PAUL SCHERRER INSTITUT

Stephan Egli:: Paul Scherrer Institut

SciCat: History and Status

SciCat Meeting Hamburg, July 6th 2022

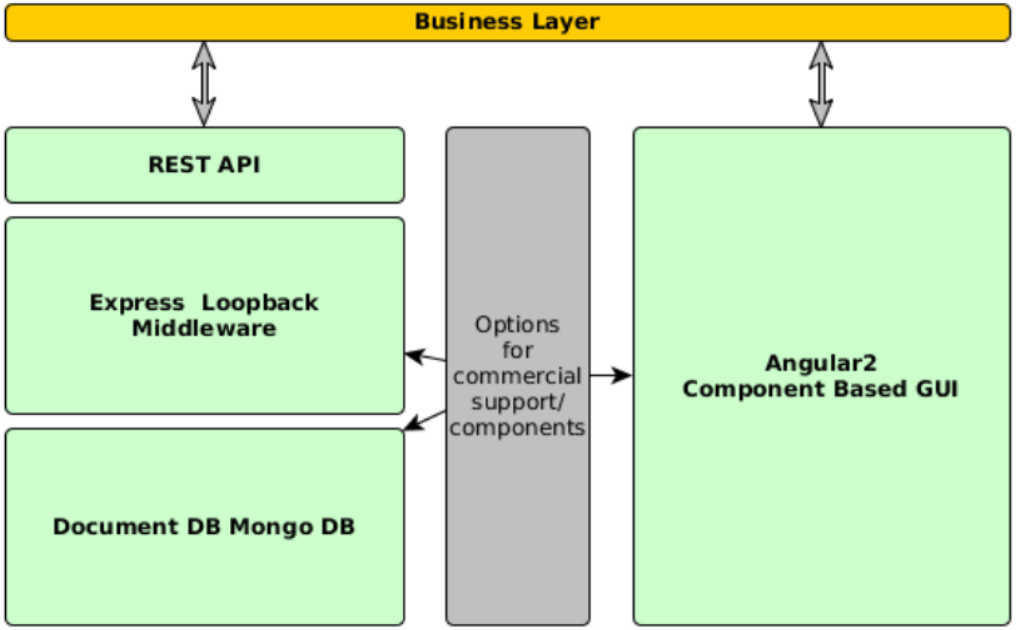
History: continuous evolution over > 5 years

- 2015-12 Initial data catalog requirement definition within DaaS project
- 2016-08-26 Decision for new development based on prototype demo
- 2016-09-16 ESS-PSI contract for «Data Curation» project signed
- 2016-12 First Kubernetes installation
- **2017-06-22:** **First PSI-ESS meeting** on “Melanie” (Mongo-Express-Loopback-Angular Integrated Enviroment)
- 2017-08 Initial release based on Loopback 3 and Angular 2
- 2017-10-06 Baptize as “SciCat”, define Logo
- 2017-10-08 Move code from PSI repos to GitHub
- 2018-01 Start of first data ingests at PSI
- **2018..2021** **Main development period during Data Curation project**
- 2021-04 Release with support for publication workflow and OAIPMH
- 2021-07 Start of evaluating **nestjs** as loopback 3 replacement
- 2021-11-26 Rename catanie to **frontend** and catamel to **backend**
- 2022-01 adding pyscicat and OIDC support
- 2022-02-14 Frontend: replace the build time configs with runtime configs
- 2022-09-01 Planned switch to new backend

Historic Slide: Architecture

MELANIE Architecture

- Idea: Do profit from the work of others, stand on shoulder of giants
- Use Mongo-Express-Loopback-Angular2-Integrated technology stack
  - Use Document Databases as basis
  - Use a standardized REST API ("API first" principle)
- Simple and modular two-pillar architecture with well defined interfaces



Historic Slide: Prototype

Show Metadata for selected Dataset

Recently created datasets for pgroup p12532:

Folder	Created	Size	Define ACIA	Status	Lifetime on disk
/home/amor/data/2015 /000/amor2015n000001.hdf	2015-04-28 09:47:50	113820	archive, 2 tapes, sign, encrypt	Stop Archiving	
/home/amor/data/2015 /000/amor2015n000004.hdf	2015-04-28 10:34:10	113820	archive, 2 tapes, sign, encrypt	Stop Archiving	
/home/amor/data/2015 /000/amor2015n000006.hdf	2015-04-28 10:35:58	113820	archive, 2 tapes, sign, encrypt	To Archive	
/home/amor/data/2015 /000/amor2015n000008.hdf	2015-04-28 10:37:27	113820	archive, 2 tapes, sign, encrypt	To Archive	

