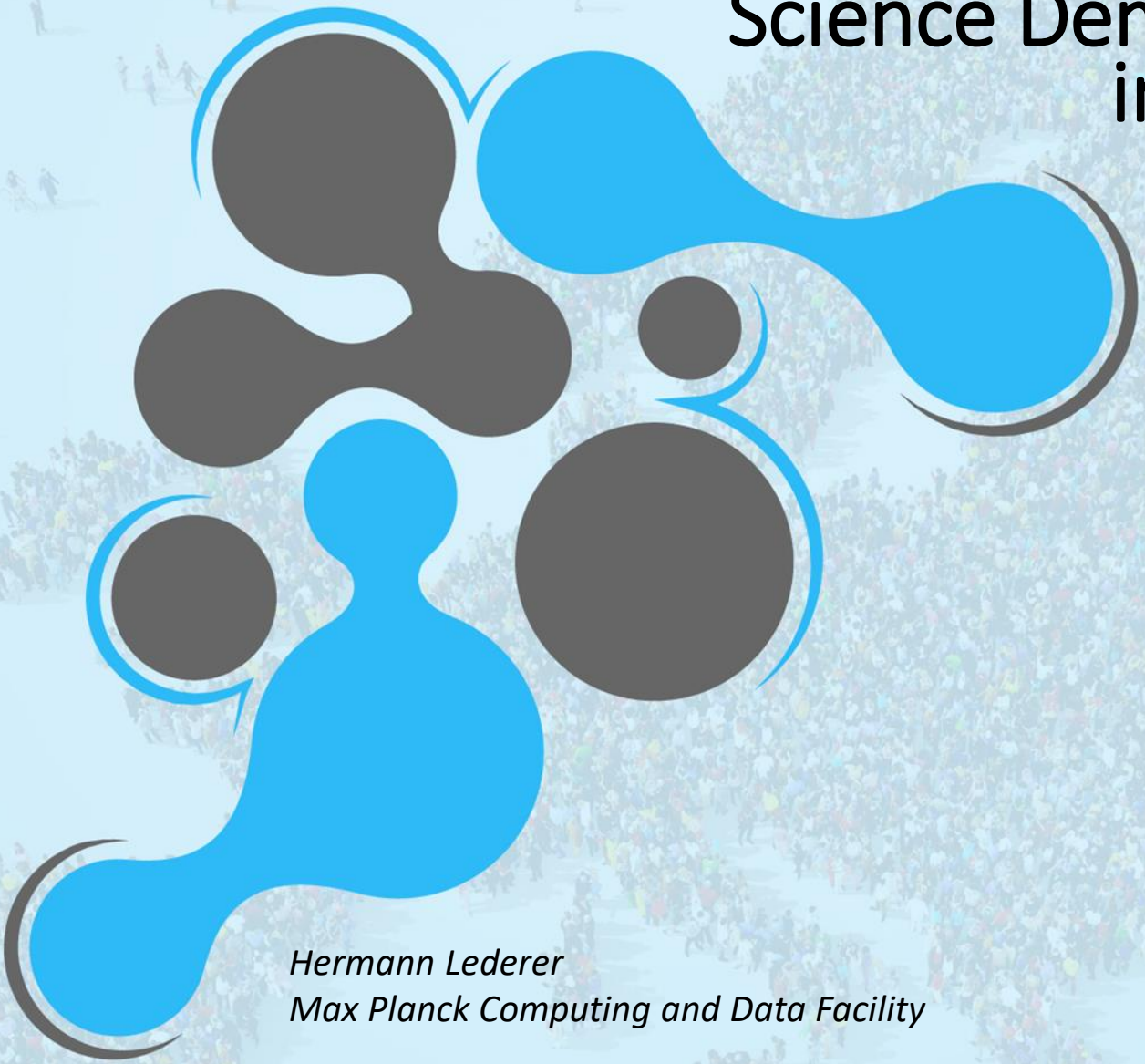
# Science Demonstrators in EOSCpilot



*Hermann Lederer*
*Max Planck Computing and Data Facility*

EOSCpilot
The European Open Science Cloud for Research Pilot Project
www.eoscpilot.eu

First Stakeholder Event, Brussels - Nov 28 , 2017

# Science Demonstrators in EOSCpilot

## What are *Science Demonstrators*?

**Challenging projects**

helping to define the infrastructure

needed by European researchers

showing the **scientific excellence and societal impact**

that could be achieved

by the European Open Science Cloud (EOSC )



The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

# Science Demonstrators in EOSCpilot

**Supporting 15 Science Demonstrators (SDs) in EOSCpilot (12 months each)**

| | |
|---|---|
| First five SDs: | Preselected prior to EOSCpilot begin<br>from 29 proposals, *Jan – Dec 2017* |
| Second five SDs: | Selected after 1st Open Call in April 2017<br>from 30 proposals, *July 2017 – June 2018* |
| Third five SDs: | Selected after 2nd Open Call in Aug/Sep 2017<br>from 26 proposals, *Dec 2017 – Nov 2018* |


The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

# Science Demonstrators in EOSCpilot

## First 5 Science Demonstrators (Jan – Dec 2017)

- **Environmental & Earth Sciences** - ENVRI Radiative Forcing Integration to enable harmonised data access and integration across multiple research communities
- **High Energy Physics** – DPHEP/WLCG: large-scale, long-term preservation and re-use of HEP data in the EOSC open to other researchers
- **Social Sciences and Humanities** – TEXTCROWD: Collaborative semantic enrichment of text-based datasets by make new software available on the EOSC.
- **Life Sciences** - Pan-Cancer Analyses & Cloud Computing within the EOSC to accelerate genomic analysis on the EOSC
- **Physics / Materials Science** - The photon-neutron community to improve the community's computing facilities by creating a virtual platform for all users


