rabbitmq-qa	Active
cert-manager	kubernetes.io/metadata.name: cert-manager name: cert-manager	Active
my-keycloak-operator	kubernetes.io/metadata.name: my-keycloak-operator olm.operatorgroup.uid/185c387e-a462-421e-91bc-6441b666596c	Active
mongo	kubernetes.io/metadata.name: mongo	Active

Services inside SciCat production

kubernetes			scicat-production ▾	Suchen
Discovery and Load Balancing > Services				
Services				
Name	Labels	Typ		
● backend	app.kubernetes.io/instance: backend app.kubernetes.io/managed-by: Helm app.kubernetes.io/name: generic-service-chart Alles anzeigen	ClusterIP		
● frontend	app.kubernetes.io/instance: frontend app.kubernetes.io/managed-by: Helm app.kubernetes.io/name: generic-service-chart Alles anzeigen	ClusterIP		
● search-api	app.kubernetes.io/instance: search-api app.kubernetes.io/managed-by: Helm app.kubernetes.io/name: generic-service-chart Alles anzeigen	ClusterIP		
● pan-ontologies-api	app.kubernetes.io/instance: pan-ontologies-api app.kubernetes.io/managed-by: Helm app.kubernetes.io/name: generic-service-chart Alles anzeigen	ClusterIP		