1 2 3 4 5

Dataset 579da05e0024ef0b6b885425

Basic data | Sample/Probe | Beam/Source | Detector | Scan | Images


```
{
  "id": "579da05e0024ef0b6b885425",
  "createdAt": "2016-07-31T06:53:18.388Z",
  "updatedAt": "2016-08-15T13:36:43.022Z",
  "License": "CC BY-SA 4.0",
  "file_time": "2015-04-28 09:47:50",
  "owner_fax_number": "UNKNOWN",
  "owner_address": "PSI",
  "instrument": "AMOR",
  "owner": "J. Stahn",
  "entry1": {
    "comment": "I feel uncommented",
    "single_detector_1": {},
    "title": "Referenz 2014_12",
    "area_detector": {},
    "amor_mode": "horizontal",
    "user": {
```

Goals

- Fill a very specific gap: provide a **metadata** catalog for **scientific** data to support the processes of **data management** (therefore supporting FAIRness rules)
- **For scientists**
  - Support all scientific areas with **flexible** data format.
  - Make it easy to **add** data to the catalog and **find** data back
  - Support dataset **lifecycle** management (data handling and metadata edit history)
  - Support to **publish** and **link** the raw data leading to publications
  - Support **sharing** of data for multidisciplinary research
  - Allow users to **extend** the system by their own tools via well defined API
- **For the operators/implementers**
  - Options to integrate with existing storage infrastructure, archive systems, digital user interface and user identity systems
  - Non-goals:
    - replace or modify existing disk storage infrastructure or archive systems
    - replace or modify existing proposal systems or user office systems
  - Make deployment easy and operation stable (micro-services)

Core Features (User relevant)

- Long-term storage for **metadata** of scientific data in form of **datasets** with unique **PID**
- **Flexible format**: allow principal investigator, beamline scientists, instrument responsible to choose the metadata they need (no developer or operator help needed)
- Define **raw** datasets and **derived** datasets including their relation
- Link datasets to **proposals** and **samples**
- Optional adding of **keywords** and **attachments**
- **Query** on all provided metadata fields to find the data stored in the system
- Link filename lists to datasets inside datablocks (no built in size limitations)
- Automated edit history, **lifecycle** book keeping
- Automated creation of **SI units** from given units
- Define **published data** including its metadata (author, abstract , title…) by combining datasets and assigning a unique **DOI**
- **Job** system to **interface with existing external storage infrastructure** (e.g. for archive and retrieve jobs)
- Allow ad-hoc **sharing** of data via email addresses

Start screen: Searchable list of datasets

Datasets /

Items per page: 25 1 – 25 of 330

Search Clear

Text Search

Location

/PSI/SmuS/piE1/muX | 1

/PSI/SWISSFEL/ARAMIS-ALV...

/PSI/SMUS/LEM | 86

/PSI/SLS/cSAXS | 9

/PSI/SLS/TOMCAT | 22

/PSI/SLS/SIM | 2

+ Add Condition

Name	Source Folder	Size	Start Time	Type	Proposal ID	Group	Data Status
/nemu	.../nemu	43 MB	2022-05-23 Mon 13:47	raw	0.11935/202(	p19064	retrievable
data_nexus/S08000-08999	...8000-08999	31 GB	2022-04-19 Tue 20:54	raw	00.11935/p1	p17880	retrievable
data_nexus/S07000-07999	...7000-07999	52 GB	2022-04-19 Tue 20:24	raw	00.11935/p1	p17880	retrievable
data_nexus/S06000-06999	...6000-06999	55 GB	2022-04-19 Tue 18:16	raw	00.11935/p1	p17880	retrievable
data_nexus/S05000-05999	...5000-05999	56 GB	2022-04-19 Tue 18:08	raw	00.11935/p1	p17880	retrievable
data_nexus/S04000-04999	...4000-04999	60 GB	2022-04-19 Tue 17:53	raw	00.11935/p1	p17880	retrievable
data_nexus/S03000-03999	...3000-03999	31 GB	2022-04-19 Tue 17:38	raw	00.11935/p1	p17880	retrievable
data_nexus/S02000-02999	...2000-02999	31 GB	2022-04-19 Tue 16:50	raw	00.11935/p1	p17880	retrievable
data_nexus/S01000-01999	...1000-01999	31 GB	2022-04-19 Tue 16:42	raw	00.11935/p1	p17880	retrievable