The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

# Science Demonstrators in EOSCpilot

- Second 5 Science Demonstrators (July 2017 – June 2018)
  - **HPCaaS for Fusion** - Culham Science Centre, UK
  - **Life Science Leveraging EOSC** to offload updating and standardizing life sciences datasets and to improve studies reproducibility, reusability and interoperability- CRG, Spain
  - **Seismology**: EPOS Virtual Earthquake and Computational Earth Science e-science environment in Europe- University of Liverpool, UK
  - **CryoEM** Linking distributed data and data analysis resources as workflows in Structural Biology with cryo-Electron Microscopy: Interoperability and reuse CSIC, Spain
  - **Astronomy Open Science Cloud** access to LOFAR data - ASTRON, NL


The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

# Science Demonstrators in EOSCpilot

**Third 5 Science Demonstrators** (Dec 2017-Nov 2018)

- **Generic Technologies** - **Frictionless Data Exchange** Across Research Data, Software and Scientific Paper Repositories - Open University (UK) and Los Alamos National Lab (without funding) with support by PIN (IT) and CERN

- **Life Sciences and Health Research**: **Mining a large image repository** to extract new biological knowledge about human gene function - University of Dundee (UK), EMBL (DE), EMBL-EBI (UK) related to Euro-Bioimaging

- **Astro Sciences** - **VisIVO**: Data Knowledge Visual Analytics Framework for Astrophysics - INAF (IT) with international engagements in SCI-BUS, ER-flow, VIALACTEA, INDIGO DataCloud, ASTERICS, AENEAS, AARC2

- **Hydrology** - Switching on the EOSC for **Reproducible Computational Hydrology** by FAIR-ifying eWaterCycle and SWITCH-ON - Delft University of Technology (NL), Netherlands eScience Center (NL), SMHI (SE), EMBnet, SURFSara (NL), EGI, CYFRONET (PL), Bavarian Academy of Sciences (DE)

- **Social Sciences and Humanities** - **VisualMedia:** a service for sharing and visualizing visual media files on the web - ISTI-CNR (IT), PIN (IT), MIBACT – ICCU (IT), Athena Research Center (GR), MPG (DE), CNRS (FR)


The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

# Science Demonstrators in EOSCpilot

## Panelists for the Plenary Session

**CryoEM -** Carlos Oscar Sorzano (CNB)

**ERFI -** Werner Kutsch (ICOS RI)

**EPOS/VERCE** - Andreas Rietbrock (U. Liverpool)

**Life Sciences Datasets -** Erik van den Bergh (EMBL)

**LOFAR -** Rob van der Meer (Astron)

**Pan-Cancer -** Sergei Iakhnin (EMBL)

**Photon-Neutron Research -** Michael Schuh (DESY)

**PROMINENCE -** Shaun de Witt (UKAEA)

**TEXTCROWD -** Franco Niccolucci & Achille Felicetti (Univ. Florence)


The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

## Questions for the plenary session

What is the excellent science the Science Demonstrator  stands for?

What are the main challenges for the SD?

What is the particular use case examined for the EOSC pilot?


The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

# Science Demonstrator PanCancer

**ORGANISATION:** Genome Biology Unit, European Molecular Biology Laboratory (EMBL)
CONTACT: Sergei Iakhnin

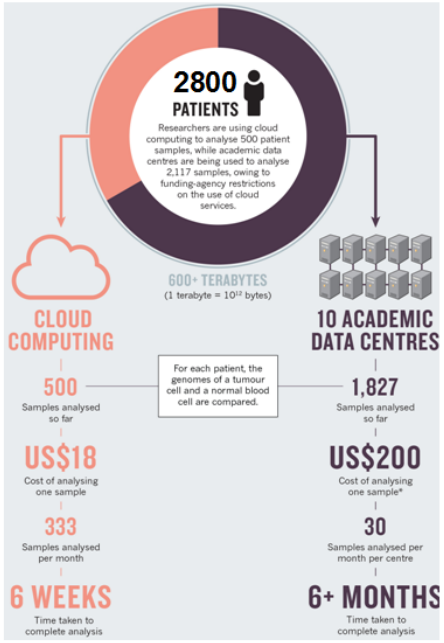
## Major European Cloud Use-Case in Genomics:
### The Pan-Cancer Analysis of Whole Genomes Project (PCAWG)

- **Mission of PCAWG phase I**: joint reanalysis of **2,800 cancer genomes**, 1 Pb of DNA data, using hybrid clouds.
- **Global network** (includes USA, Canada, Japan, Korea, Europe)
- **European leadership** (until recently co-funded by EC / Eurocan-Platform).

- Our initial focus has been on **standardizing** genomic **data processing**, and data redistribution, **on the cloud**.
- We **aim to make these processes inter-operable for the EOSC**.

- Outcomes of PCAWG phase I are intended to be published in **2017**.

**The International Cancer Genome Consortium (ICGC)**




Stein, Knoppers, Campbell, Getz & Korbel, *Nature* 2015

## *Demonstrator Overview*

Leveraging open science analysis models around controlled access data sets developed in collaboration with researchers elsewhere in the world.

These analysis frameworks could also be re-used to analyse cardiovascular and neuro-degenerative diseases as well as stimulating biotech/pharmaceutical industries to use public cancer genomic data in R&D.


