The XDC project

Data Management for extreme scale computing

Daniele Cesini
Alessandro Costantini
Cristina Duma

eXtreme DataCloud is co-funded by the Horizon2020 Framework Program – Grant Agreement 777367

Outline

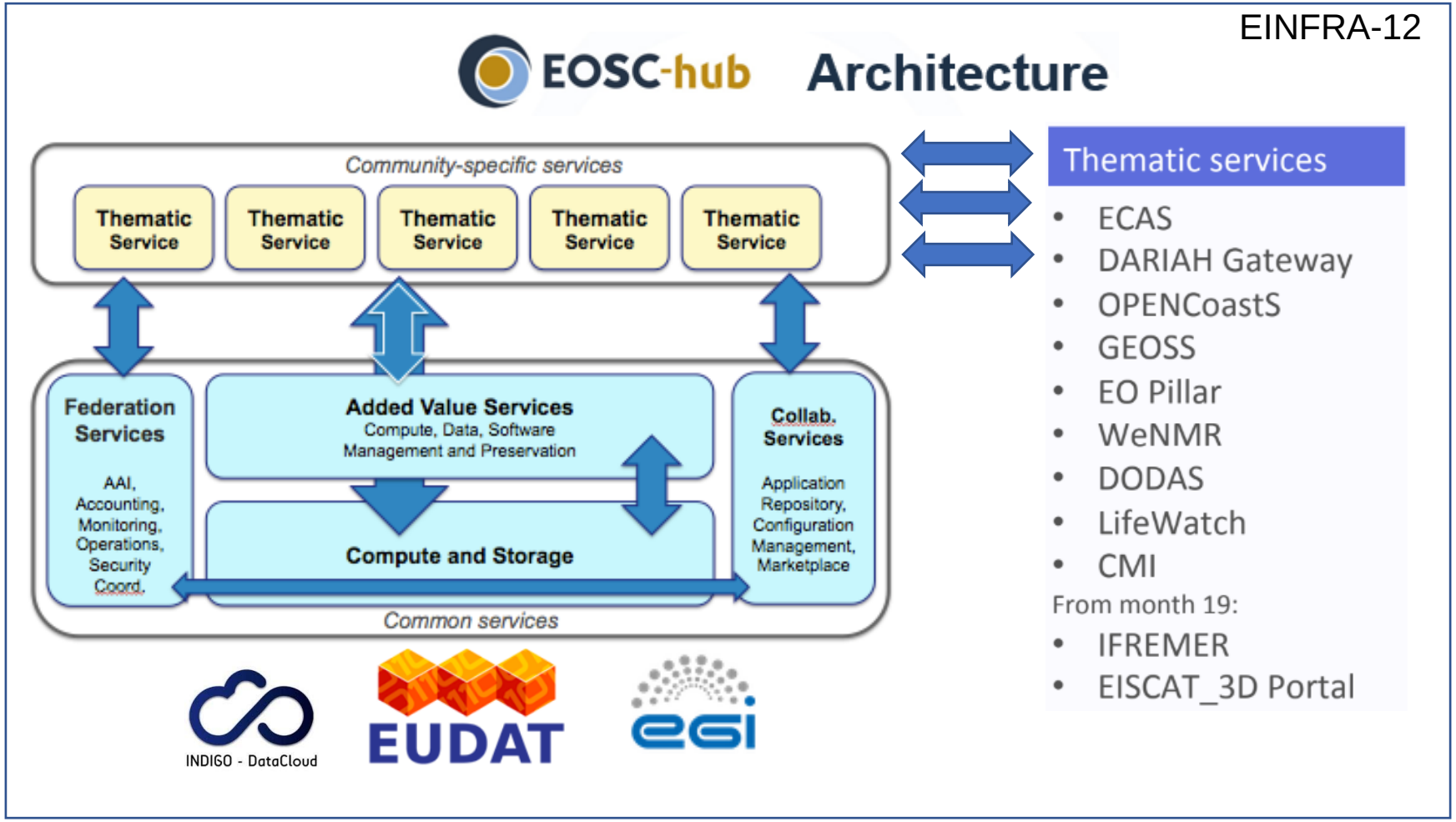
- The EOSC ecosystem
- XDC Introduction
- Overview of the XDC main Achievements
- Conclusion part1
  - Part2 at the end of the DEEP presentation

EOSC Ecosystem (back to 2017)…the oversimplified story

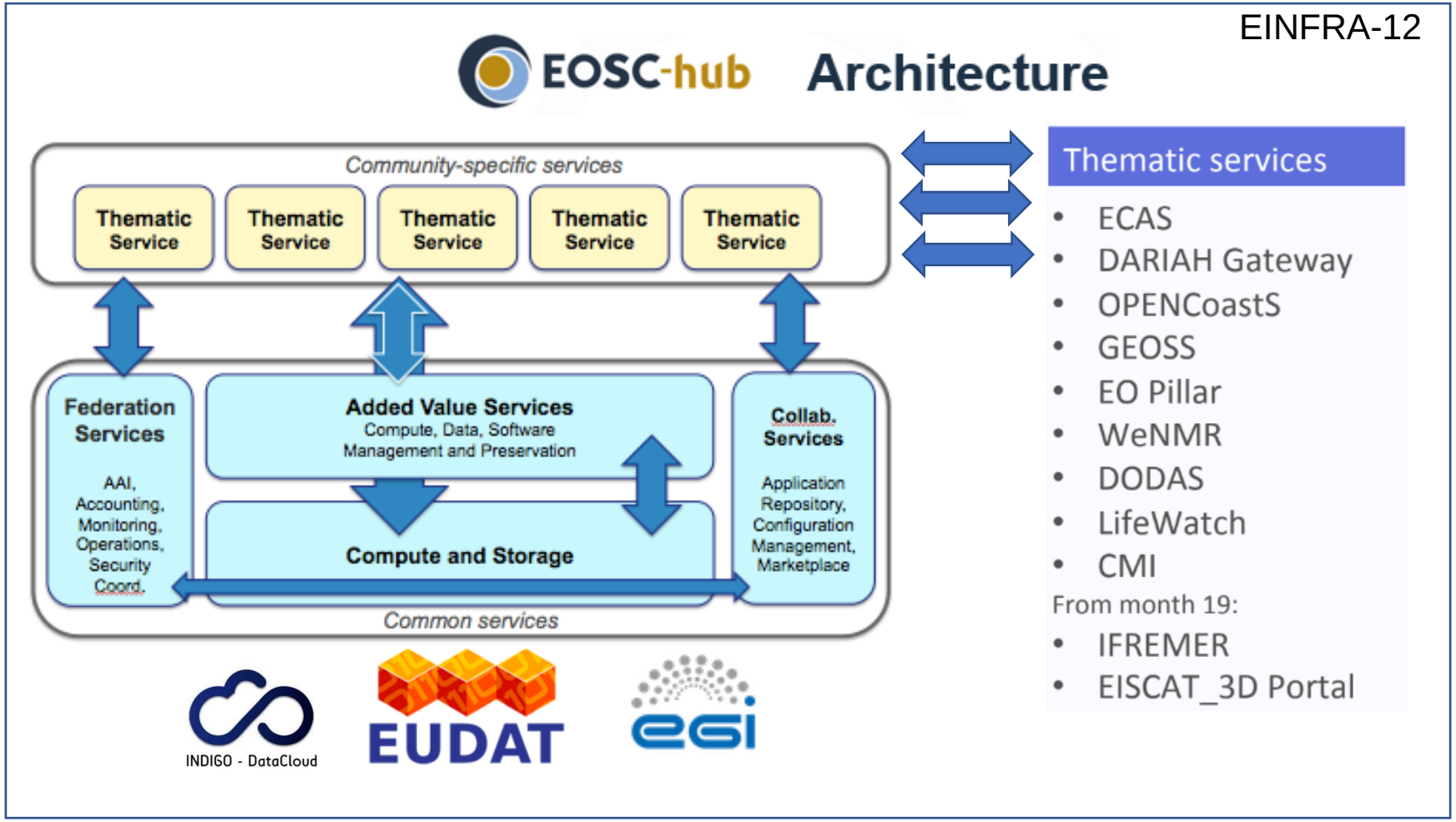
EOSC-hub mobilizes providers from European major digital infrastructures, EGI, EUDAT CDI and INDIGO-DataCloud jointly offering services, software and data for advanced data-driven research and innovation

- 100 Partners, 76 beneficiaries (75 funded)
- €33M total budget
- 36 months: Jan 2018 – Dec 2020