Published datasets before login

Your (embargoed) datasets after login

Example dataset definition

- Metadata is stored in human and computer readable JSON format.
- Administrative and scientific metadata, see screenshots:

Example administrative metadata

Datasets / 20.500.11935/975130e6-6697-4709-a785-51be31b90ce5 /

Details | Datafiles | Attachments | Lifecycle

General Information

Name	Type3_03_
PID	20.500.11935/975130e6-6697-4709-a785-51be31b90ce5
Type	raw
Creation Time	2022-06-12 22:57
Keywords	Add Keyword +

Creator Information

Owner	Christian Schlepuetz
Owner Group	p15741
Access Groups	slstomcat

File Information

Source Folder	/sls/X02DA/Data10/e15741/20220612_Heymann_PrintedNozzles/Type3_03_
Size	14 GB
Data Format	Tomcat pre 2017

Related Documents

Proposal	
Creation Location	/PSI/SLS/TOMCAT

Scientific Metadata

Example scientific meta data

Scientific Metadata


```
▾ beamlineParameters:
  ▸ Beam energy: Object {"u":"keV","v":11.999,"valueSI":1.9224516603435e-15,"unitSI":"(kg m^2) / s^2"}
    FE-Filter: "No Filter 100%"
    Monostripe: "Ru/C"
    OP-Filter1: "No Filter"
    OP-Filter2: "No Filter"
    OP-Filter3: "No Filter"
  ▸ Ring current: Object {"u":"mA","v":400.925,"valueSI":0.40092500000000003,"unitSI":"A"}
▾ detectorParameters:
  ▸ Actual pixel size: Object {"u":"um","v":0.33,"valueSI":3.3e-7,"unitSI":"m"}
    Camera: "PCO.Edge 5.5"
  ▸ Delay time: Object {"u":"ms","v":0,"valueSI":0,"unitSI":"s"}
  ▸ Exposure time: Object {"u":"ms","v":200,"valueSI":0.2,"unitSI":"s"}
    Microscope: "Opt.Peter MB op"
  ▸ Microscope x position: Object {"u":"mm","v":-182.23,"valueSI":-0.18223,"unitSI":"m"}
  ▸ Microscope y position: Object {"u":"mm","v":-167.86,"valueSI":-0.16786,"unitSI":"m"}
  ▸ Microscope z position: Object {"u":"mm","v":93.4,"valueSI":0.09340000000000001,"unitSI":"m"}
    Millisecond shutter: "not used"
    Objective: 20
    Scintillator: "LuAg:Ce 20um (C20-79)"
    X-ROI End: 2560
    X-ROI Start: 1
    Y-ROI End: 1761
    Y-ROI Start: 400
▾ scanParameters:
  ▸ Angular step: Object {"u":"deg","v":0.09,"valueSI":0.0015707963267948964,"unitSI":"rad"}
    File Prefix: "Type3_03_"
```

Lifecycle information


```
▾ history:
  ▾ 0:
      id: "a879f3ab-074f-48ad-b120-74f8324c23d5"
    ▾ datasetlifecycle:
      ▾ currentValue:
          archivable: "false"
          retrievable: "false"
          archiveStatusMessage: "started"
      ▾ previousValue:
          archivable: false
          retrievable: false
          publishable: false
          archiveRetentionTime: "2032-06-12T00:00:00.000Z"
          dateOfPublishing: "2025-06-12T00:00:00.000Z"
          isOnCentralDisk: true
          archiveStatusMessage: "scheduledForArchiving"
          retrieveStatusMessage: ""
          retrieveIntegrityCheck: false
      updatedBy: "archiveManager"
      updatedAt: "2022-06-16T16:26:36.725Z"
  ▾ 1:
      id: "5ae24109-4b26-48dd-b718-fc3cc4967047"
    ▾ datasetlifecycle:
      ▾ currentValue:
          archivable: "false"
          retrievable: "true"
          archiveStatusMessage: "datasetOnArchiveDisk"
      ▾ previousValue:
```