The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

# Science Demonstrator TEXTCROWD

Collaborative semantic enrichment of text-based datasets

ORGANISATION: PIN srl; CONTACT: Franco Niccolucci

***Demonstrator Overview:*** Cultural heritage and humanities datasets are largely based on texts. Enabling the semantic enrichment of text sources through cooperative, supervised crowdsourcing, based on shared semantics, and then to make this work available to others via EOSC.

## Issues to solve

- Access to these data (partially) solved with online publication
  - Different repositories, access rules, fragmentation
- Discovery still problematic:
  - Resource registries: only the general scope of the text is described, according to a dataset model
  - Free-text search within specialized repositories: cumbersome, not always available, different repositories, incoherent results
  - In general, poor metadata for content
  - Need to add semantics of content
  - Indexing and metadata creation by hand is time-consuming and may be inconsistent/unreliable
  - Multilingualism
- Text mining with NLP and Information Extraction a solution

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

# Science Demonstrator ENVRI/ERFI

## ENVRI Radiative Forcing Integration

Organisations & Contacts: Werner Kutsch, Alex Vermeulen (ICOS ERIC), Ari Asmi (ENVRIplus) Paolo Laj (ACTRIS), Stefan Kindermann, IS-ENES2 (DKRZ), Sylvie Joussaume, Sébastien Denvil, IS-ENES2 (IPSL)

***DEMONSTRATOR:***

Focus on dynamics of greenhouse gases, aerosols and clouds and their role in radiative forcing, Interoperability between observations and climate modeling; cooperation between environmental research infrastructures.

Improvement of data integration services based on metadata ontologies, model-data integration by use of HPC, Petascale data movement, innovative services to compile and compare model output from different sources, especially on semi-automatic spatiotemporal scale conversion

***FAIR CHALLENGES:***

Findability: Metadata ontologies matching between NETCDF-CF and in-situ metadata, data quality indicators.

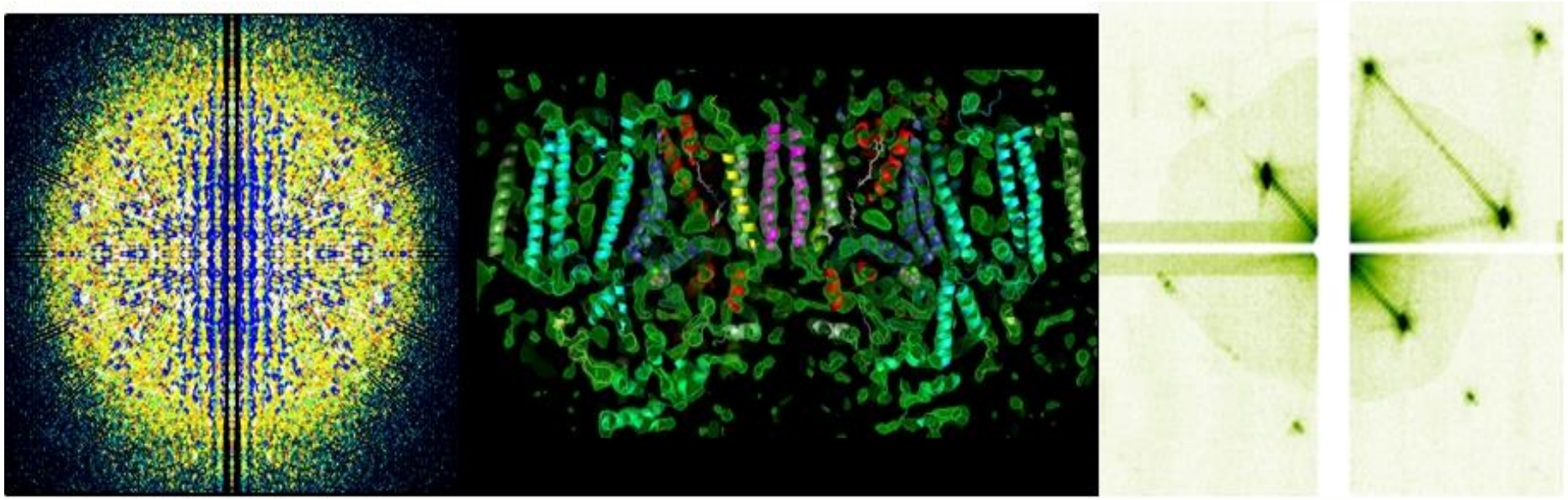
Accessibility: Automated access routines between the RI repositories. For fully open data, this is not immediately problematic, but might require analysis on needed resources and APIs.

Interoperability: APIs, service integration, large data transfers, where to do processing (how to document?)

Reusability: Citing and persistently identifying scale-changed data-sets? How to transfer knowledge of data versions used.