EINFRA-12



EOSC Ecosystem (back to 2017)…the oversimplified story



EOSC Ecosystem (back to 2017)…the oversimplified story

EINFRA-21

Hybrid DataCloud

eXtreme DataCloud

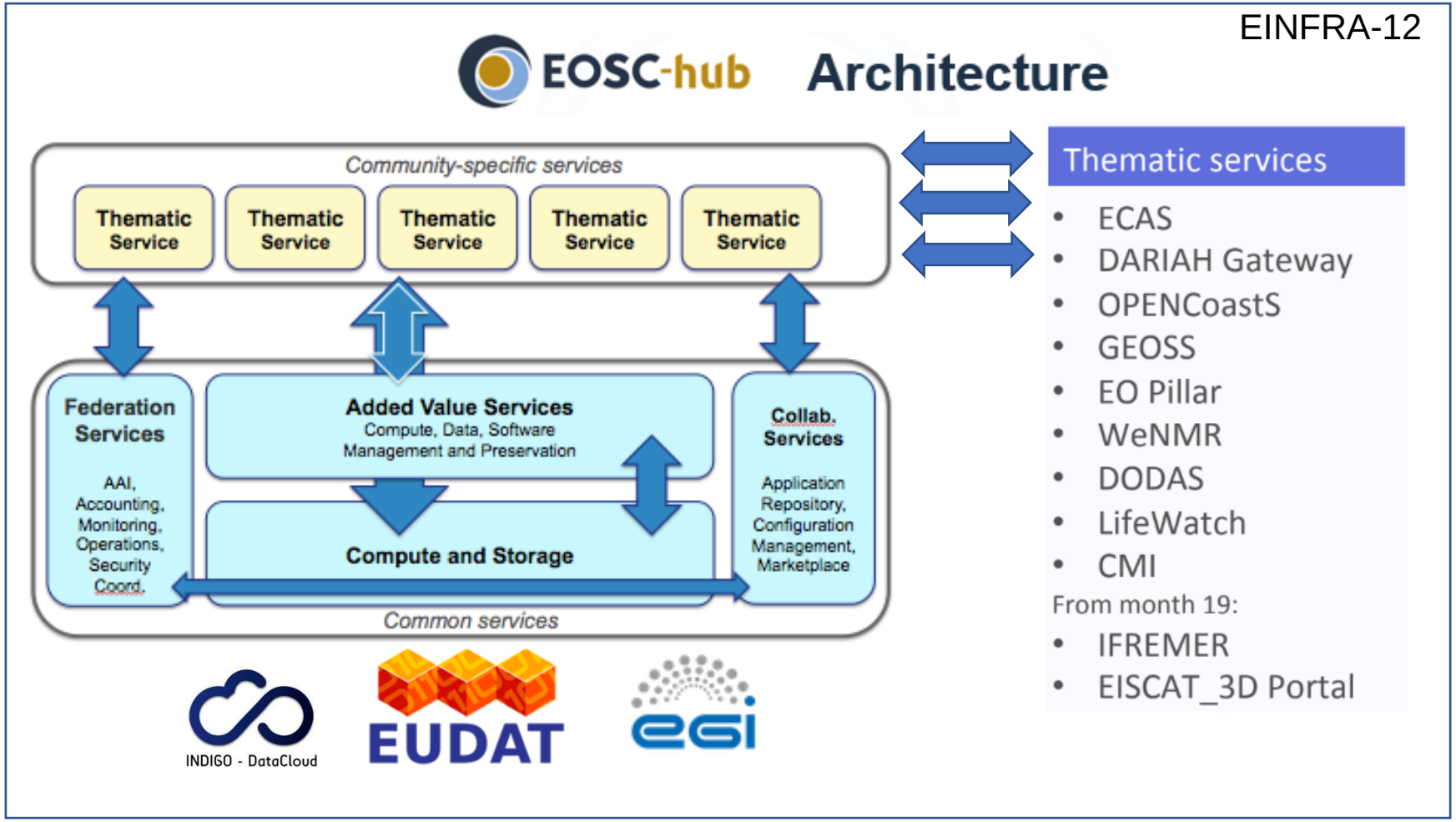
DARE

EUXDAT

PROCESS

FREYA

EINFRA-12



XDC

XDC Foundations

- **XDC take the move from**
  - the INDIGO-DataCloud Data management activity
  - the experience of the project partners on data-management

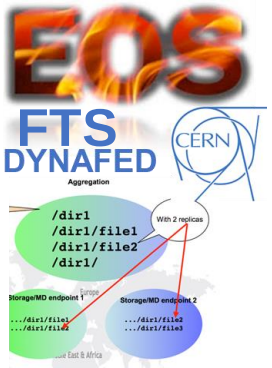
- **Improve already existing, production quality, Federated Data Management services**
  - By adding missing functionalities requested by research communities
  - Must be coherently harmonized in the European e-Infrastructures
  - **TRL 6+ → TRL8** (as requested by the H2020 EINFRA-21 call)

The Approach

- **Improve already existing, production quality Data Management services**
  - By adding missing functionalities requested by research communities
  - Based mainly on technologies provided by the partners and by the INDIGO-Datacloud project
  - Must be coherently harmonized in the European e-Infrastructures

INDIGO PaaS Orchestrator

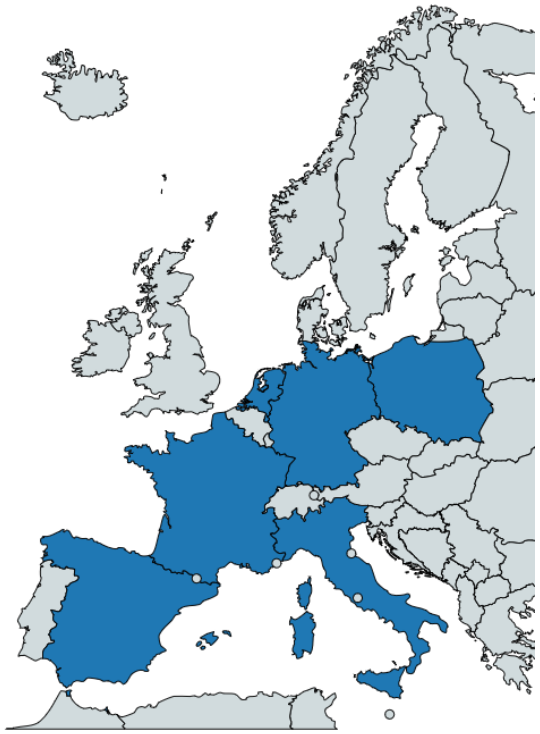
INDIGO CDMI Server



XDC Consortium

ID	Partner	Country	Represented Community	Tools and system
1	INFN (Lead)	IT	HEP/WLCG	INDIGO-Orchestrator
2	DESY	DE	Research with Photons (XFEL)	dCache
3	CERN	CH	HEP/WLCG	EOS, DYNAFED, FTS, RUCIO
4	AGH	PL		ONEDATA
5	ECRIN	[ERIC]	Medical data	
6	UC	ES	Lifewatch	
7	CNRS	FR	Astro [CTA and LSST]	
8	EGI Foundation	NL	EGI communities	

- 8 partners, 7 countries
- 6 research communities represented + EGI
- XDC Total Budget: 3.07Meuros