Filelisting, search for filenames

Files /

Hornby

Items per page: 10 1 – 10 of 2684

Filename	Size	Created at	U I D	G I D	Owner Group
contains	greaterThan	Start date – End date	contains	contains	contains
HDF2Tif.py	1952	Dec 17, 2018, 10:01:56 AM	marone	p15869	p11218
HDF2TifSelectedTimeStep.py	2551	Nov 20, 2019, 4:51:40 PM	marone	p15869	p11218
Jetraw	4096	Apr 25, 2022, 7:09:21 PM	marone	p15869	p11218
Jetraw/JetrawWithDPCore-21.05.26.6	4096	Aug 12, 2021, 5:30:39 PM	marone	p15869	p11218
Jetraw/JetrawWithDPCore-21.05.26.6/bin	4096	Jun 17, 2021, 6:46:30 PM	marone	p15869	p11218
Jetraw/JetrawWithDPCore-21.05.26.6/bin/dpcore	1221440	Jun 17, 2021, 6:46:29 PM	marone	p15869	p11218
Jetraw/JetrawWithDPCore-21.05.26.6/bin/jetraw	665240	Jun 17, 2021, 6:45:44 PM	marone	p15869	p11218
Jetraw/JetrawWithDPCore-21.05.26.6/include	4096	Jun 17, 2021, 6:46:30 PM	marone	p15869	p11218
Jetraw/JetrawWithDPCore-21.05.26.6/include/dpcore	4096	Jun 17, 2021, 6:46:30 PM	marone	p15869	p11218
Jetraw/JetrawWithDPCore-21.05.26.6/include/dpcore/dpcore.h	4081	Mar 23, 2021, 9:56:26 AM	marone	p15869	p11218

Publishing data

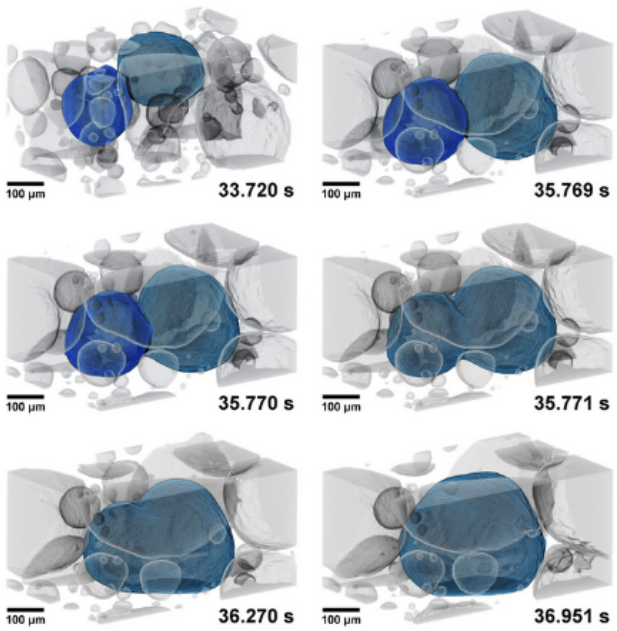
Published Datasets / **10.16907/d7582cb6-7850-42bc-ad76-e845b998e9ca** /

Publication Status

Status	registered
Registered Time	2021-08-09, 10:28

General Information

Title	Tomoscopy: Time-resolved tomography for dynamic processes in materials
Abstract	The article accompanying this dataset gives a brief overview of the recent developments of time-resolved X-ray tomography that have led to what we now call "tomoscopy". A novel setup is presented and applied that pushes temporal resolution down to just 1 ms, i.e. 1000 tomograms per second ('tps') are acquired, while maintaining spatial resolutions of micrometres and running experiments for minutes without interruption. Applications recorded at different acquisition rates ranging from 50 to 1000 tps are presented. We observe and quantify the immiscible hypermonotectic reaction of AlBi10 (in wt%) alloy and dendrite evolution in AlGe10 (in wt%) casting alloy during fast solidification. We analyse the combustion process and the evolution of the constituents in a burning sparkler. Finally, we follow the structure and density of two metal foams over a long period of time and derive details of bubble formation and bubble ageing including quantitative analyses of bubble parameters with millisecond temporal resolution.
DOI	10.16907/d7582cb6-7850-42bc-ad76-e845b998e9ca
URL	doi.psi.ch/detail/10.16907/d7582cb6-7850-42bc-ad76-e845b998e9ca
Publication Year	2021



Publishing data continued…

Creator Information

Creator	Francisco Garcia-Moreno,Paul Hans Kamm,Tillmann Robert Neu,Felix Bülk,Mike Andreas Noack,Mareike Wegener,Nadine von der Eltz,Christian Matthias Schlepütz,Marco Stampanoni,John Banhart
Publisher	PSI

