

Repositories for Cancer Imaging

Ignacio Blanquer

Universitat Politècnica de València

Institute of Instrumentation for Molecular Imaging



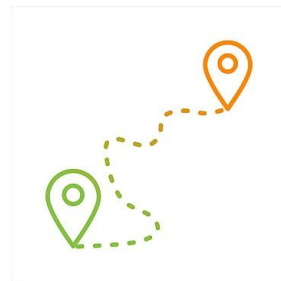
Instituto de Instrumentación
para Imagen Molecular



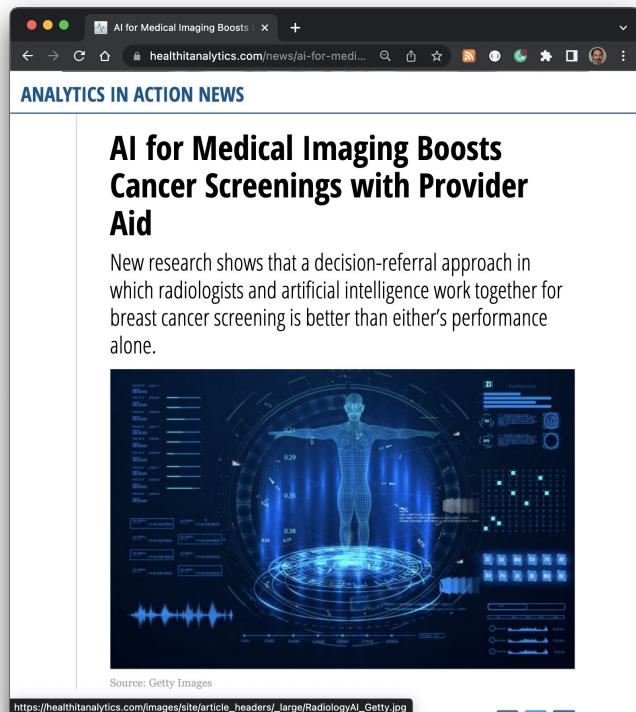
UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Outline

- Motivation and challenges
- The PRIMAGE and CHAIMELEON projects
- The AI4HI Network
- Towards a pan-European Federation: EUCAIM

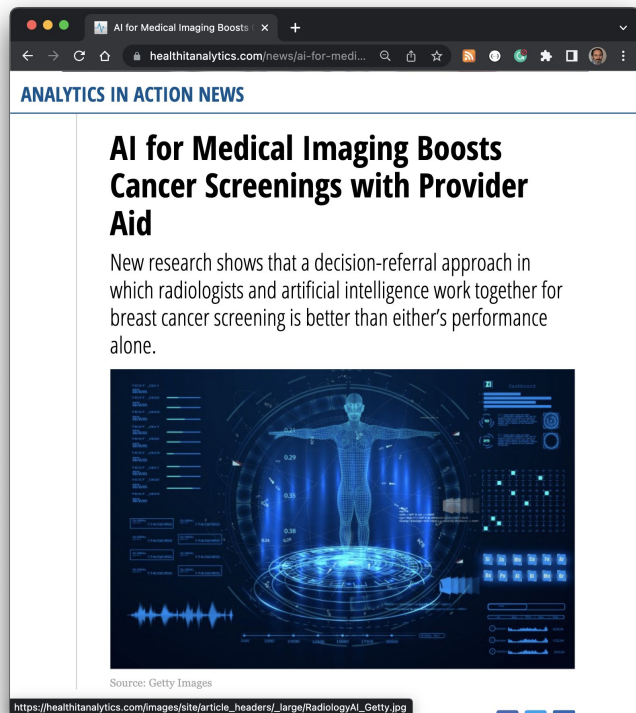


Motivation



- A success story of an AI model trained with 1,193,197 Digital Mammographies
 - In the Valencian Screening program, ~750.000 women are cited each two years.
 - It will take less than 4 years to create such a dataset.

Motivation



<https://healthitanalytics.com/news/ai-for-medical-imaging-boosts-cancer-screenings-with-provider-aid>

- A success story of an AI model trained with 1,193,197 Digital Mammographies
 - In the Valencian Screening program, ~750.000 women are cited each two years.
 - It will take less than 4 years to create such a dataset.
- DIPG is one of the deadliest brain tumours in children
 - With 1,5 cases per 100.000 inhabitants per year.
 - We would need 170.000 years.
- Cancer imaging datasets can be a world-scale challenge.

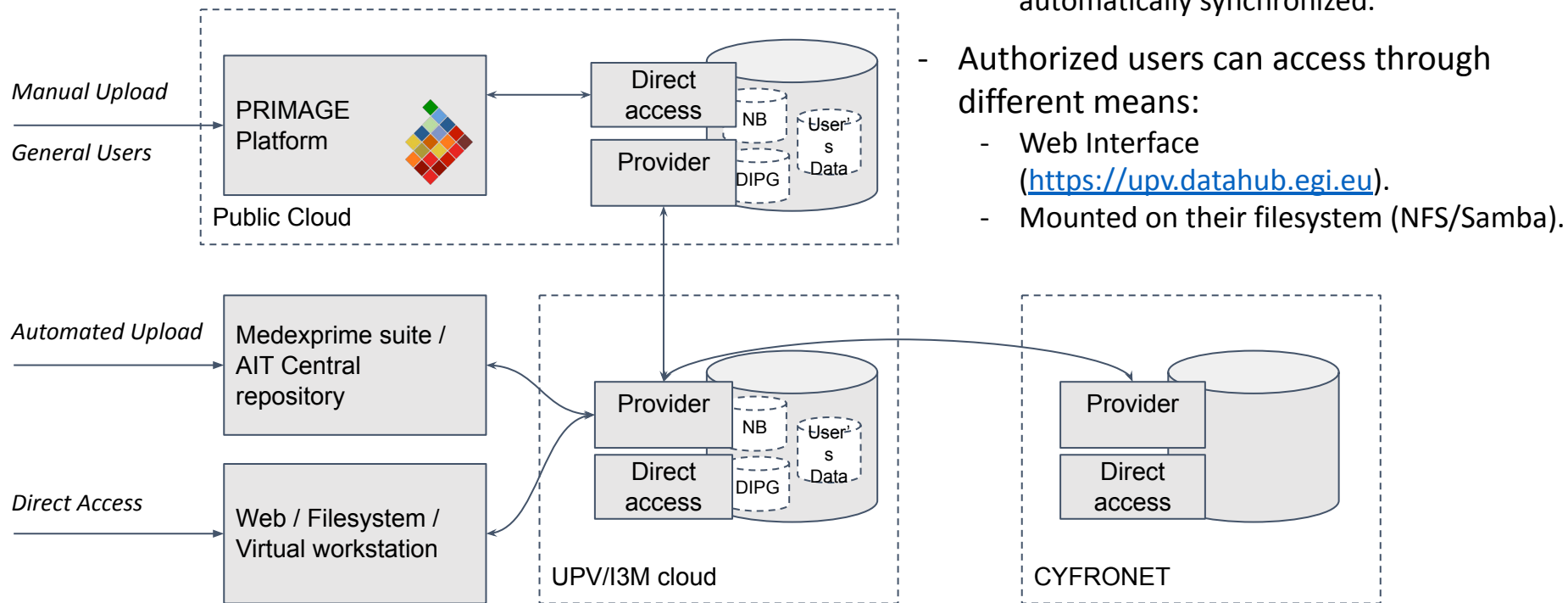
Motivation and Challenges

- The development of advanced methods for Cancer diagnosis, prognosis and treatment requires large, homogeneous and high-quality imaging data.
- Medical Imaging data comes from Clinical Trials but mainly from Real World Data.
- Collecting a representative annotated dataset is complex due to multiple dimensions:
 - The legal constraints on the secondary use of Medical Data.
 - The risk of patient reidentification.
 - The heterogeneity of multicentre datasources.
 - The reluctance of providers to share data.
- Cancer Imaging Repositories aim at reducing the barriers for those challenges.