The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

# Photon & Neutron Science Pilot



DESY

ESS

European XFEL

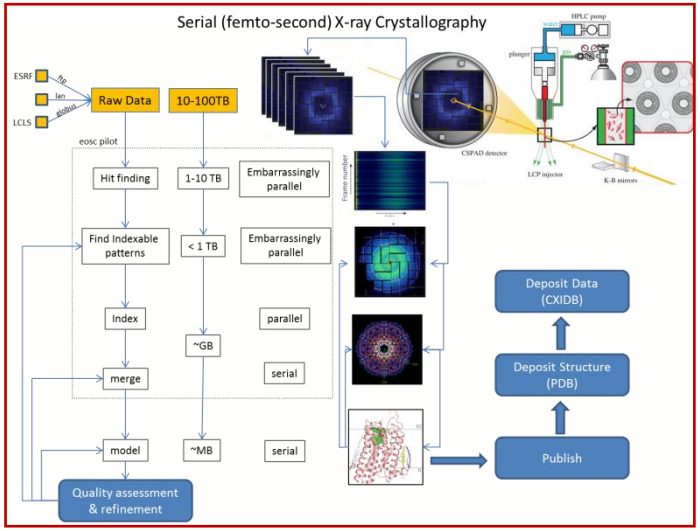


The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

# Photon & Neutron Science Pilot

## *Photon & Neutron Sciences*

- User dedicated and user driven facilities
  - >30 facilities with over ~50k users/yr worldwide
- Big Data
  - volume-velocity-variety-variability-value
  - Plethora of diverse instruments and sensors
  - Precious and often irreproducible (at reasonable costs)
- Big Impact
  - In very many areas of applied and fundamental sciences
  - ~11k peer reviewed publications/yr
  - Technology & innovation driver e.g. FELs, Lasers, Detectors, Techniques, etc




## *Challenges*

- Identity and Access Management
- Big Data
  - Several 10PB in trillion of files every year
  - Aggregation from thousands of data sources: hard to manage
  - Write Once; read a few times; rarely re-use: hard to cache & distribute
- Competitive Science
  - Fine-grained rights management on everything from software to data
  - Policies & standards supporting FAIR Open Science: a tedious process
- Application Frameworks
  - Complex software stacks varying from experiment to experiment
  - Rights management: Substantial parts are proprietary or non-redistributable

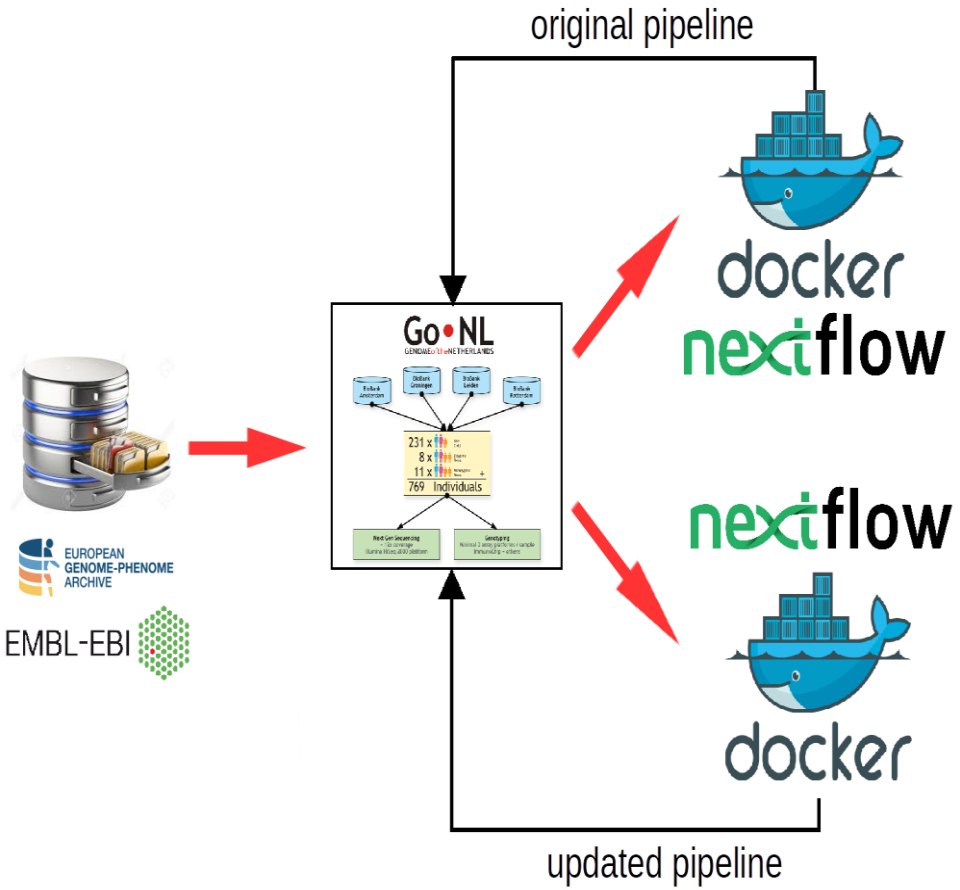
## *Science demonstrated*

- Prototypical: Serial (femto-second) X-ray Crystallography
- Extremely demanding in terms of volume, velocity and value
- Complex non-distributable software framework & complex workflow
- Successful containerization & cloud deployment
- Also targeting: Online Data Analysis and data streaming for X-FELs
- Also targeting: collaborative cloud configuration management
- Also targeting: DaaS



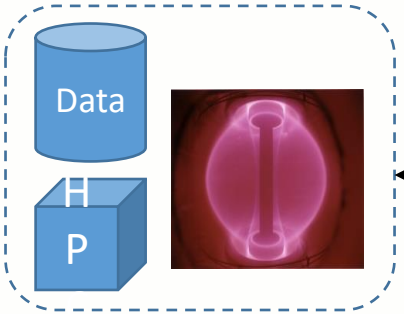
The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