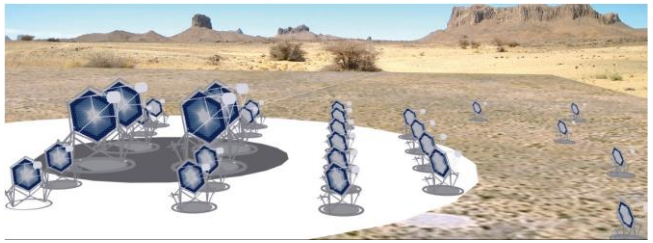


A User Driven Project

eXtreme DataCloud

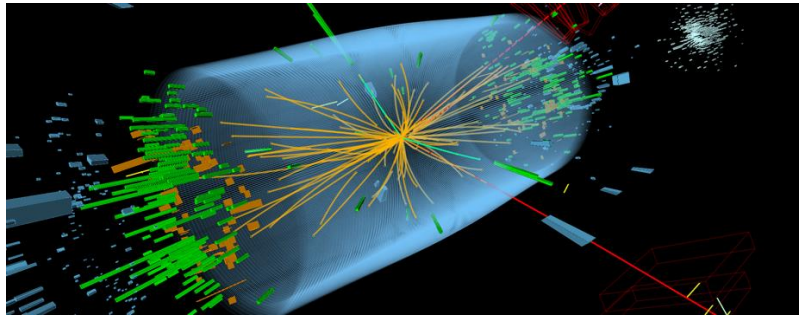
cta
cherenkov telescope array

LSST



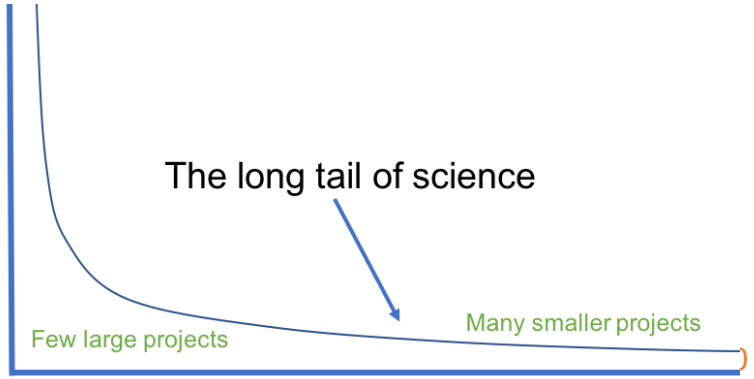
European XFEL

WLCG
Worldwide LHC Computing Grid

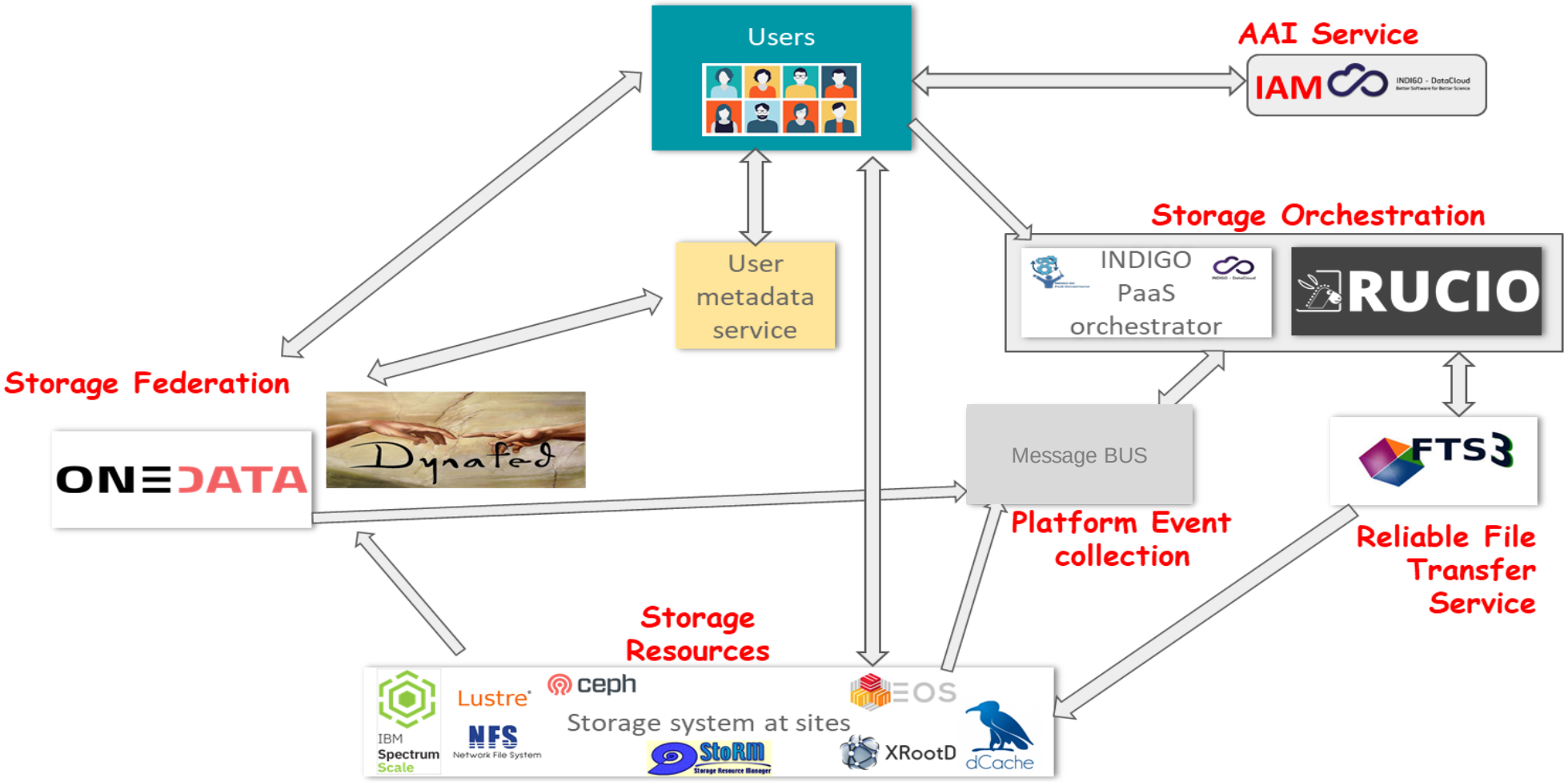


LifeWatch ERIC

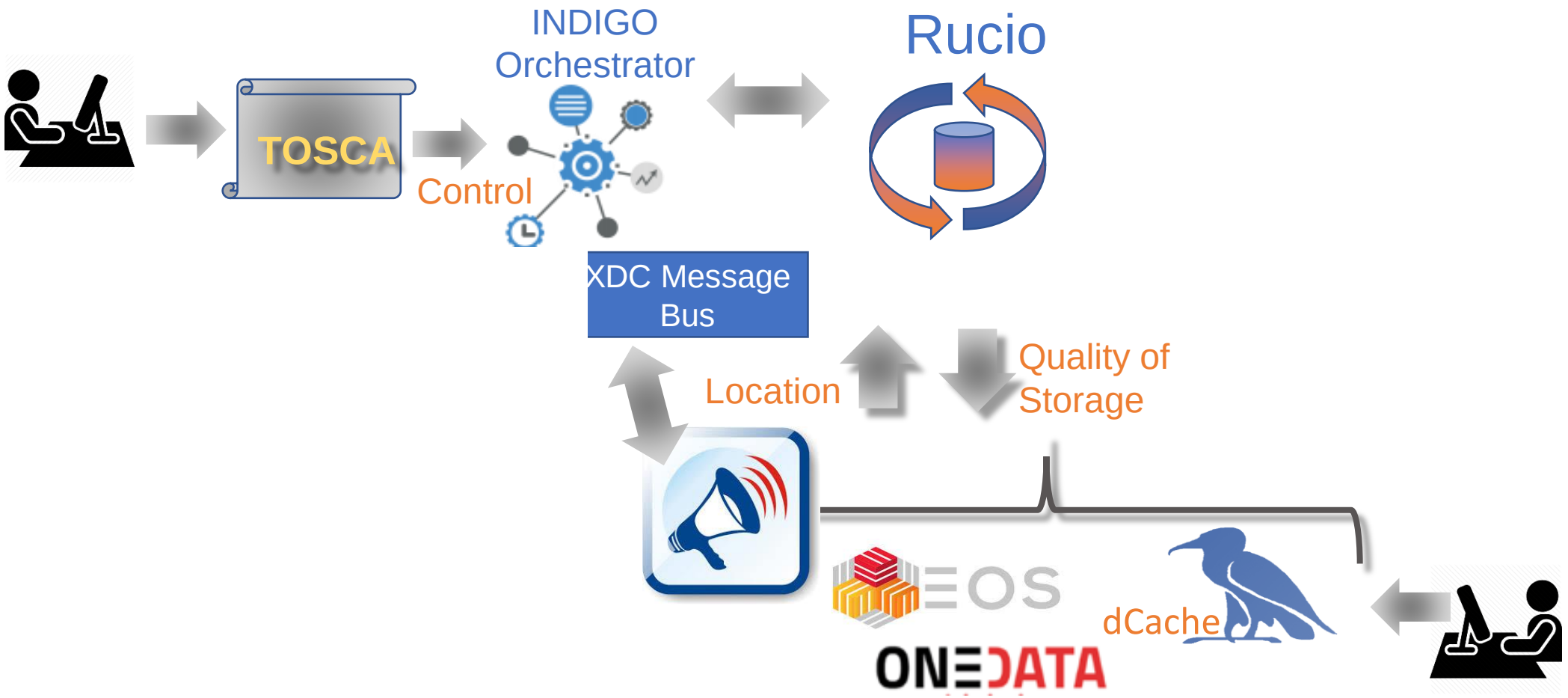
ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK



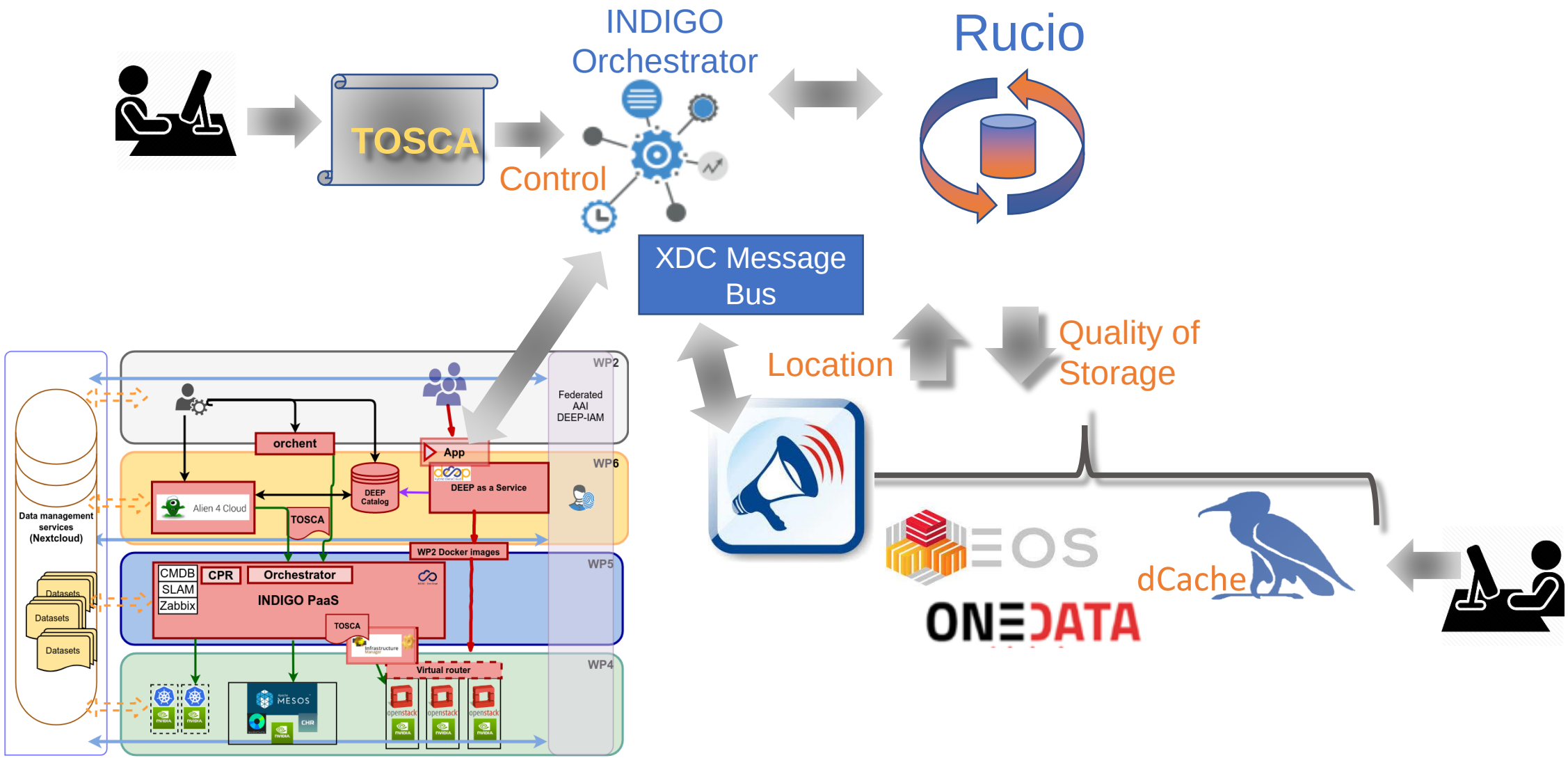
Architecture definition



Policy and event driven data management

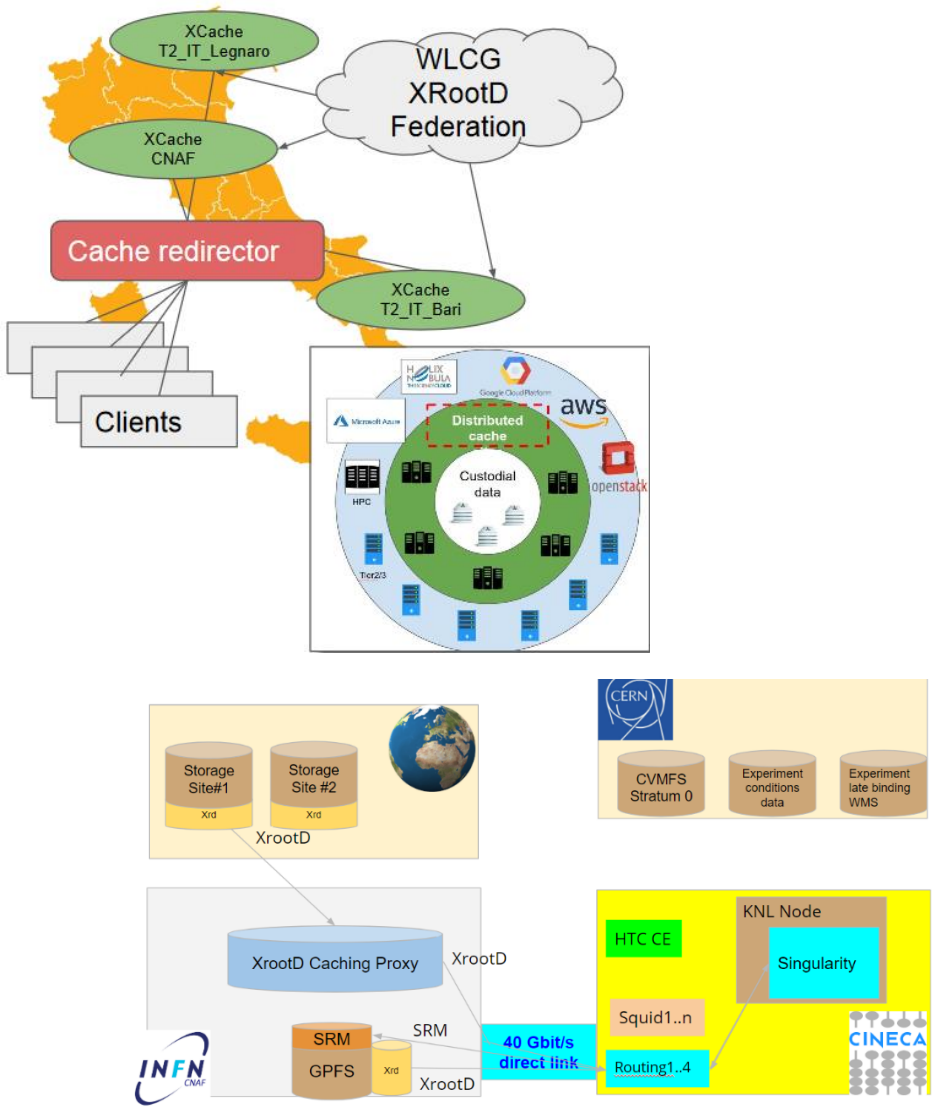


Policy and events driven data management

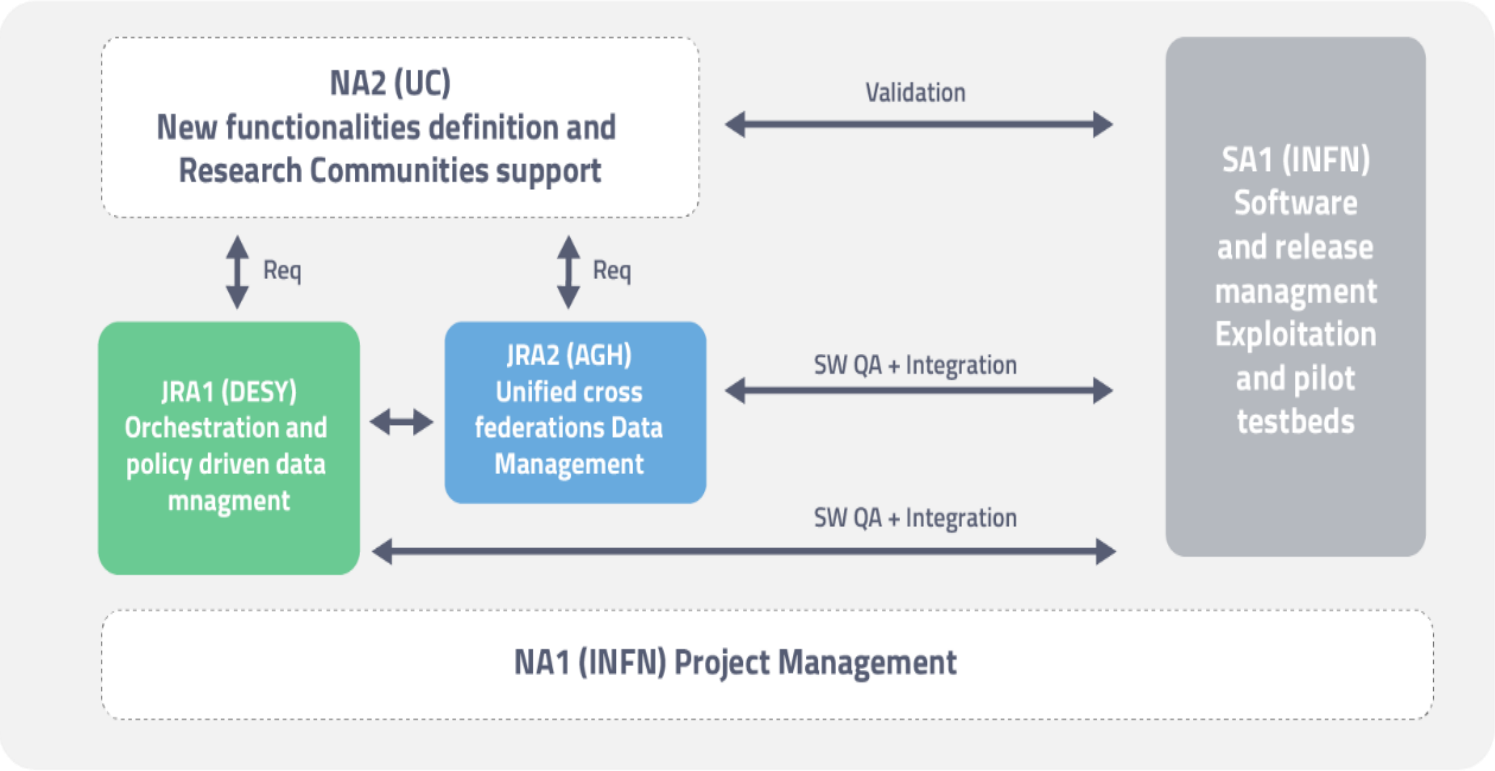


XDC Caching Solutions

- Deployment of Geo-distributed caches
- Network of unmanaged storage for hot data
- On-demand cache resources instantiation recipes
  - Both for XrootD and HTTP protocols
    - XCache and XRootD redirectors
    - NGINX caching system extend to support X.509 and OIDC authentication
    - DYNAFED used for HTTP cache federation



Project Structure



- Project structure
  - 5 WPs
  - 3 Management Bodies
    - Collaboration Board
    - Project Executive Board
    - Technical Coordination Board
  - Two bodies to coordinate the releases production and exploitation
    - Engineering Management Team
    - Service Providers Board

XDC Main Achievements

User communities requirements collection

- Adoption of “Champions”
  - To reduce the gap between the final researchers/users and the developers
  - Member of Research Communities/Infrastructures. “Person in the middle”
  - Understand the needs of the Use Case + general understanding of the available solutions and features. Technical - Scientific

- Agile approach

Use Case ➔ User Stories ➔ New functionalities ➔Technical Requirements ➔ Tests definition

High-level Functionalities Requested from User Stories

Topic	Requested Funtionality	Service(s) involved	Technical User Requirements from user stories (76 in total)
Policy Driven Data Management	Notifications and Monitoring Error handling	PaaS Orchestrator, dCache, Onedata	3 from LFW, 3 from CTA, 2 from XFEL 5 from ECRIN