File information

Download Link	https://doi2.psi.ch/datasets/
Resource Type	derived
Data Description	The data collection contains the recorded (raw) projection images of the continuous measurements and selected reconstructions. Raw data are stored in the scientific data exchange schema for hdf5 files (De Carlo et. al, 2014), while reconstructions are stored as tiff stacks. The data set for the solidifying AlBi10 sample (50 tps) contains horizontal and vertical volume sections over time, as well as a selection of 100 reconstructions in full temporal resolution as individual *.tif stacks. The data set for the solidifying AlGe10 sample (200 tps) also contains previously mentioned overview sections, as well as 300 selected reconstructions in full temporal resolution. The data set for the burning sparkler (400 tps) consists of the horizontal sections over time, as well as 500 selected reconstructions in full temporal resolution. The dataset for the foaming AlSi6Cu4 sample (650 tps) consists of the horizontal slices over time and every 1000th reconstruction. The data set for the AlSi8Mg4 foam (1000 tps) consists of the horizontal slices over time, as well as 1000 selected reconstructions in full temporal resolution. Data were collected and processed at the TOMCAT beamline X02DA of the Swiss Light Source.

Related Documents

Related Publications	F. García-Moreno, P. H. Kamm, T. R. Neu, F. Bülk, M. A. Noack, M. Wegener, N. von der Eltz, C. M. Schlepütz, M. Stampanoni, and J. Banhart, "Tomoscopy: Time-resolved tomography for dynamic processes in materials", Advanced Materials 23 September 2021 (online), DOI: https://doi.org/10.1002/adma.202104659
Dataset IDs	20.500.11935/ebff0e97-8c8f-4380-bb4c-473a49975bc3,20.500.11935/e73df1b8-6441-44cd-98b3-3ea2ce037c5b,20.500.11935/24321367-ac96-4ec2-9359-d1d64f5934fa,20.500.11935/577b841c-b56b-4d84-b327-bda561d4a5cf,20.500.11935/7238b5a8-62c2-42f0-9d6b-c46b773e209d,20.500.11935/f52afae9-06eb-493c-ac2b-2ae225fd6816,20.500.11935/1ad13757-20b9-4567-9464-2aa8b23cd362,20.500.11935/9ecbe2a8-098e-4689-b588-67c213c6357e,20.500.11935/6285a066-8d02-4b7e-94f6-4f2efcdcb8af,20.500.11935/1c6e20f4-0f49-473e-89c8-524c1a3d295a