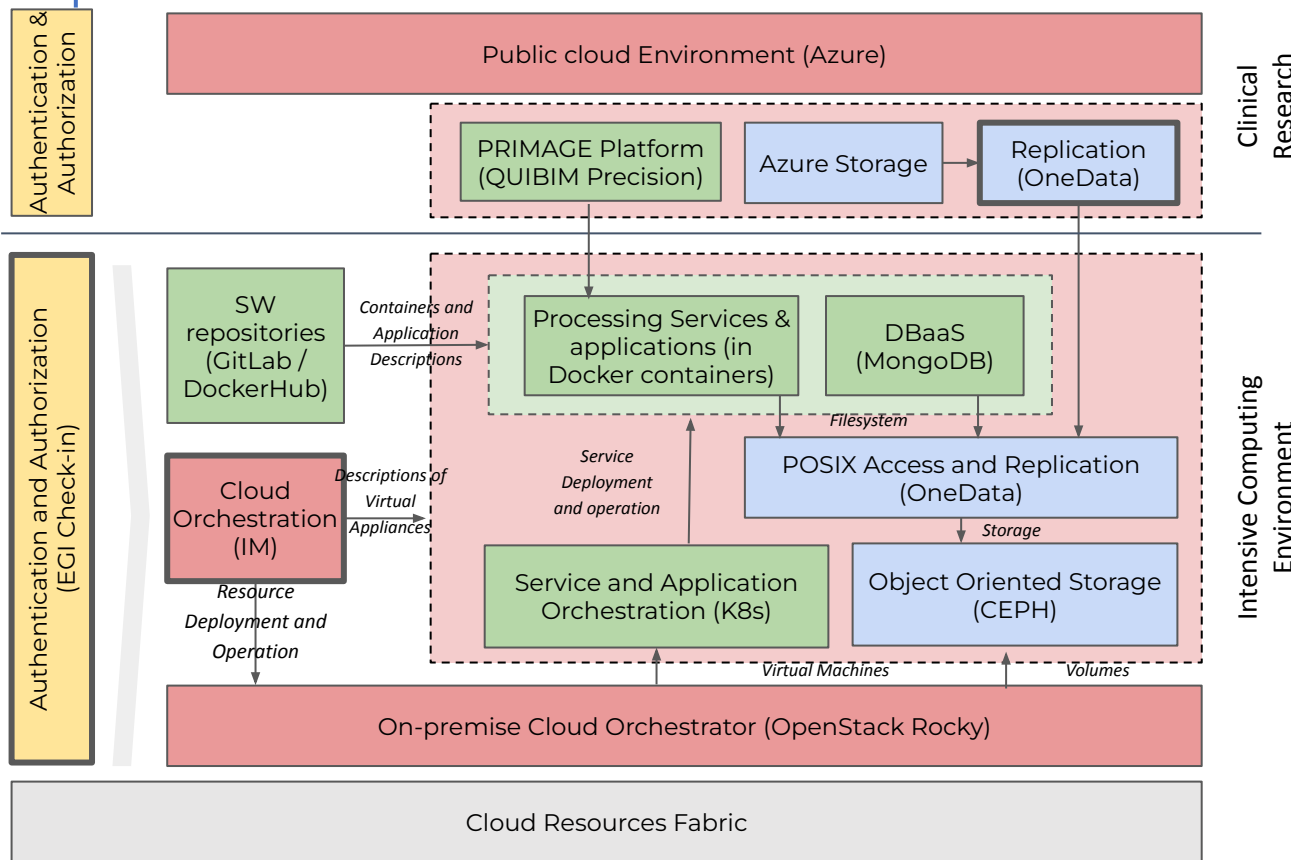
PRIMAGE

-

Integration with the Data Management



- Data is uploaded through a web
 - Data in the cloud storages are automatically synchronized.
- Authorized users can access through different means:
 - Web Interface (<https://upv.datahub.egi.eu>).
 - Mounted on their filesystem (NFS/Samba).

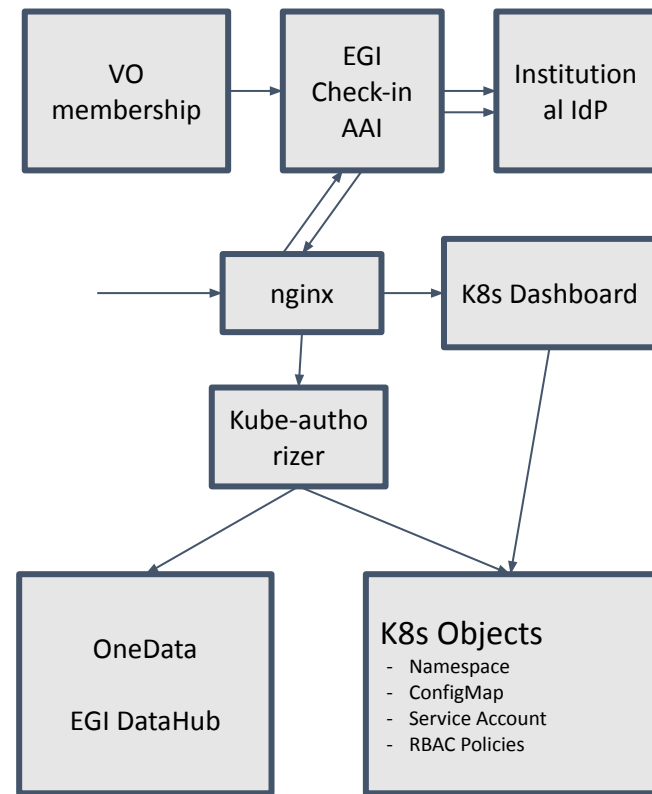


The architecture splits between the Intensive Computing research environment, where AI models training and simulation take place and the Clinical Research Environment where those models are usable for medical researchers.

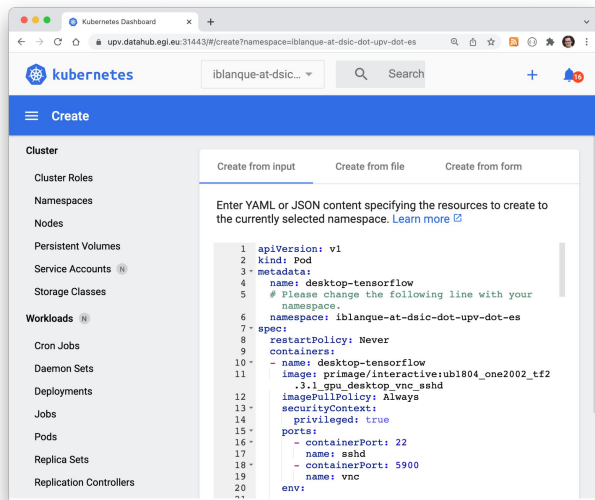
- The Cloud backend is platform agnostic and can be deployed in different cloud providers.
- The Cloud backend comprises an Object-Oriented Storage and a Service Orchestrator.
- The Service Orchestrator deploys and manages the applications available in trusted repositories.



- A VO has been created to manage access permissions at the level of the Kubernetes (K8s).
- The K8s Dashboard is put behind a proxy that authenticates via EGI-Checkin
 - Only authenticated users that belong to the vo.primage.eu can access the resources.
 - First time a user connects to the Dashboard, a service (Kube-Authorizer) triggers the creation of a set of objects
 - In K8s A namespace, a configmap with the generic configuration, a service account and RBAC policies.
 - Users can only create and browse objects within his/her namespace except for a few specific objects in the default namespace.
 - In EGI DataHub OneData, runs a container that creates a “home” directory and a set of permissions.
 - Further accesses go directly to the K8s Dashboard.
- This provides a seamlessly management of multiple sites with the same permission schema
 - Focusing on Repository Federation.