	Archive Policy definition and management	PaaS Orchestrator, Onedata, dCache, FTS	8 from CTA, 3 from XFEL, 2 from CTA
	Archived data access	dCache, EOS, Rucio, FTS	1 from XFEL
Pre-processing, Processing during Ingestion	Job-like deployment analysis	PaaS Orchestrator, Rucio	2 from LFW, 1 from XFEL
	Data pre-processing workflows	PaaS Orchestrator, Rucio Onedata	1 from LFW, 1 from CTA, 1 from XFEL
	Data and Metadata Ingestion	PaaS Orchestrator	1 from LFW, 1 from XFEL
	Automated custom workflows (data-aware PaaS scheduling)	PaaS Orchestrator + storage systems	1 from LFW, 1 from XFEL, 3 from ECRIN
Smart Caching	Smart Caching	EOS, dCache, Dynafed, Caching-on-Demand, xdc-http-cache	4 from WLCG
Metadata and Data Life Cycle Management	Metadata Discovery and Data Access	Onedata dCache	2 from LFW, 1 from CTA, 1 from XFEL, 8 from ECRIN
		Onedata	1 from LFW, 1 from ECRIN, 1 from CTA
	Metadata attachment	Onedata	2 from LFW, 3 from ECRIN, 1 from CTA
	PID minting	Onedata	7 from ECRIN
	Metadata Management Interface	Onedata, PaaS Orchestrator	2 fom ECRIN, 1 from CTA
	Metadata Checking		
Data Security and Encryption	AAI based on OIDC	All	1 from CTA, 1 from XFEL
	Data encryption	Onedata	None

Highlights of XDC-1

Key technical highlights

- **OpenIDConnect support for token-based authentication**
- new QoS types integration and support in dCache, FTS, GFAL
- Orchestrator integration with other components
- Performance improvements in Onedata
- Support for groups and roles in Onedata
- EOS-dCache integration
- **Caching systems instantiation based on XROOTD protocol**
- **Storage events notification in dCache**
- EOS caching with XCache for geographic deployment
- EOS external storage adoption

