Future Architecture for ESGF

Philip Kershaw

Centre for Environmental Data Analysis, STFC

The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Motivation

- Ten years since the original system was designed: a need to re-architect ESGF in the light of
  - Growth and diversification of use– number of projects and disciplines it supports
  - New technologies and changing paradigms for access and analysis of data
  - Challenge of legacy software - operations and maintenance of system

The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Process

- Initiated through ESGF Executive Committee
- Presented ideas at 2018 F2F
- Dedicated meeting in November in UK
- Report compiled: https://bit.ly/esgf-fut-arch-report
- Priorities identified and roadmap set

The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Rewind: High-level Vision and Objectives

1) To provide access to climate data primarily *MIP activities[+]

2) Address the challenge of hosting and access to large volume climate datasets through the provision of a distributed system of collaborating data provider organisations linked together in a federation

[+]ESGF may support access to other Earth science data but the core community to support is *MIP

The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

High-Level System Requirements

1) Provide access to large volume *MIP data
2) Federate access to data across participating data providers in a federation
3) Data providers (see definition in following section) retain control of data (not obligated to publish to whole federation)
4) Data providers must maintain an agreed level of operational service in order to integrate with the federation
5) **Prioritise the provision of a very robust core of functionality for the infrastructure which satisfies a basic set of functions and services well**
6) **Provide a stable API to facilitate developers and third parties integrating their applications and services with the ESGF infrastructure**
7) Provide a good design which will inspire others to adopt it
8) **Adopt community standards e.g. cataloguing, metadata to enable easy integration with third party applications, systems and infrastructure**
9) Provide a baseline reference web-facing user interface to facilitate access to data for the user community
10) Support Open access and FAIR Data principles

Gather more requirements downstream from science community (e.g. IPCC). Are we focusing on the right areas (e.g., data analytics, compute)? Can we achieve those requirements?

The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Actors and Stakeholders

Data providers

Modelling centres; Satellite-based data services; Data workflows; Sensor/observational data; Other data services

Analysis users

Big data: require a large computing facility;
Medium data: node or a modest cluster
Small data: can do the work on their laptop
Software as a service / applications developers