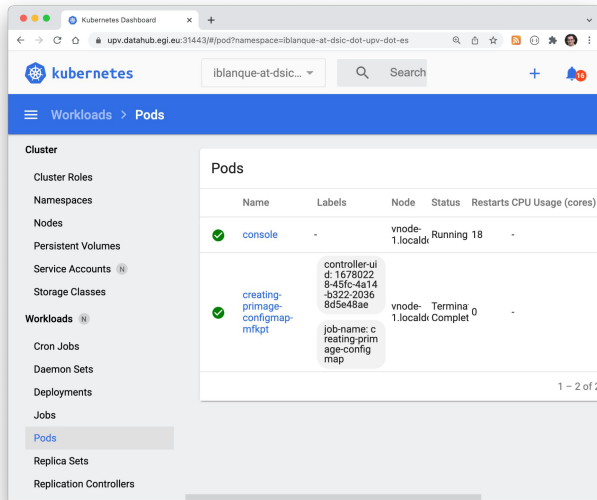


Processing backend



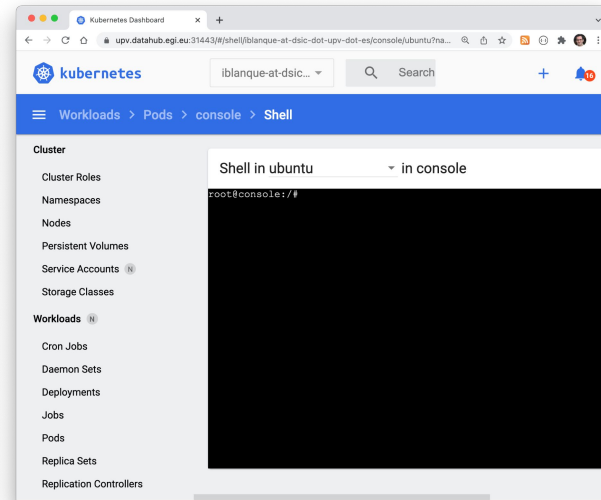
```
1 apiVersion: v1
2 kind: Pod
3 metadata:
4   name: desktop-tensorflow
5   # Please change the following line with your
6   namespace.
7 spec:
8   namespace: iblanque-at-dsic-dot-upv-dot-es
9   restartPolicy: Never
10  containers:
11  - name: desktop-tensorflow
12    image: primage/interactive:ubi804_one2002_tf2.3.1_gpu_desktop_vnc_sshd
13    imagePullPolicy: Always
14    securityContext:
15      privileged: true
16    ports:
17    - containerPort: 22
18      name: sshd
19    - containerPort: 5900
20      name: vnc
21  env:
```

Creating applications



Name	Labels	Node	Status	Restarts	CPU Usage (cores)
console	-	vnode-1.localdk	Running	18	-
creating-primage-configmap-mkpt	-	vnode-1.localdk	Terminating	0	-

Deploying Applications

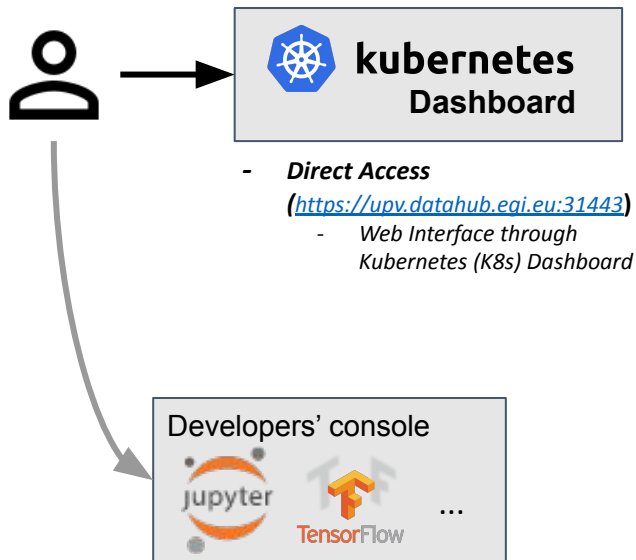


```
root@console: /#
```

Accessing Applications



Processing backend



- **Access to deployed interactive applications:**
 - Ports 30050-30070 for Jupyter and similar applications
 - Guacamole web client for desktop applications with SSH and Remote Desktop protocols <https://upv.datahub.eqi.eu:32443/guacamole/>

Deployment by template:

```
primage-batch-job.yaml 1013 Bytes
1  apiVersion: batch/v1
2  kind: Job
3  metadata:
4    name: #batch_job_name#
5    namespace: #namespace_name#
6  labels:
7    name: primage-batch-job
8  spec:
9    template:
10     metadata:
11       name: primage-batch-job
12     spec:
13       containers:
14         - name: primage-batch-job
15           image: #docker_image#
16           imagePullPolicy: Always
17           args: ["#command_line_arguments#"]
18       securityContext:
19         privileged: true
```

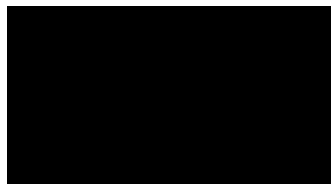
- **Build and Run PRIMAGE Applications (TR2.2 & D2.5 Documents)**
<https://gitlab.com/primageproject/documentation>
- **Types of applications:**
 - o Batch
 - o High Throughput Computing (HTC)
 - o Interactive
- **Templates and canonical examples:**
<https://gitlab.com/primageproject/applications>
- **Docker Containers (DockerHub PRIMAGE Organisation)**
<https://hub.docker.com/u/primage>
 - o 15 repositories, 42 containers.
- **API (python, Java) (GitLab PRIMAGE Repository)**
 - o <https://gitlab.com/primageproject/api-python-k8s>
 - o <https://gitlab.com/primageproject/api-java-k8s>



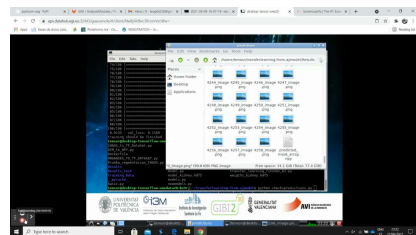
Relevant Links

- <https://gitlab.com/primageproject>
- <https://www.linkedin.com/company/primageproject/>
- https://twitter.com/primage_project