https://releases.extreme-datacloud.eu/en/latest/releases/pulsar/index.html

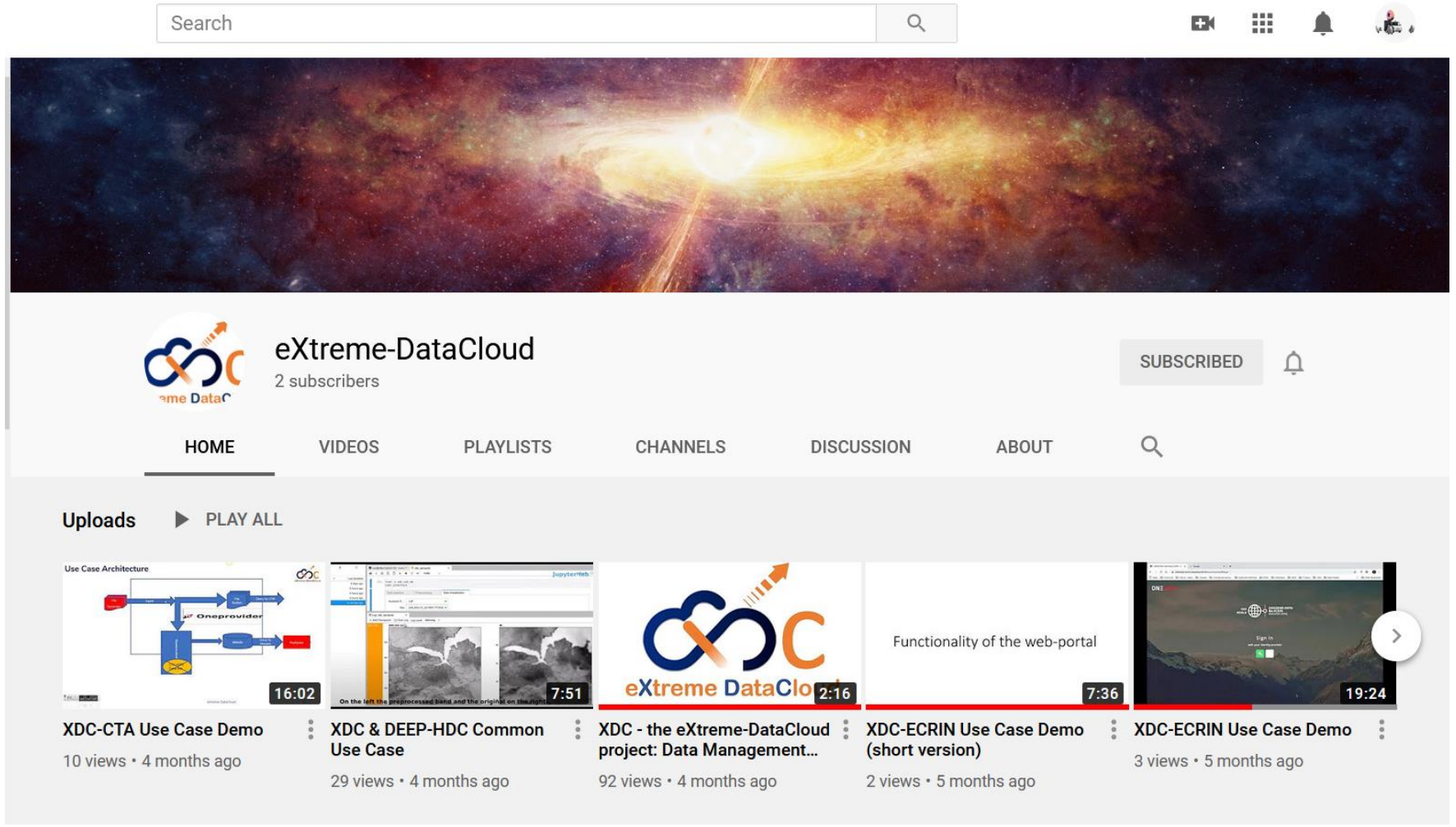
Highlights of XDC-2

- OpenIDConnect support for token-based authentication for all the components
- **CachingOnDemand extended to support the HTTP protocol**
- **QoS bulk transition support in EOS**
- Orchestrator integration with RUCIO
- “Fourth party copy” feature via DYNAFED
  - This allows services without third party copy support (such as S3) to participate fully in the data distribution infrastructure
- Improved support for Storage Events Notifications
- PaaS Orchestrator Dashboard released
- Python bindings for Onedata in a form of onedataFS
- Onedata performance improvement
- **Onedata MDR plugin to support the ECRIN use case**

https://releases.extreme-datacloud.eu/en/latest/releases/quasar/index.html

Some XDC Use Case Demos

https://www.youtube.com/channel/UCMnwdgI6YyqBfOABJE-qY1w



Software quality process and testbeds

- XDC defined a very rigorous process for ensuring software quality
  - Definition of roles
  - SQA Policies and Procedures
  - Development testbed
  - Integration testbed
  - Pilot testbed
  - Maintenance of Tools and Repositories

Strict Software Quality Assurance Process

1. Complete list of NEW services and components — DONE
2. Check source code availability — DONE
3. Check licensing information — DONE
4. Check CI status — DONE
5. Collect release notes — DONE
6. Documentation verification — DONE
7. Deployment automation verification — DONE
8. Pilot Preview deployment verification — DONE
9. Functional and Integration test reports — DONE

XDC - 2 (Quasar) updates - QC reports

Product	VCS		License	CodeStyle	Unit Testing (%)	Functional Testing	Docs	Code Review	Automated Deployment	Security	Artefacts
	code	tag									
dCache	✓	6.1.3	✓	★	✓ = XX	✓	✓	✓	✓	✓	rpm & deb ✓
PaaS Orchestrator	✓	2.4.0-FINAL	✓	✓	✓ = XX	✓	✓	✓	✓	✓	docker image: ✓

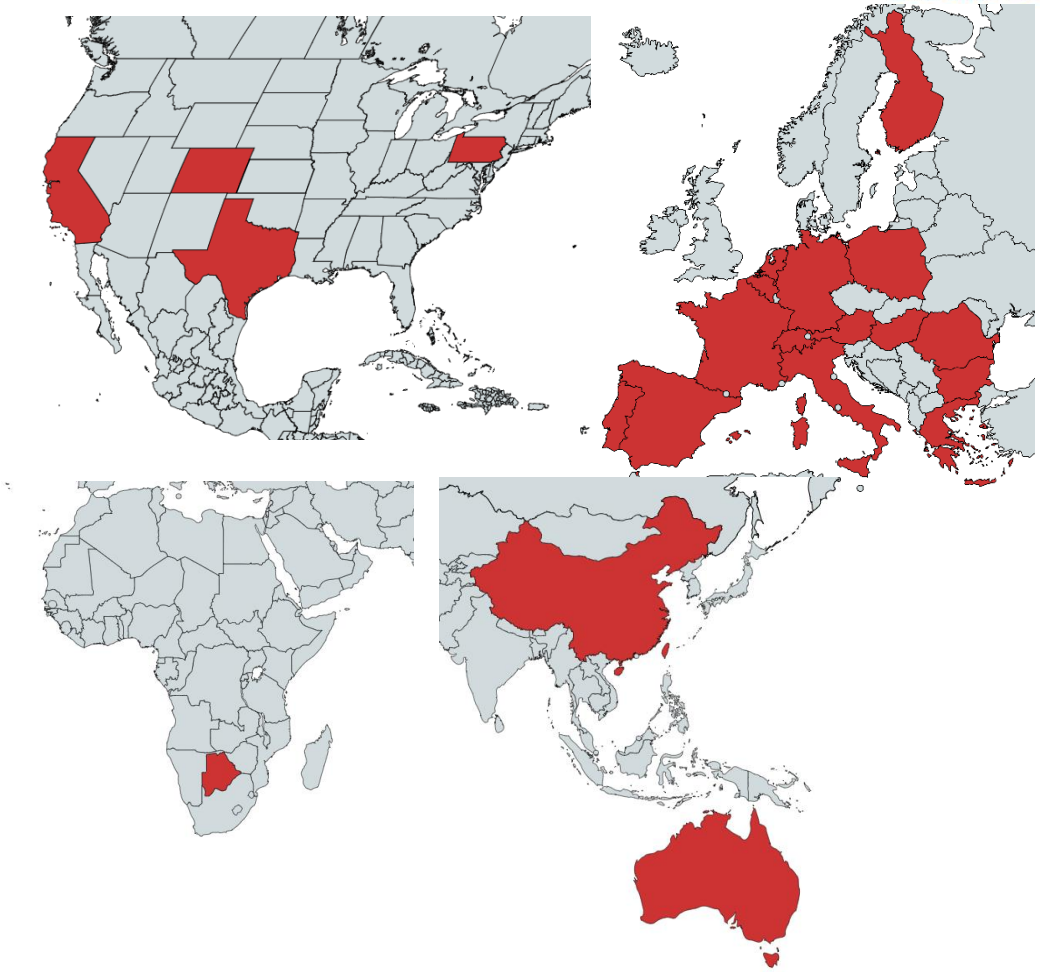
XDC - 2/Quasar QC reports

Product	VCS		License	CodeStyle	Unit Testing (%)	Functional Testing	Docs	Code Review	Automated Deployment	Security	Artefacts
CachingOnDemand	✓	1.0.0	✓	✓	✓	✓	✓	✓	✓	✓	docker image ✓
Dynafed	✓	1.3.3	✓	★	★ = 50%	✓	✓	✓	★ = manual	✓	rpms ✓
EOS	✓	4.4	✓	★	★ = 38%	✓	✓	✓	✓	✓	rpms ✓
Onedata	✓	19.02.1	✓	★	✓ = 70,7%	✓	✓	✓	✓	✓	rpm & docker image ✓
PaaS Orchestrator	✓	2.3.0-FINAL	✓	✓	✓ = 72,5%	✓	✓	✓	✓	✓	docker image ✓
PaaS Orchestrator Dashboard	✓	1.1.0	✓	✓	✓ = 75%	✓	✓	✓	★ = manual	✓	docker image ✓
Rucio	✓	1.22.0rc2	✓	✓	✓ = 90% (cover only XDC changes)	✓	✓	✓	★ = manual	✓	rpm & docker image ✓
TOSCA Types & Templates	✓	4.0.1	✓	N/A	N/A	★	★	✓	★ = manual	✓	tarballs ★