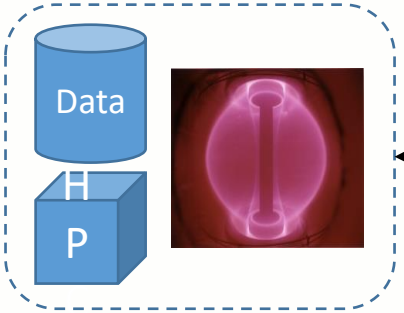
# EGA Life science datasets

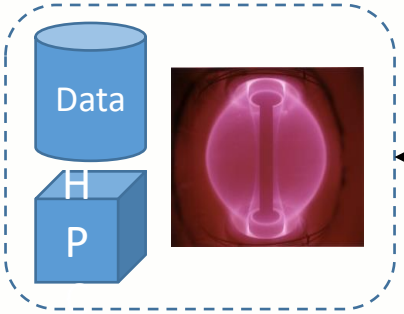



- A third part dataset (GoNL project) as use case
- Reproduction of the original pipeline
- Production of an updated pipeline
- Containerized versions of both pipelines using NextFlow
- Test both pipelines on the use case dataset

# Prominence: HPCaaS




Local Data and Resources

Analytics&Modelling

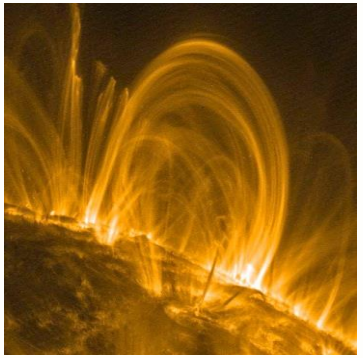
Analytics&
Modelling

Analytics&Modelling

Researcher

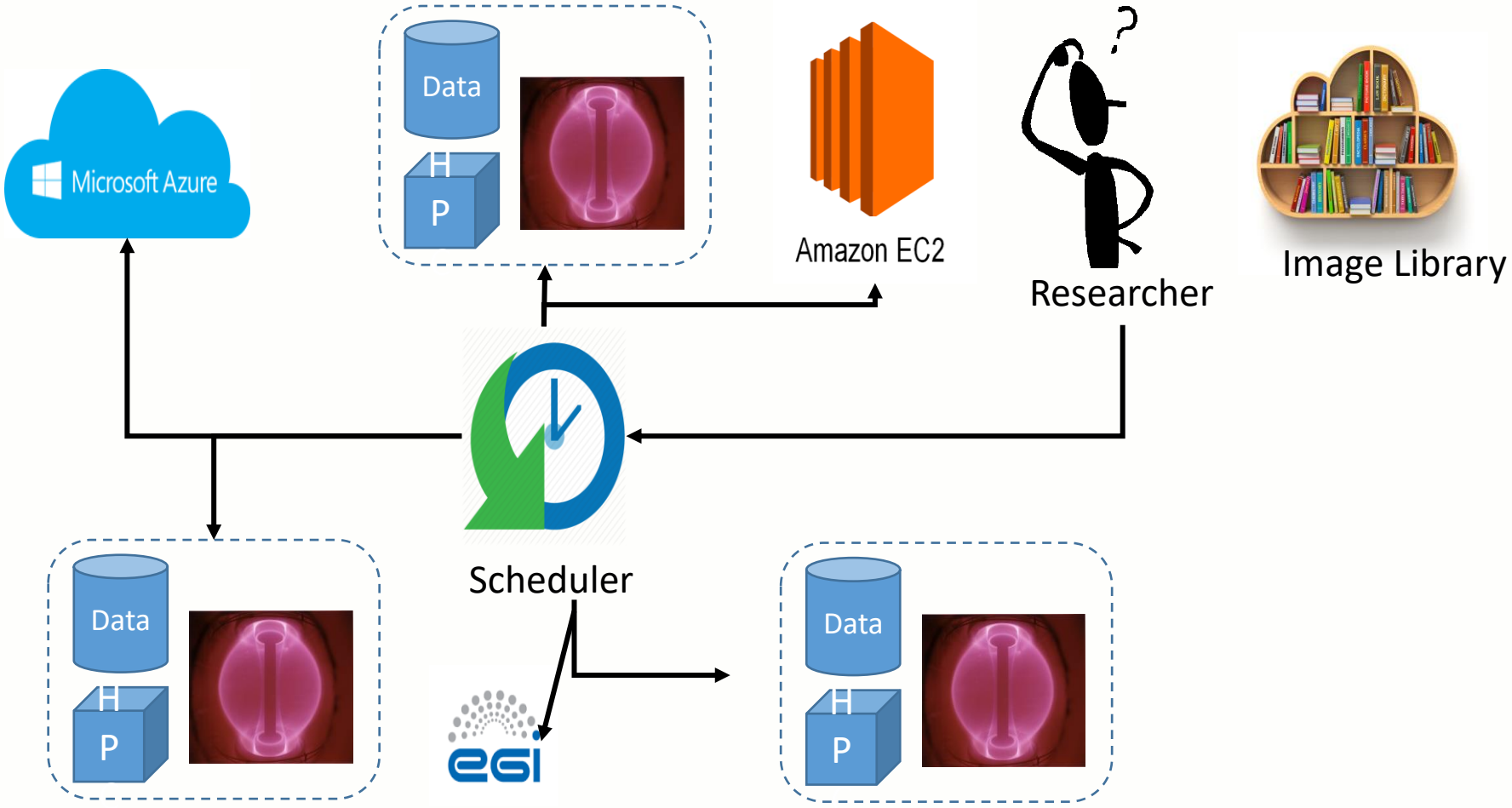
Modelling

Tier-0 and Tier-1 Compute Resources

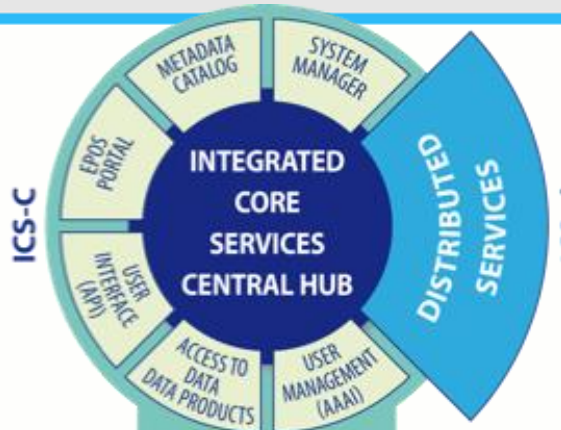


**Fusion: The Current Situation**

# PPROMINENCE: Demonstrator



# EPOS Computational Earth Sciences




CES

Computational Seismology

Satellite Data processing