<https://www.primageproject.eu/>

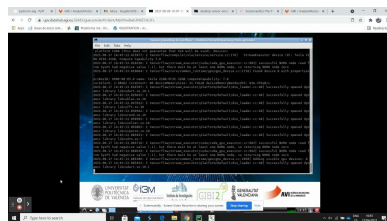


The PRIMAGE Backend

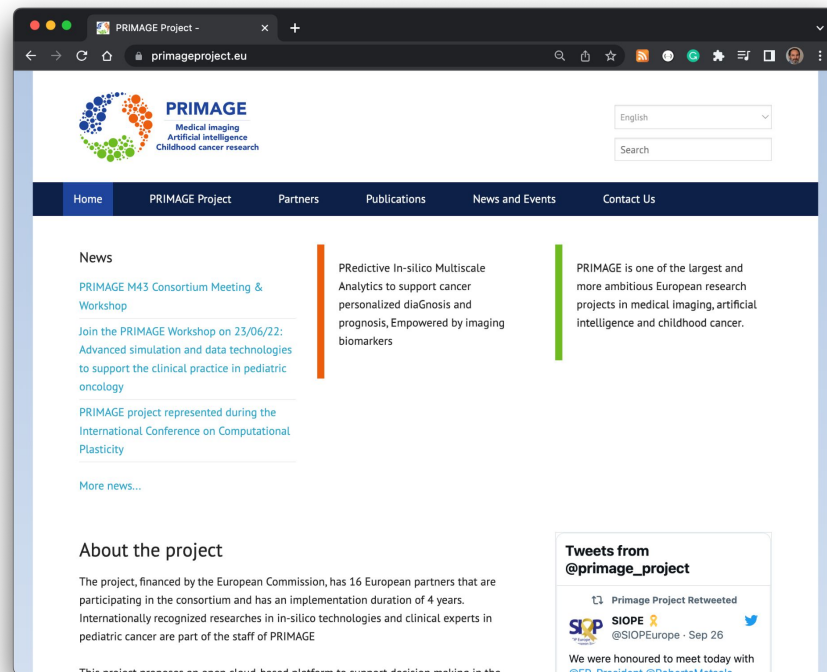


Using a previously trained model

Registering in the PRIMAGE VO



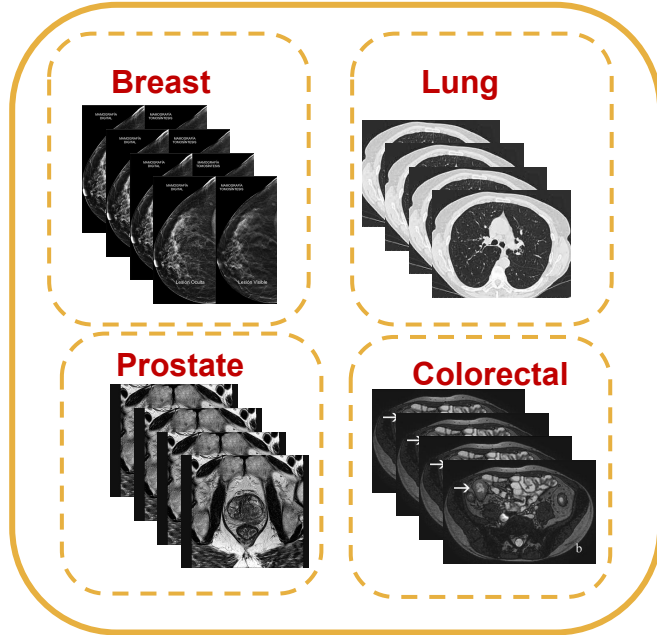
Training a model



CHAIMELEON



CHAIMELEON Project



Images + Related
clinical data (e-form)

Cloud-based cancer imaging repository as an online resource for the AI community working on the development of cancer management solutions

Not just a data warehouse...

- Incorporating all necessary functionalities to allow AI experimentation on the cloud (without downloading the data).
- Powered with automation tools.
- Interoperable with other existing initiatives.

Instituto de Investigación
Sanitaria La Fe

UNIVERSITÀ DI PISA

SAPIENZA
UNIVERSITÀ DI ROMA

centro hospitalar
do Porto

Gruppo
San Donato

CHARITÉ
UNIVERSITÄTSKLINIK BERLIN

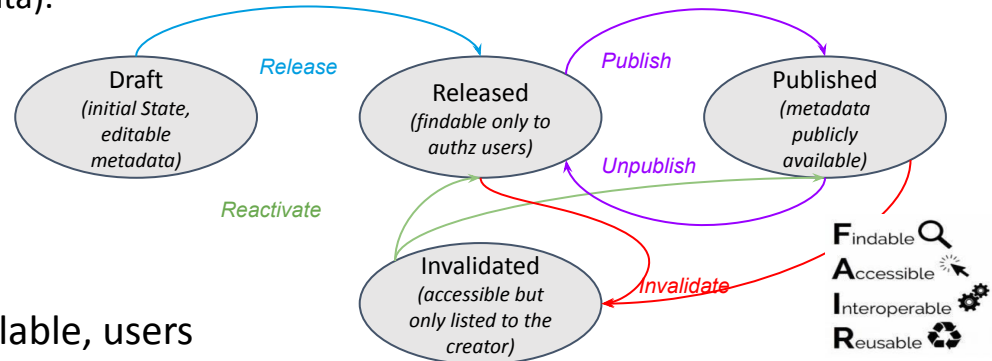
cerf
collège
des
enseignants
de
radiologie
de
france

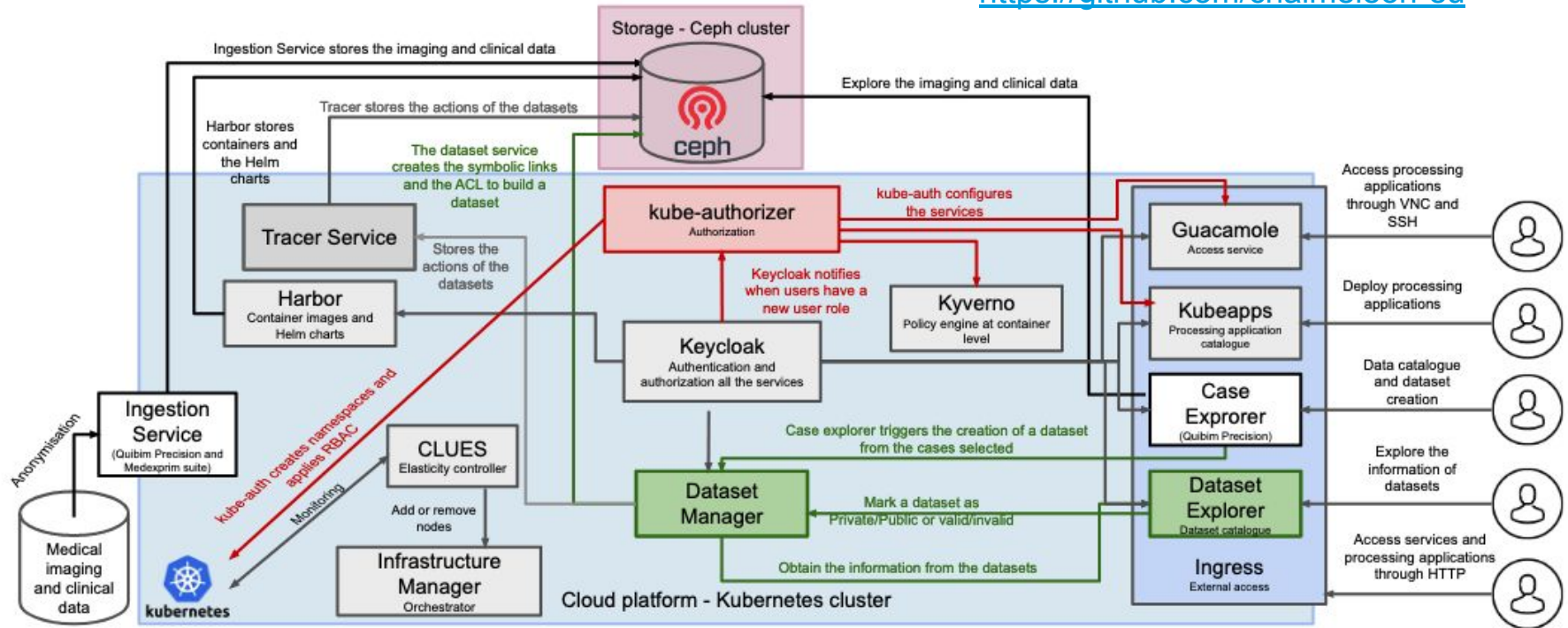




Organization of data: A Dataset

- A CHAIMELEON Dataset is a coherent set of annotated image studies and the associated clinical data that have a persistent identifier.
- A Dataset is a research object that can be citable and fulfils the FAIR principles.
 - Datasets have a metadata that contains aggregated information, following the MIABIS specification, which could be made public (just metadata).
 - Released datasets are discoverable (not necessarily accessible) by the users registered in CHAIMELEON.
 - Published Datasets have their metadata publicly accessible.
- By having the aggregated metadata available, users could raise interest for a specific dataset.
- Access will be granted upon request and after the approval of the Data Access Committee.