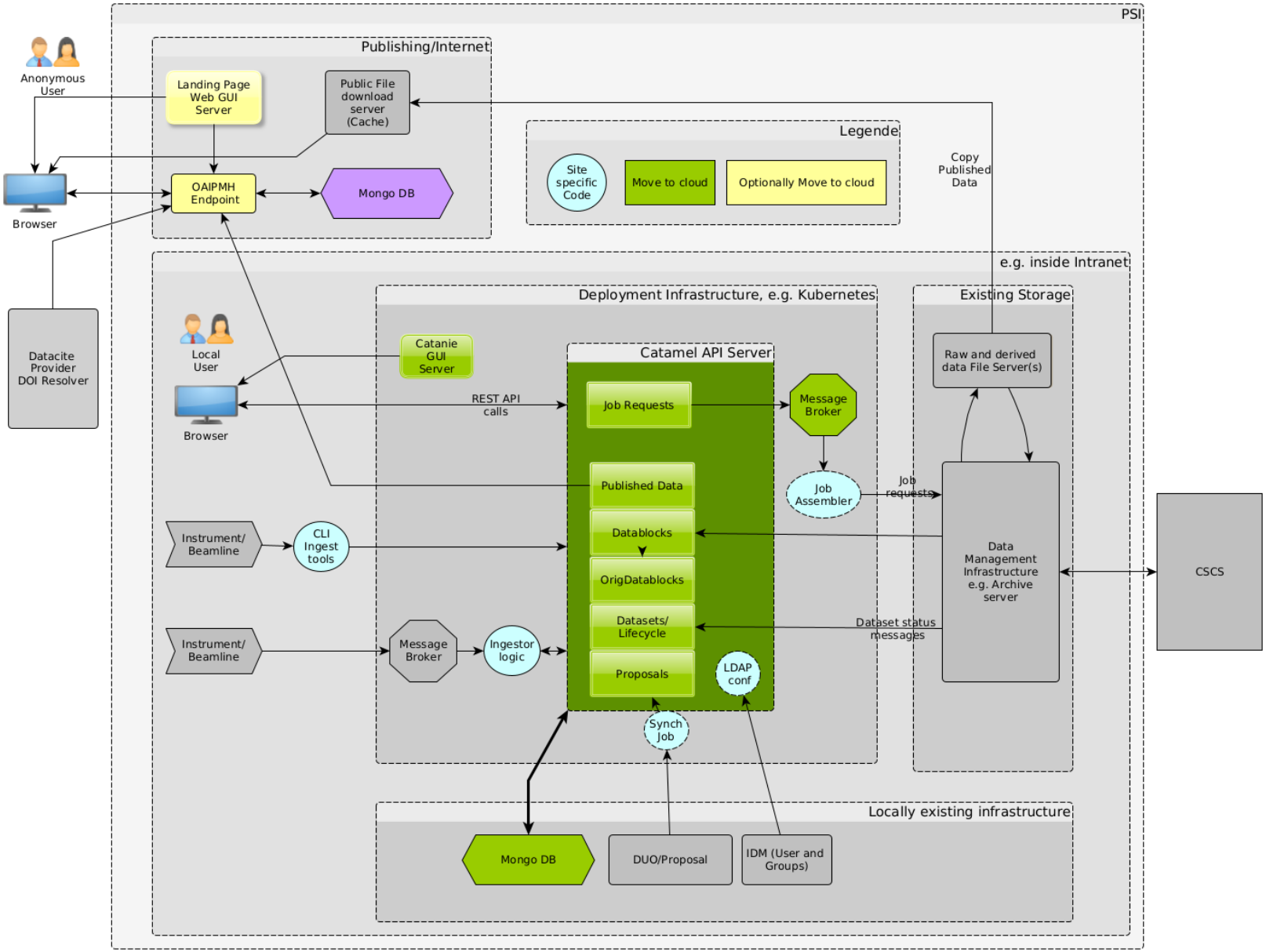
Customizable GUI features

- Allows to make GUI adjust to site specific needs:
  - Enable archive/retrieve workflow
  - Enable direct download options
  - Allow for editing of datasets and or sample data
  - Configure columns to be presented in dataset tables
  - dataset detail display options, display of raw json data
  - Etc.

Technical concepts

- **API first** concept: All functionality via the SciCat API . (Multiple) graphical UIs and CLIs can be developed independently .
  - REST based API (currently based on Loopback 3, migration to NESTJS ongoing)
  - Language agnostic interfacing (Python pyscicat, JS, Go…)
  - Dedicated or auto-generated language specific **SDKs**
- Storage in **Document Databases** (MongoDB): flexible format, full query power
- **User and access** handling managed by linking to existing LDAP systems (AD, OpenLdap) or to identity providers (OIDC, Keycloak)
- Modern **web framework** (Angular >=10) for GUIs
- **Micro-service** architecture for container based deployments (Kubernetes)
- SciCat Core: «**passive**» system: data must be pushed: task of «external» **ingest** tools, with various levels of automation. Therefore gain flexibility in connecting existing storage systems.

Technical components around SciCat API



Job definition: separate “what” from “how”