Anthropogenic Hazard

Integrated Core Services
**ICS**

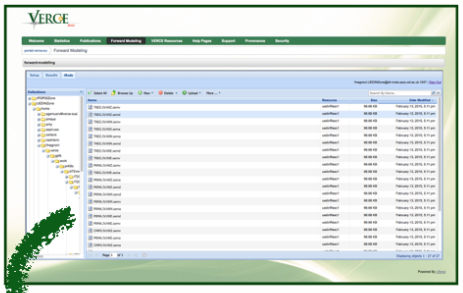
Thematic Core Services
**TCS**

National Research Infrastructures
**NRI**

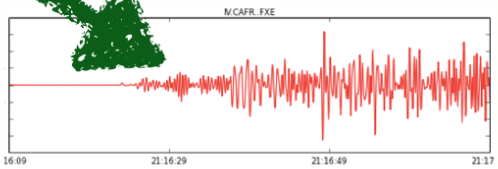


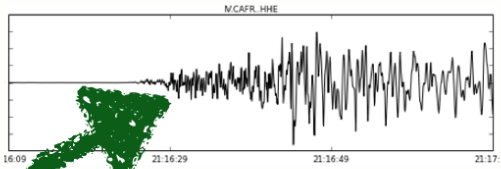
THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON 2020 RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT N° 676564
Disclaimer: the content of this presentation reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains

# EPOS/VERCE: Towards Inverse Modeling



Simulated Synthetics

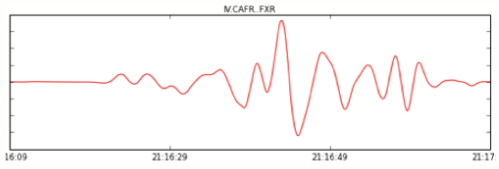


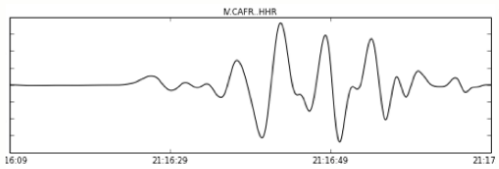




Data download

EIDA nodes

PRE-PROCESSING

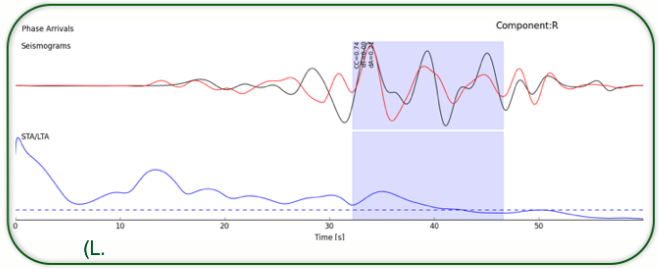


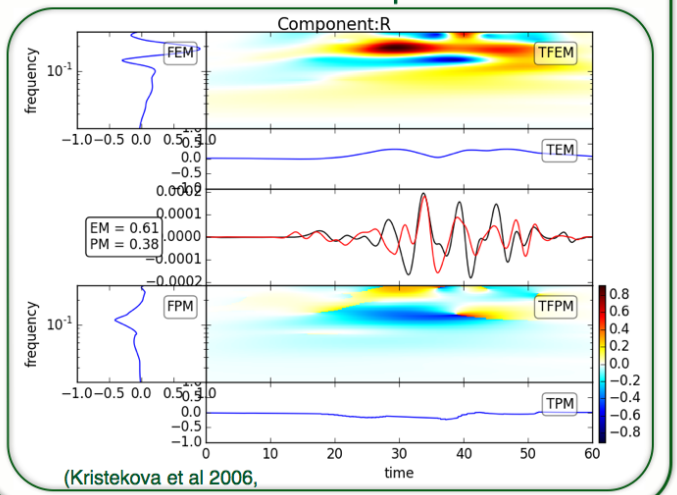

(filters, instrument correction, resampling, rotation, …)