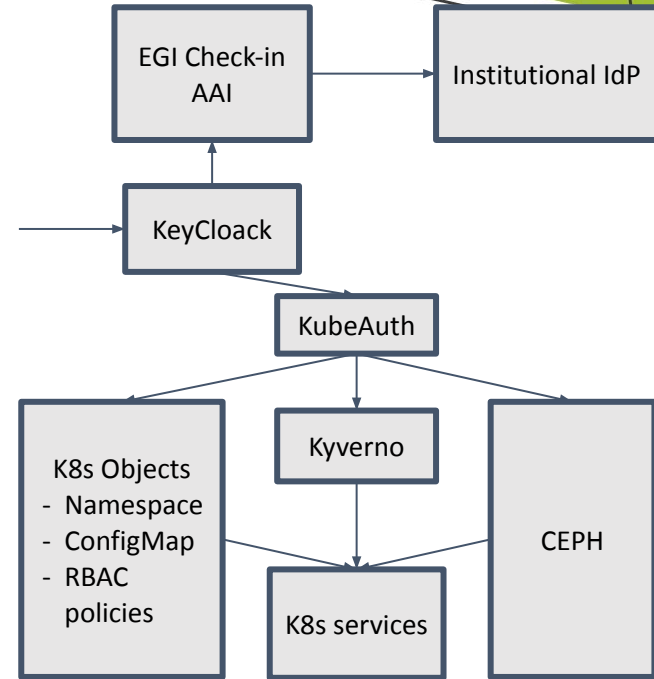






Integrating Complex Policies and CEPH

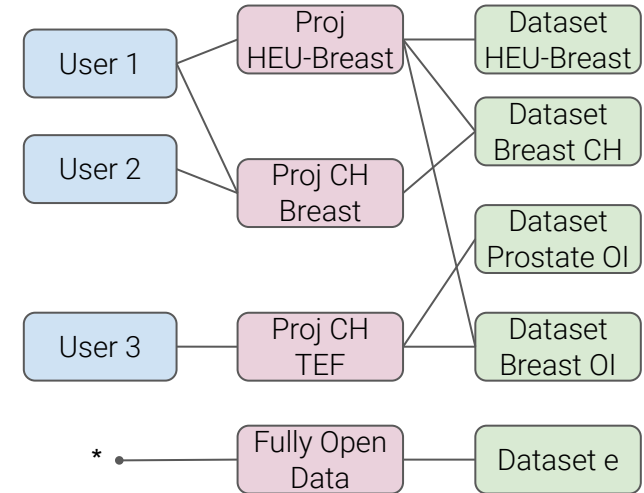
- RBAC in K8s is sometimes insufficient
 - Fine-grain permissions in CEPH, dedicated ingress routes, etc.
- Access to resources may not be only through K8s Dashboard.
- A solution based on KeyCloack and Kyverno has been implemented
 - Kyverno creates the policies for fine-grain permissions.
 - A container creates directly in CEPH the access permissions and the shadow volumes with part of the information
 - Users' home, Read-only permissions for general data and RW for home, etc.
 - Extending the existing functionality of KubeAuth (Namespaces, configmaps and regular RBAC policies).





Authentication and Authorisation

- Authentication is performed through Keycloak¹ and relies on EduGAIN² Identity Providers (IdPs)
 - This reduces the burden of managing additional credentials and relies on institutional third party IdP.
- Authorization is based in two concepts
 - Groups represented by projects (tenants)
 - A project has access granted to a set of datasets.
 - Users of a project share the same access permissions.
 - Access to a dataset may be granted to multiple projects.
 - Policies implemented through Kyverno³
 - It defines in a declarative way which resources and operations are allowed for a specific project.
 - Access to data is performed through RBAC (Role Based Access Control) policies.



¹ <https://www.keycloak.org>

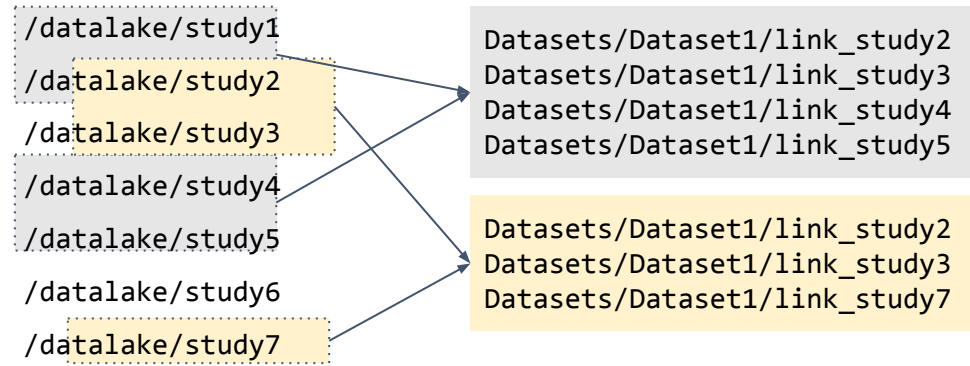
² <https://technical.edugain.org/status>

³ <https://kyverno.io/>



Access model in CHAIMELEON

- Users have access to the data only through controlled virtual environments running on the platform.
- A user can browse and process data in-situ, but cannot download data from the platform
 - Users can upload data and code though.
 - Access is performed through a proxy.
- Datasets are from the data Lake are not replicated, but symbolic links are provided instead
 - Faster, reduces resource wasting and flexible.





Datasets and Metadata

Dataset Explorer 1.1.7-BETA Datasets Fair Principles API Specs

Search: 3 records...

ID	Dataset	Flags	Author	Created	Subjects	Actions
22b357dc-96e7-4c97-beac-12cece51f30	Maastricht Lung1 v2	Published	Pau Lozano	04/04/2022, 11:04:23 CEST	422	422 More
53251c3f-0776-4a57-a10f-eae7aa33a0bd	TestDataset108	Published	Pau Lozano	21/01/2022, 11:34:09 CET	1	1 More
c7e69e21-e8b2-4ca5-96bf-4e0267c57d1d	TestDataset113	Published	Pau Lozano	04/02/2022, 14:12:55 CET	3	3 More

[Previous](#) [Next](#)

Dataset Explorer 1.1.7-BETA Datasets Fair Principles API Specs

Home / Dataset information

Maastricht Lung1 v2 (22b357dc-96e7-4c97-beac-12cece51f30) [Actions](#)

Created on 04/04/2022, 11:04:23 CEST **Published**

Description: Test dataset from Maastricht University.

Details [Studies](#) [History](#)

Author: Pau Lozano

PID URL: <https://doi.org/10.5072/zenodo.1070138>

Contact Information: Responsible from Maastricht University (unknown@maastrichtuniversity.nl)

License: [Maastricht University custom license](#)