Dissemination & Communication Achievements

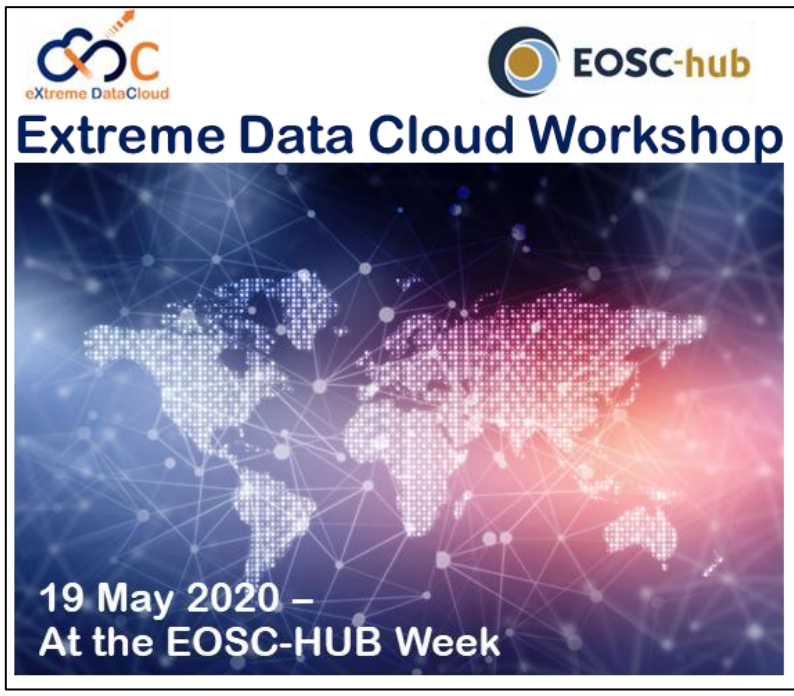
XDC Dissemination

- **18 contributions** to events in 2018
  - Talks, posters, presentation, ws sessions
- **26 contributions** to events in 2019
- **6 contributions** to events in 2020
- **9 events organized or co-organized**
  - 2 internal
  - 5 initiatives with other projects/communities
- **2 Training Events**
  - **1 summer course** organized with DEEP in Santander in 2018
    - "New challenges in Data Science: Big Data and Deep Learning on Data Clouds"
  - **1 training session** at the EOSC-Hub Week 2019
- Participation to the **Common Dissemination Booster** with DEEP and INDIGO-DC
- **9 XDC related scientific papers**

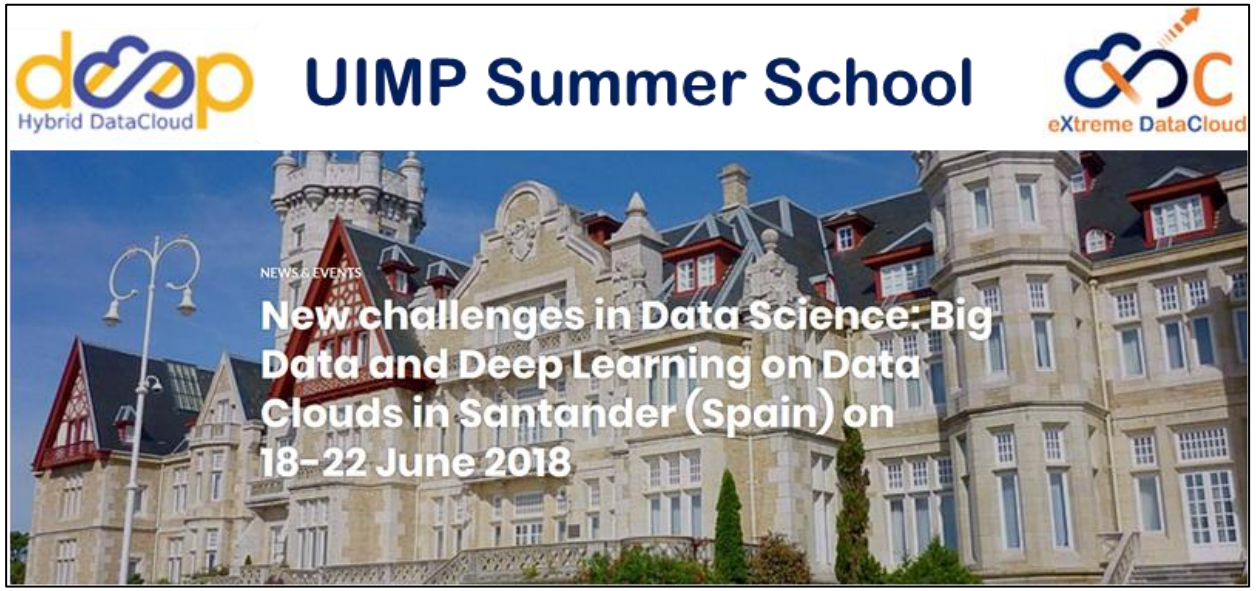


http://www.extreme-datacloud.eu/dissemination/

Events organized by the project



Training Events co-organized



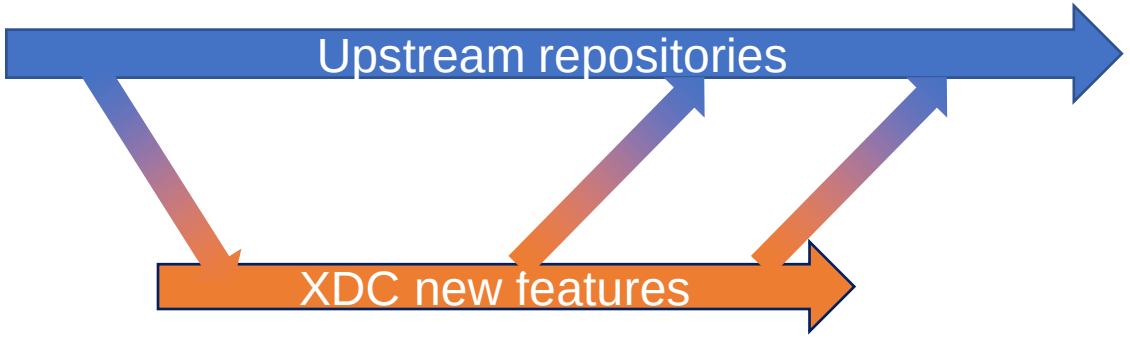
Project Communication

- XDC in Social Media
  - Website
    - http://www.extreme-datacloud.eu/
  - Twitter account
    - https://twitter.com/XtremeDataCloud
  - Youtube Channel
    - https://www.youtube.com/channel/UCMnwdgI6YyqBfOABJE-qY1w
  - LinkedIn group
    - https://www.linkedin.com/groups/12181004/
- Service Catalogue Brochure
  - http://www.extreme-datacloud.eu/wp-content/uploads/2020/05/sc_xdc.pdf
- Promotional video
  - https://www.youtube.com/watch?v=E8CAjnqqj8k
- XDC branded Gadgets

Exploitation & Sustainability Achievements

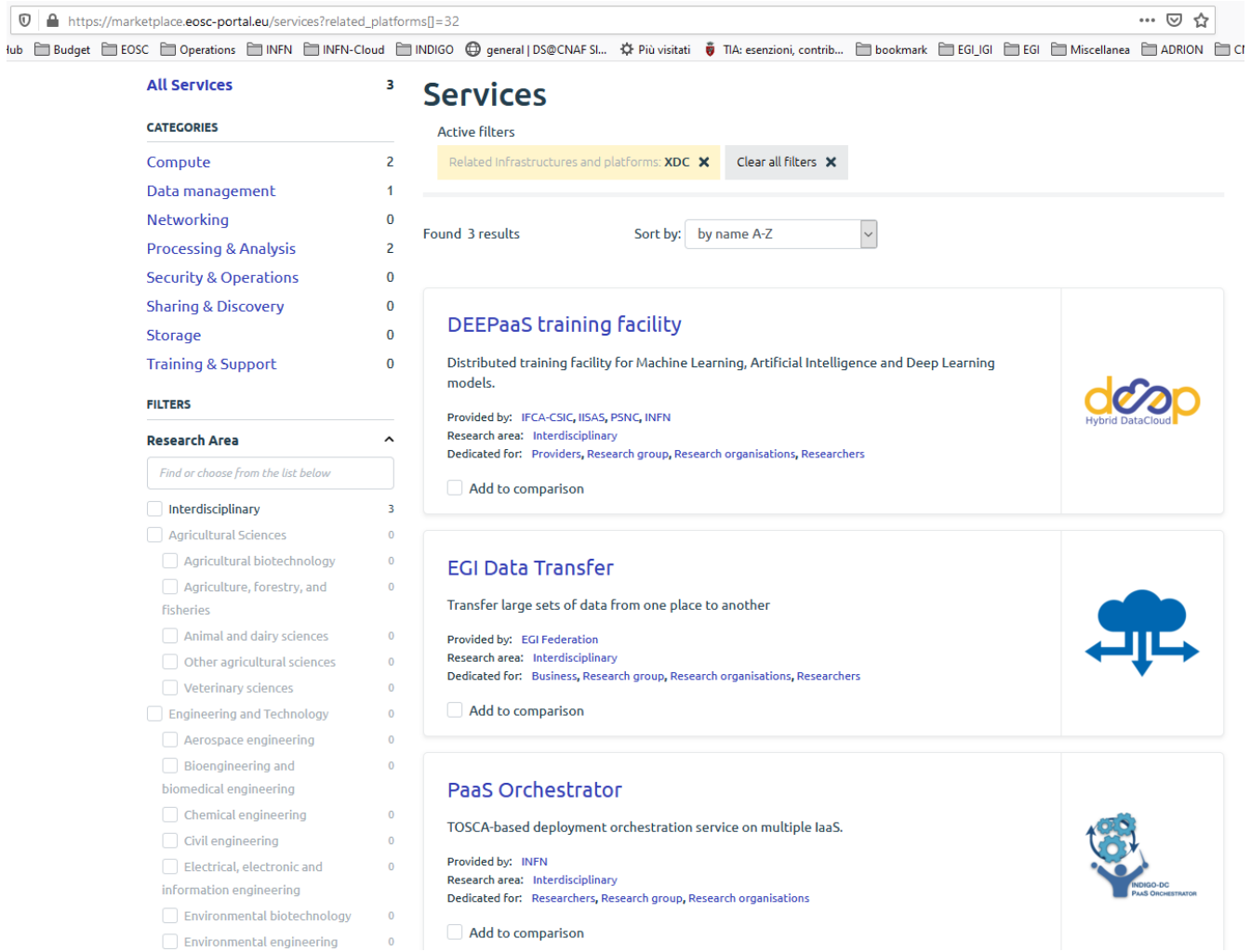
Exploitation and sustainability strategy