POST-PROCESSING

Time Window Selection -




Phase and Amplitude



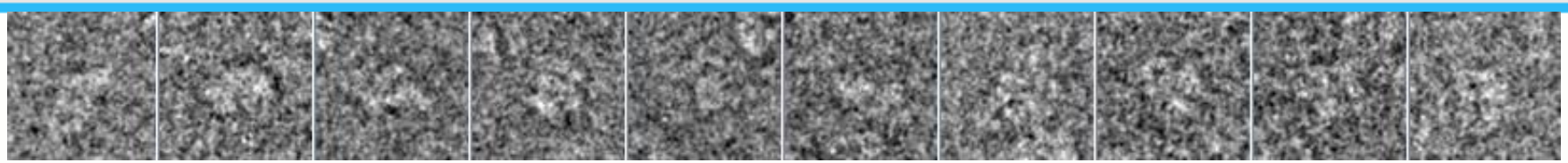

**HPC** **Server** **Cloud / EOSC**

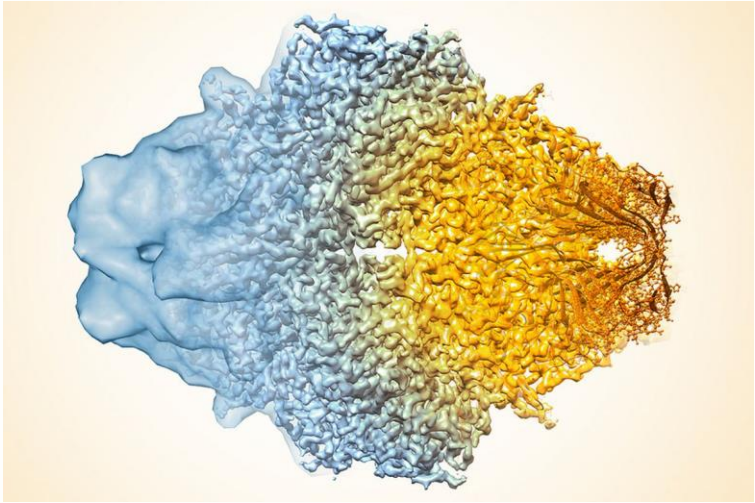


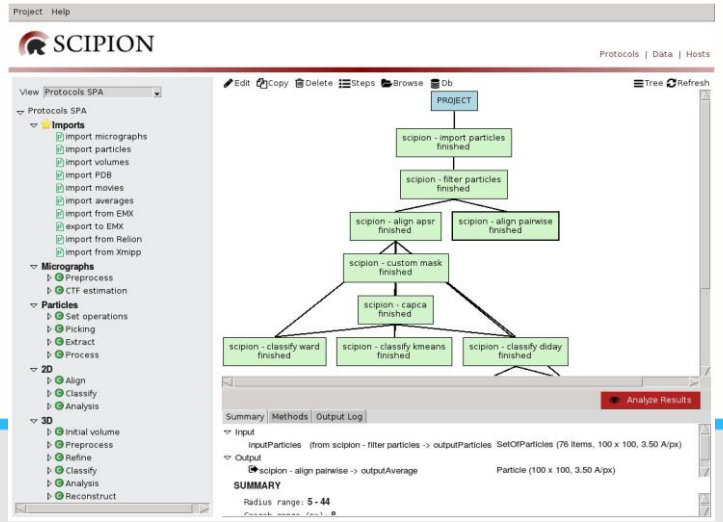
THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON 2020 RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT N° 676564
Disclaimer: the content of this presentation reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains

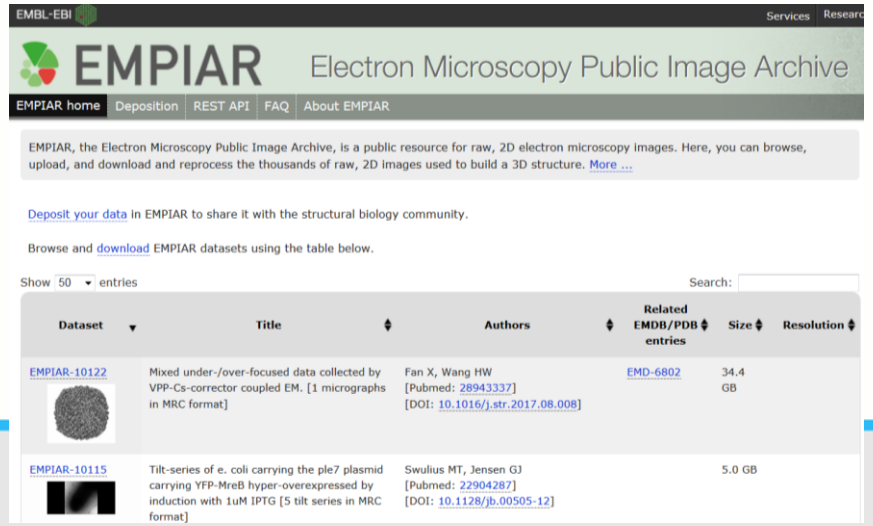
EOSCpilot
The European Open Science Cloud for Research Pilot Project
www.eoscpilot.eu

# Cryo-Electron Microscopy Science Demonstrator







# LOFAR Science Demonstrator



## *Scientific Problem*

**Reduce and analyse Radio astronomy data**:

From antenna signal to visibilities to images

## *Challenges*

**Large volume, complex (multi-step analysis)**

- Power users need compute at their data because of volume
- Unexperienced user need guidance with parameters and data sets
- Make it work across platforms and data centers