Studies count: 422

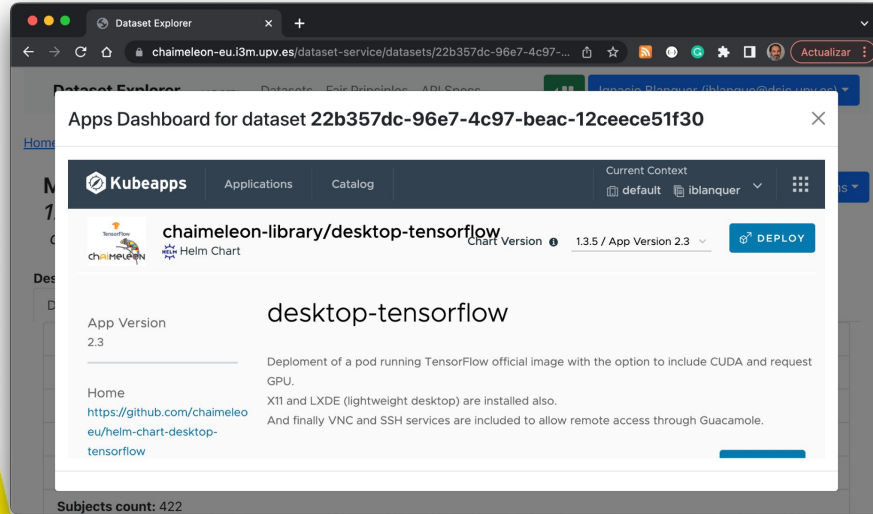
Subjects count: 422

<https://chameleon-eu.i3m.upv.es/dataset-service/datasets>

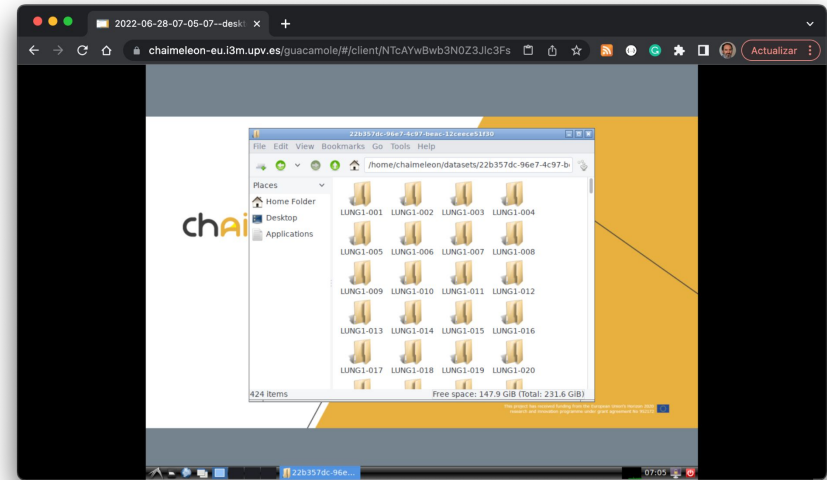




Virtual Environments and in-situ access



<https://chameleon-eu.i3m.upv.es/apps>

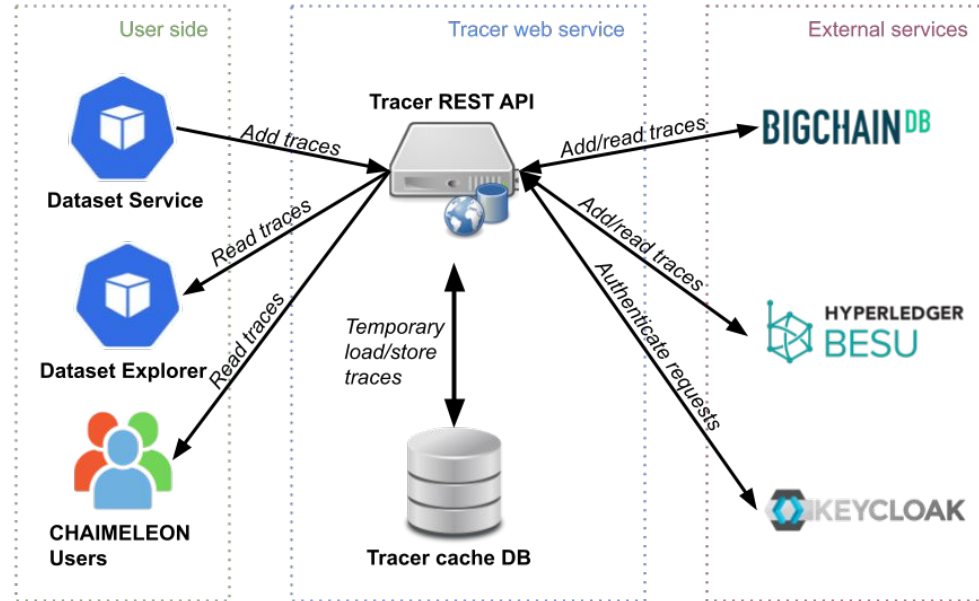


<https://chameleon-eu.i3m.upv.es/guacamole/#/>



- It logs user's actions on the cluster (the traces)
 - Create / update (properties of) / use datasets
 - Create / use models
- It stores them in blockchains
 - BlockchainDB
 - *Hyperledger Besu*
- It does not store user / patient private information
 - Only hashes (for files), ids
- It provides the only entry point to access the blockchain(s)
 - *Through Kube network policies*
- It ensures traces do not get lost
 - *By storing them in cache until the blockchain(s) incorporate(s) them successfully*

Use datasets in a pod: {id: a6dd9a18-c5e0-4fce-bed3-ae7dfab94f28, callerId: e9f5c5cc-cb9c-4c2e-a661-7b6470c7c0da, timestamp: 1663327393, version: V1, userId: f0e9d561-ce6b-44b5-8bdb-2bc9253eea86, userAction: USE_DATASETS, datasetsIds: [384e1361-09ed-4311-9f19-6bcf9b280fca, d087a6dd-9fce-4b33-82b2-a662bc9ffe47, e9f5c5cc-cb9c-4c2e-a661-7b6470c7c0da]}



[/chaimoleon-eu/tracer](https://github.com/chaimoleon-eu/tracer)



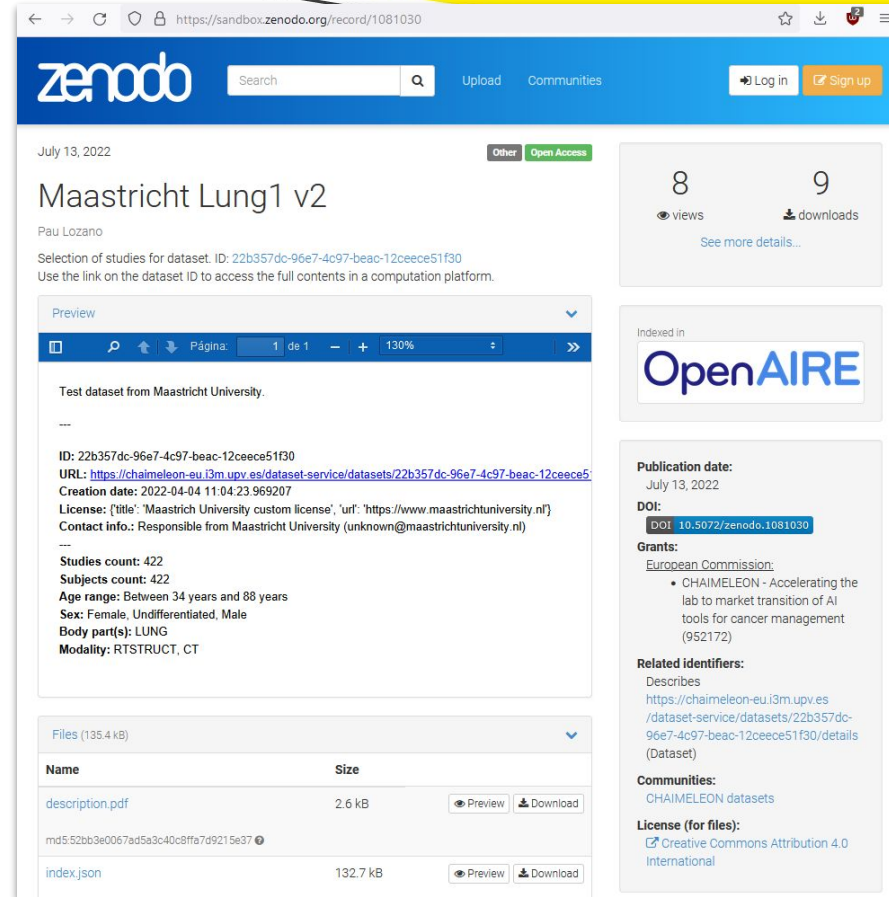
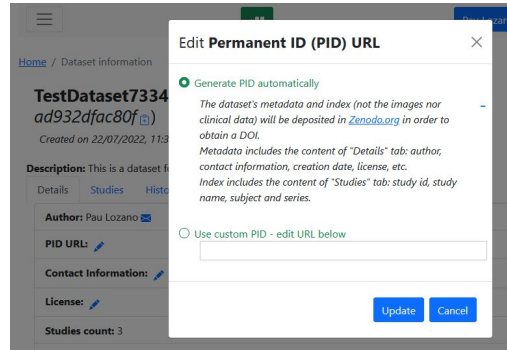
[/apis/UPV-CHAI MELEON/Tracer/](https://apis/UPV-CHAI MELEON/Tracer/)



Integration with Zenodo

- Dataset service now use the Zenodo API to automatically create a deposition of dataset metadata. This will be done for **published** datasets and optionally for **released** datasets.
- The publication at Zenodo gives visibility to the dataset and allow us to obtain a **DOI**, which is a permanent reference (Permanent ID Url) to the dataset that can be included in publications
- Only metadata will be published in Zenodo, never the contents of the dataset (images nor clinical data)
- On **invalidated** datasets, the Zenodo deposition will be closed (not listed in searches and showing the “closed” label for someone who access through a reference)

Currently the Zenodo sandbox endpoint is used, ready to change to production.