- The Service Providers Board (SPB) goal is to link XDC with Service Provider within and outside of the project consortium in order to have a regular dialog among XDC and all the Service Providers
  - Internal members + 10 external service providers
- Making XDC products available through EGI distribution channels
  - UMD release already delivers dCache, FTS, GFAL, XRootd
- Identification and interaction with the EOSC-HUB Communities and competence centers
- Pushing developments in the upstream repositories of all the services to ensure sustainability beyond the project
  - Double path
    - XDC repositories
    - Upstream repository



Promoting XDC software via EOSC marketplace

- It is only possible to register service endpoints in the EOSC Marketplace
- Nevertheless 3 services supported by XDC software are already published
  - EGI Data Transfer (DataHub)
  - PaaS Orchestrator
  - DEEPaaS training facility
- XDC is linked via a TAG in the search mask



Piloting activities with external use cases

- **EOSC-Hub Marine Competence Centre**
  - implementation of pilots for SeaDataNet activities
- **EOSC-Hub Fusion Competence Centre**
  - evaluation and report on Data Replication and Access testing using Onedata
- **Photon and Neutron** in the context of the PaNOSC project
  - various presentations have been made and piloting activities are being implemented
- **Earth Observation, ASTRON, ESCAPE**
  - joined a workshop on data management to present their use cases and evaluate possible solutions
- **LOFAR, EISCAT_3D, the Project MinE, the Tropomi project, The African Rainfall Project, the "Plankton, Aerosol, Cloud ocean Ecosystem" (PACE) project**
  - All are using or experimenting with dCache storage events notifications
- **Workshops with the other EINFRA-21 initiatives**
  - Common use case addressed with DEEP

MDR for the COVID-19 crisis

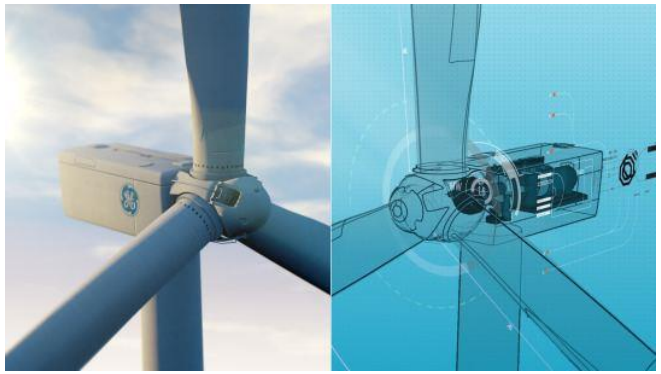
- **MDR exposed to communities as a production service for the ECRIN task force on COVID-19, accessible from the ECRIN webpage** (https://www.ecrin.org/clinical-research-metadata-repository)

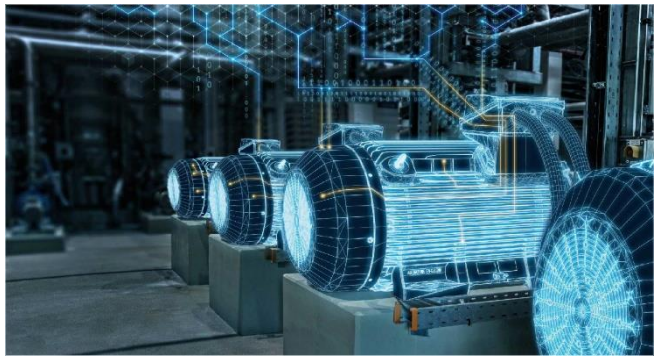
- **MDR presented in the discussion about the implementation of the European COVID-19 Research Data Platform**

- **MDR proposed for inclusion in the Infection Diseases Data Observatory (IDDO)**

Sustainability via other H2020 projects

- XDC is providing solutions to build the computing infrastructures of other important H2020 initiatives
  - IoTwins
    - Computing architecture for the creation **digital twins for industries and facilities management**
      - H2020-ICT-2018-2020, 16M€ EU contribution, 23 partners, started Sept. 2019, 36 months
  - ESCAPE
    - Creation of distributed e-infrastructures and datalakes for astroparticle communiites
      - Data management for extreme scale experiments
      - INFRAEOSC-2018-2, 16M€ EU contribution,30 partners, s tarted in Jan 2019 - 42 Months
  - EOSC-Life
    - MDR further developments





Effort, Budget & Co.

Effort, Budget & Co.

- 3.07Meuro Total EC Contribution
  - INFN
    - 583k (CNAF, BA, PD, NA, PG) from EU
  - CNAF
    - 235k workforce
    - 25k Other direct costs
- 50k contribution from CCR
  - CNAF: 30k

- 385 PMs Total
  - INFN
    - 80 PMs
  - CNAF
    - 47 PMs

WP	INFN PMs	CNAF PMs	Notes
WP1	16	14	WP1 Leader: Costantini Project Coordinator: Cesini Technical Coordinator: Donvito
WP2	8	4	WLCG Champion: Falabella, Ciangottini
WP3	27	17	WP3 Leader: Duma
WP4	24	10	T4.3 Leader: Donvito
WP5	5	0	

CNAF workforce

Name	Effort (avg)	WPs
Cesini D.	40%	1,2
Costantini A.	20%	1,2,3
Dell'Agnello L.	15%	1,2
Duma C.	20%	3
Falabella A.	25%	2,4
Fattibene E.	10%	4
Fornari F.	80%	4
Grandi G.	10%	1
Michelotto D.	20%	3
Morganti L.	20%	1,2
Salomoni D.	15%	1,2
Vianello E.	20%	3,4

Conclusion

Conclusions (the good ones)…

- XDC has been a complex project (despite the small consortium)
  - From both the coordination and technical perspectives
    - Many interacting products, many development teams
    - Many constraints (from the EC call, from the user communities, from the target infrastructures, from legacy technologies, etc.)
- However, it passed two EC reviews with excellent results:
  - *“Project has delivered exceptional results with significant immediate or potential impact”*
  - We showed that all the foreseen objectives have been achieved
  - Among the praised achievements:
    - the WP3 work on Software Release Management and Software Quality Control
    - The huge effort spent on Dissemination and Exploitation of the technical results
- It contributed to the improvement of several widely used tools in the current e-Infrastructures
  - Including INFN developed services (INDIGO-PaaS-Orchestrator, Caching-on-Demand, HTTP-Caches)
- It funded several fixed-term contracts of young researchers
  - Contributed to their professional development and internationalization
    - New technologies, new working environment
- Contributed to keep close relationships with other important actors and partners in the e-infra game
- Another tile of the INDIGO-DC and INFN participation to the development of the e-Infrastructures in Europe

…however…

- …more thoughts at the end of the today presentations….