User / **Jobs** /

global word-search

Items per page: 5

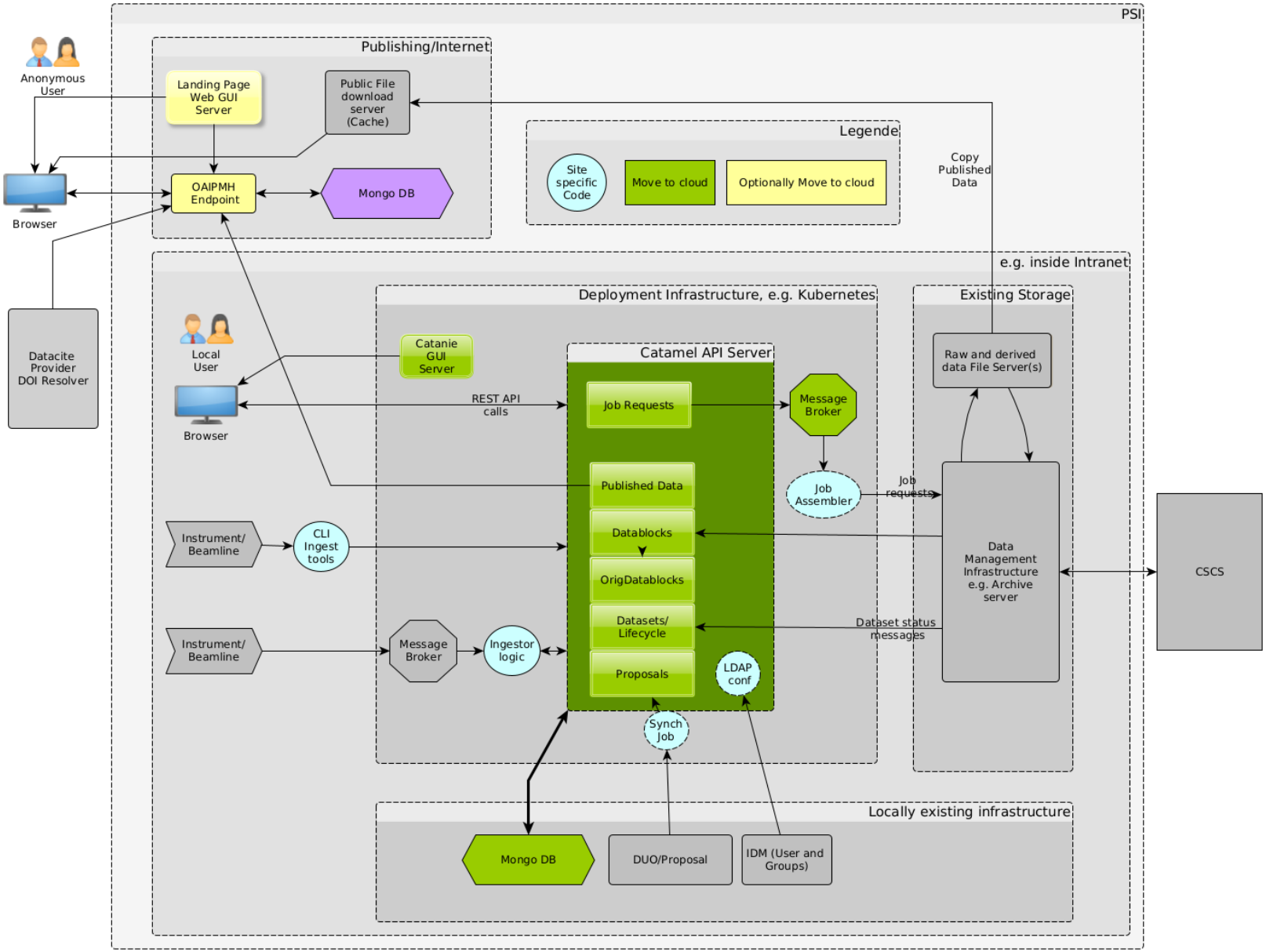
	ID	Initiator	Type	Created at local time	Parameters	Status
	contains	egli	retrieve	Start date – End date		contains
	f20011cc-df22-4001-94cf-33d7b1759412	egli@loopback.ms-ad.com	retrieve	Oct 30, 2018, 8:22:32 AM	{ "username": "ms-ad.egli", "tapeCopies": "one", "destinationPath": "/archive/retrieve" }	

Optional additional services

SciCat as crystallization point for ecosystem of expanding service infrastructure

- Landing page server for published data
- OAIPMH Server to publish data to the worldwide DOI system
- File download service via zip-service
- Link service to chat rooms (scichat)
- Federated Search API (from PaNOSC/EXPANDS project)
- Scoring service for search results
- PaNET API (experiment technology ontologies) (from PaNOSC/EXPANDS project)

Integration of Optional Services at PSI (2021)



The scicatproject codebase on GitHub

https://github.com/orgs/SciCatProject/repositories?type=all

Pull requests Issues Marketplace Explore

SciCat Project

Overview | Repositories 23 | Projects 1 | Packages | Teams 1 | People 15 | Settings

Find a repository… | Type | Language | Sort | New repository

frontend Public
SciCat open data catalogue web client
data-catalog data-management metadata-catalog doi
TypeScript BSD-3-Clause 13 12 27 2 Updated 3 days ago

pyscicat Public
Forked from als-computing/pyscicat
Python 7 0 10 3 Updated 3 days ago