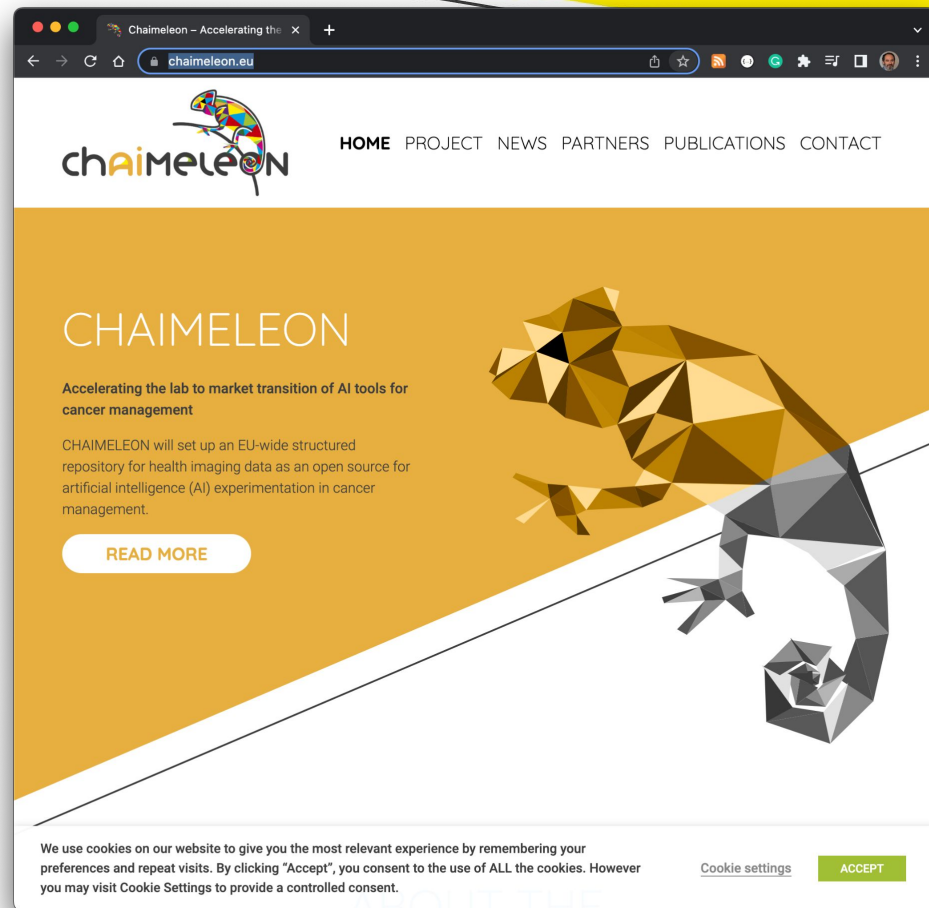


<https://chameleon.eu/>



Further Information

- <https://github.com/chameleon-eu>
- <https://chameleon.eu/>
- <https://www.youtube.com/channel/UC9jwaCgVs-RpSDIBvXencEQ>
- https://twitter.com/chameleon_eu
- <https://es.linkedin.com/showcase/chameleon>



AI 4 HI Network

AI for Health Imaging

1. **EuCanImage:** A European Cancer Image Platform Linked to Biological and Health Data for Next-Generation Artificial Intelligence and Precision Medicine in Oncology
2. **INCISIVE:** A multimodal AI-based toolbox and an interoperable health imaging repository for the empowerment of imaging analysis related to the diagnosis, prediction and follow-up of cancer
3. **ProCancer-I:** An AI Platform integrating imaging data and models, supporting precision care through prostate cancer's continuum
4. **CHAIMELEON:** Accelerating the lab to market transition of AI tools for cancer management
5. **PRIMAGE:** PRedictive In-silico Multiscale Analytics to support cancer personalized diaGnosis and prognosis, Empowered by imaging biomarkers



AI4HI

- The development of AI models requires a huge amount of high-quality, harmonized data.
- Summing up, the 5 AI4HI imaging repositories cover 5 cancer types (breast, lung, colorectal, prostate cancer, liver and paediatric tumours) over 91,000 patients in total.
- The challenge is to make this data interoperable and accessible
 - Interoperable goes beyond data formats and metadata.
 - Accessibility is enormously complex in a context where secondary usage, consent, purpose, data minimization, anonymization and ethic management impose legal obligations and a severe liability.
- The AI4HI Network has been organised into 8 working groups (Ethical and legal issues, Metadata interoperability, Data storage and management, Data annotation, AI development, AI validation, Clinical Working Group and Outreach Working Group), each consisting of 15 experts representing the 5 projects complemented with a wide range of stakeholders, perspectives, approaches and disciplines.



<https://eucanimage.eu>



<https://chaimoleon.eu>



<https://incisive-project.eu>

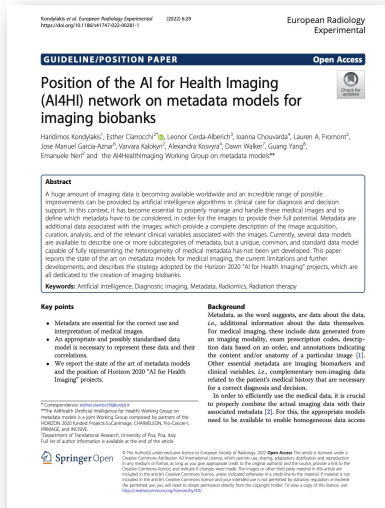


<https://www.procancer-i.eu>



<https://www.primageproject.eu>

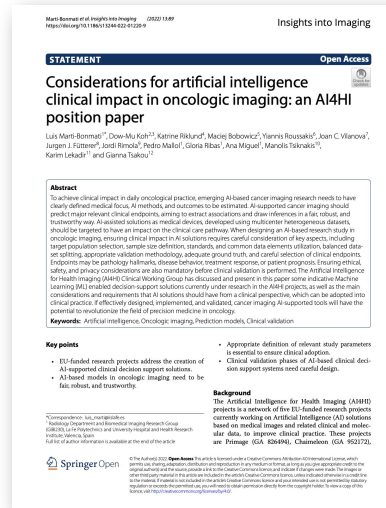
AI4HI Position Papers



Position of the AI for Health Imaging network on metadata models for imaging biobanks by Haridimos Kondylakis, et. al., European Radiology Experimental Journal.
doi.org/10.1186/s41747-022-00281-1



FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging by Karim Lekadir, et. al. (2021). arXiv preprint arXiv:2109.09658.



Considerations for Artificial Intelligence Clinical Impact in Oncologic Imaging by Luis Marti-Bonmati, et. al. Insights Imaging (2022).
doi.org/10.1186/s13244-022-01220-9

Under Review

Data Infrastructures for AI in Medical Imaging: A report on the experiences of five EU projects by Haridimos Kondylakis, et al., Journal of Biomedical Informatics