# LOFAR Science Demonstrator



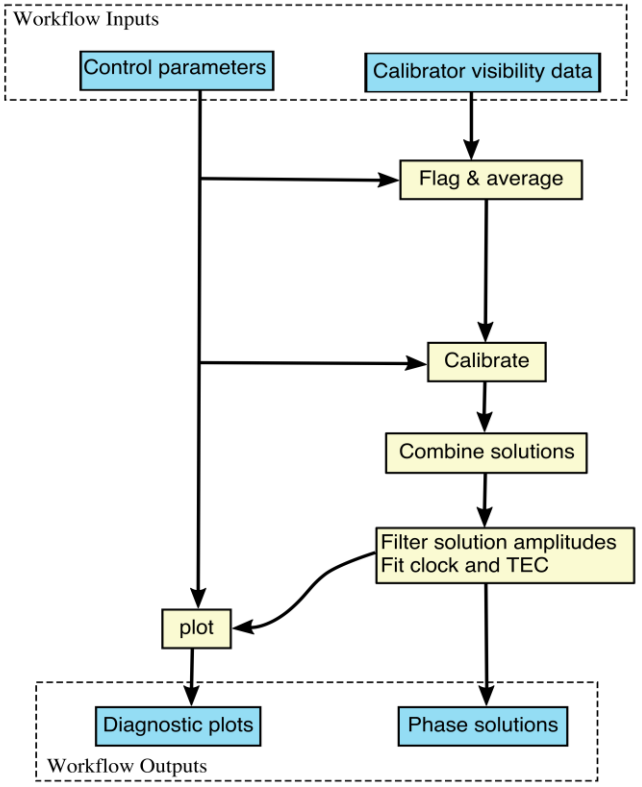
***Approaches to solution***

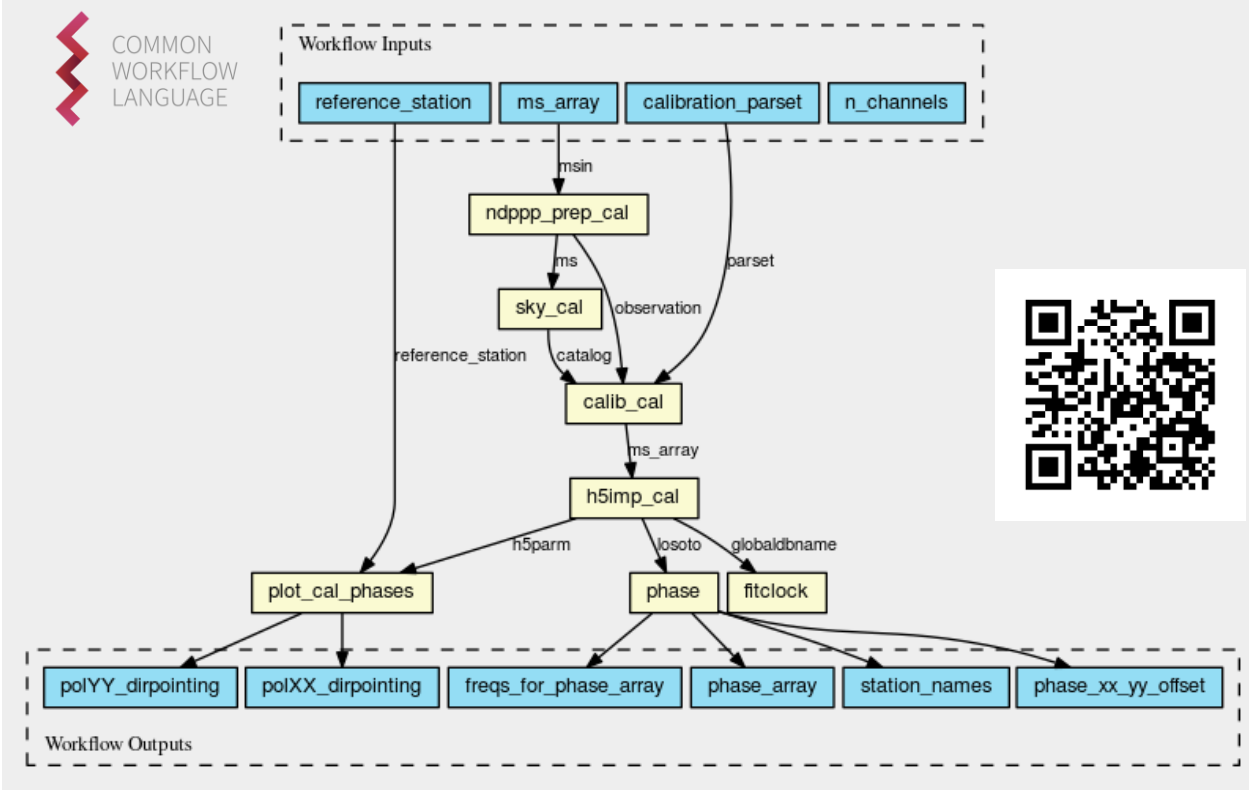
- implementation of Common Workflow Language (CWL) based pipelines
- make them deployable as Singularity containers
- run on various systems (data centers)

# LOFAR Science Demonstrator



## The Prefactor pipeline in CWL



# Science Demonstrator High Energy Physics

WLCG/DPHEP ORGANISATION: CERN CONTACT: Jamie Shiers

***OVERVIEW***:

The preservation of data from CERN's Large Hadron Collider poses significant challenges. The demonstrator shall show how existing, fully generic services can be combined to meet these needs in a manner that is discipline agnostic, i.e. can be used by others without modification.

***Demonstrator:*** Deploy services that tackle the following functions:

- Trusted / certified digital repositories where data is referenced by a Persistent Identifier (PID);
- Scalable "digital library" services where documentation is referenced by a Digital Object Identifer (DOI);
- A versioning file system to capture and preserve the associated software and needed environment;
- A virtualised environment that allows the above to run in Cloud, Grid and others.

Goal is to use non-discipline specific services combined in a simple and transparent manner (e.g. through PIDs) to build a system capable of storing and preserving Open Data at a scale of 100 TB or more.


The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563