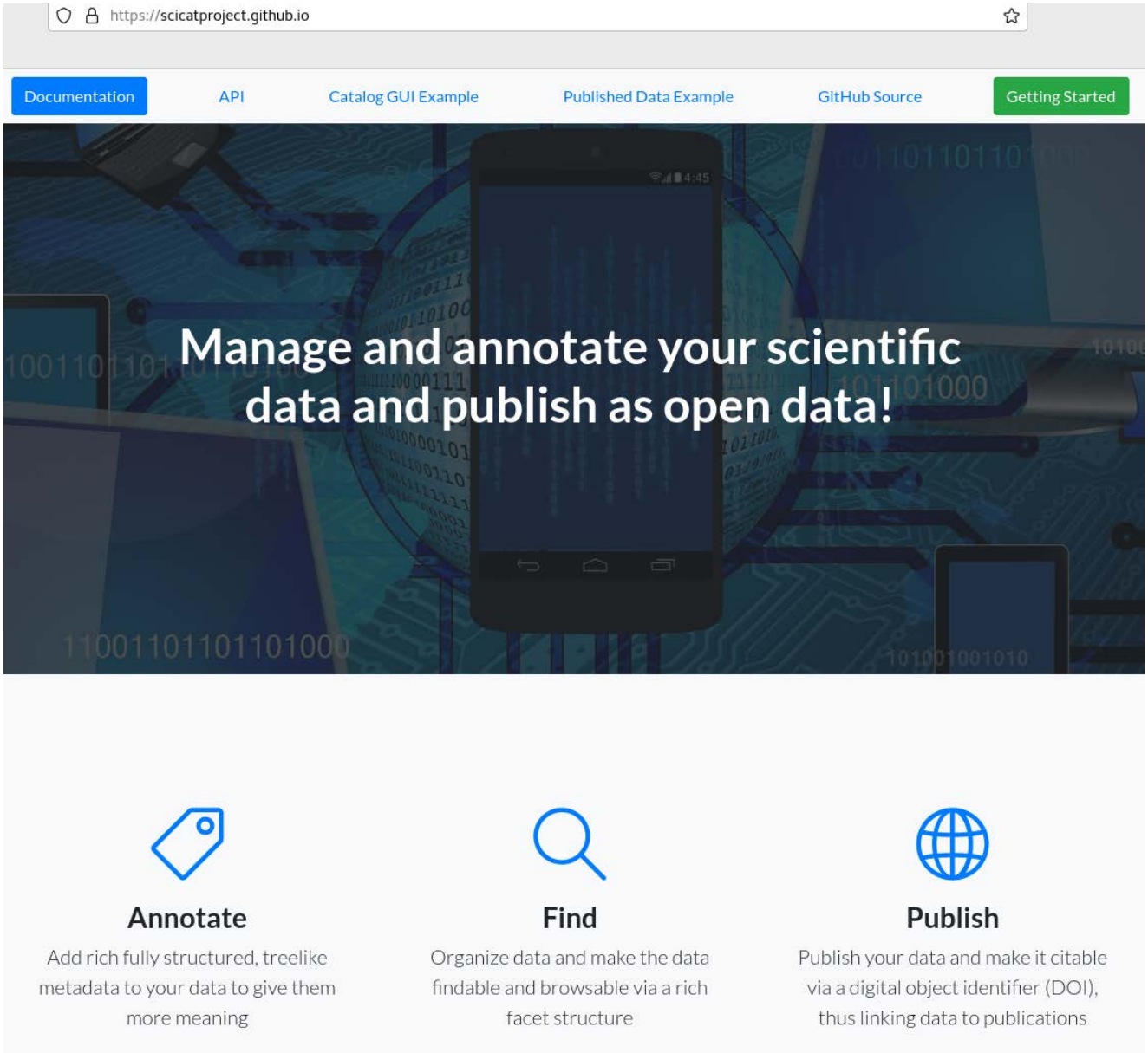
backend Public
SciCat Data Catalogue Backend
kafka rabbitmq swagger dataset loopback ingestor data-catalog-backend
JavaScript BSD-3-Clause 12 10 24 1 Updated 4 days ago

zip-service Public
Service for zipping and downloading file bundles
TypeScript 2 1 1 1 Updated 4 days ago

scicat-backend-next Public
SciCat Data Catalogue Backend
TypeScript BSD-3-Clause 2 2 2 1 Updated 4 days ago

panosc-search-api Public
SciCat integrated search API enabling federated searches for research facilities within the PaNOSC project
JavaScript BSD-2-Clause 1 0 0 0 Updated 4 days ago

Documentation: https://scicatproject.github.io



Supporting Institutions and Projects

- Main contributing institutions so far
  - ESS
  - MaxIV
  - PSI
- Supporting projects
  - *Data analysis as a service* project (swissuniversities)
  - PSI-ESS *Data curation* project
  - EU projects *EXPANDS/PANOSC*
  - Potential upcoming support within *ETH-ORD* (open research data) program

The people make the difference - Thanks to all contributors !

In order of “appearance”:

- Tobias Richter
- Frederik Bolmsten
- Gareth Murphy
- Chris Gwilliams
- Luke Gorman
- Hannes Petri
- Henrik Johansson
- Linh Nguyen
- Marco Leorato
- Laura Shemilt
- Dylan McReynolds
- Max Novelli
- Linus Pithan
- Anastasiia Pylypenko

It is a big pleasure for me to work in such a lively community !