EUCAIM

federated EUropean infrastructure for CAncer IMages data

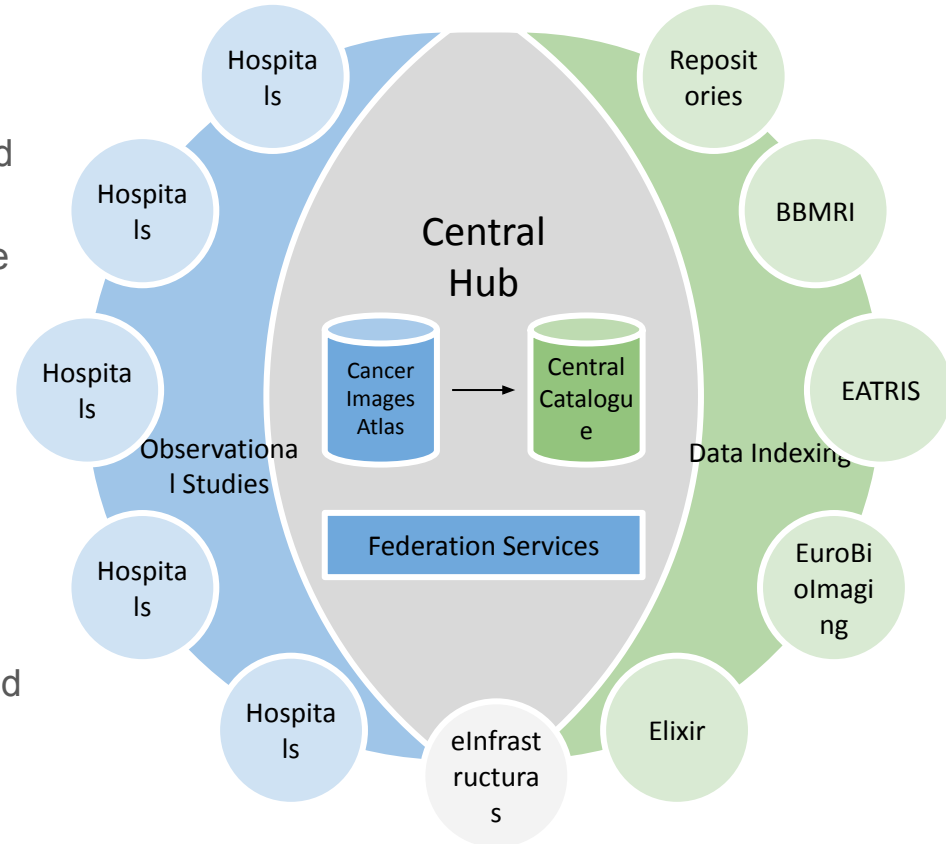


- **The EUropean Federation for CAncer IMages (EUCAIM) project originates from an unprecedented body of work and expertise of the “AI for Health Imaging” Network (AI4HI), which consists of 86 affiliated institutions from 20 countries involved in 5 large EU-funded projects on big data and AI in cancer imaging.**
- **EUCAIM project is not yet another proposal to build a new infrastructure from scratch, but an integrated architecture carefully designed by the AI4HI Network, and major European Research Infrastructures (Euro-Biolmaging, BBMRI, EATRIS and ELIXIR) on real-world achievements, as detailed in this proposal. The main concept of EUCAIM is of a central hub that federates distributed nodes (repositories, Research Infrastructures, and hospitals) to build up a hybrid distributed and centralized infrastructure on cancer images, including all types of cancer.**

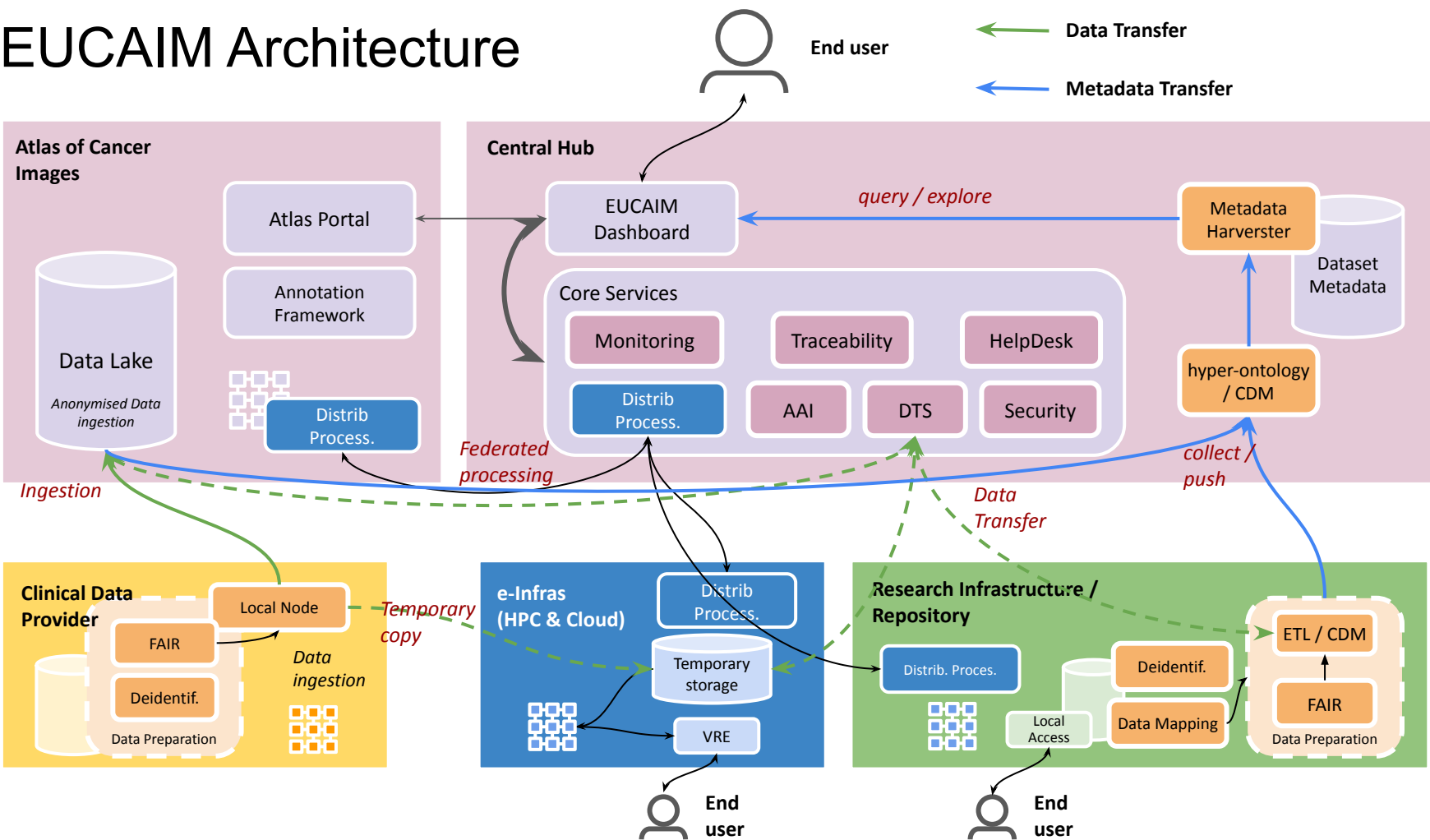
[illegible]

Data Scenarios

- **Catalogue of existing Datasets**
 - Metadata of available datasets is pushed into a central catalogue (push mode)
 - Datasets metadata is harvested from the providers side.
 - Data is homogenised to follow a Common Data Model.
 - Projects use the existing data as is.
- **Observational Studies**
 - New data is collected from the research data warehouses at the hospitals.
 - Project-driven approach, data is collected according to the selection criteria.



EUCAIM Architecture



Challenges on EUCAIM

- **Data Interoperability**
 - Data coming from clinical practice and from Observational Studies.
 - Data Preparation and deidentification stages at local providers.
 - Transformation through a Hyperontology and Image Harmonisation.
- **Federated Access to Repositories**
 - Push / Pull Dataset Metadata Harvester.
 - Data Governance procedures.
 - Redirection and federated access.
- **Federated Processing**
 - To avoid bulk data transfers and to maximise privacy preservation.
 - In-situ and in trustfull third-party processing.
- **Sustainability**
 - Recognition models for providers. A complex trade-off between privacy preservation and public recognition.

Conclusions

- Medical Imaging repositories constitute a great challenge for e-Infrastructures.
- Data privacy preservation together with ethical issues impose non-trivial access restrictions.
- The high computational demand of model training (and even the inference stage for complex Deep Learning models) require the availability of large computing facilities close to data.
- Beyond the FAIR principles, data sovereignty, data interoperability, data harmonisation, traceability and data processing capacity are enormous challenges to tackle the needs of applying AI techniques.

More information

Ignacio Blanquer

Institute of Instrumentation for Molecular
Imaging (I3M)
Universitat Politècnica de València

iblanque@dsic.upv.es

www.grycap.upv.es

github.com/grycap

chaimeleon.eu/

www.primageproject.eu/