Infrastructure/data service providers and operators

Sysadmin / network admin / infrastructure
ESGF stack administrators

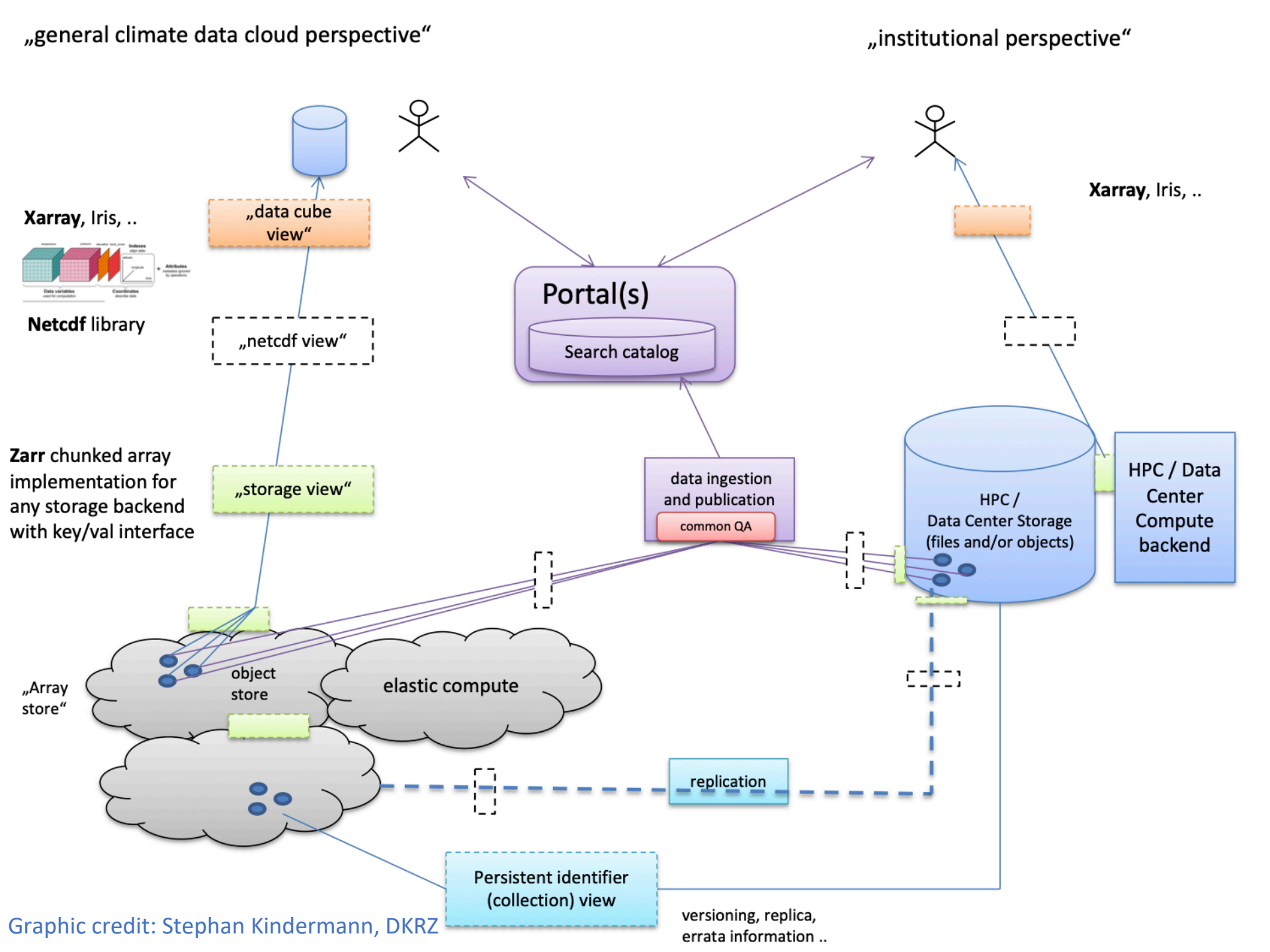
Publishers - publish data into ESGF

The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Review Landscape

- Projects, programmes, initiatives
  - Pangeo
  - ESA Thematic Exploitation Platforms
  - EO Exploitation Platform Common Architecture
- Hosting – on-premise and commercial cloud
- Search, cataloguing: Intake, ESM Collection, STAC, OpenSearch
- Data access: Xarray, zarr, object storage
- DevOps: Ansible, Containers, Kubernetes

The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084



Graphic credit: Stephan Kindermann, DKRZ

Four aspects to analysis

1) User Experience

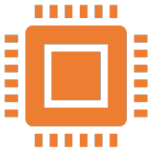
2) Data Repository and management

3) Compute on data

4) Platforms and system administration

The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Findings



Platforms and systems administration:

Define interfaces, keep functionality modular

Use of standards for interoperability

Embrace new technologies for DevOps, infrastructure-as-code approach – Containers, Kubernetes, …

Search services

Centralise search – single index aggregating all content

Integrate search with other services such as ES-Doc, PID service

New search technologies to explore OpenSearch, STAC, Intake-esm

New modes for data access and storage

Client-side aggregation model

Analysis ready data or data caching model?

UKRI Science and Technology Facilities Council Natural Environment Research Council

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Findings (2)

ID Management and Access Entitlement

Less important with open access for data but use cases to secure compute services

Use new architectural patterns: AARC Blueprint

New standards assist: OpenID Connect

Critical need to update and simplify

Compute services

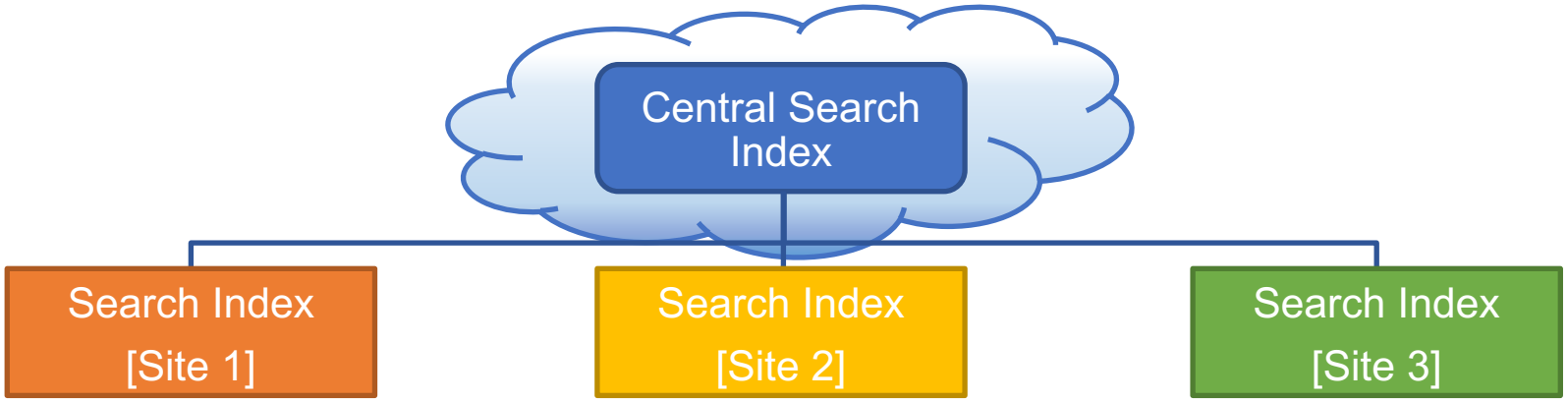
Important but no consensus for ESGF-wide standard offering yet

Metrics Collection

Leverage advances in industry with standard tooling to exploit - Prometheus and InfluxDb, Grafana

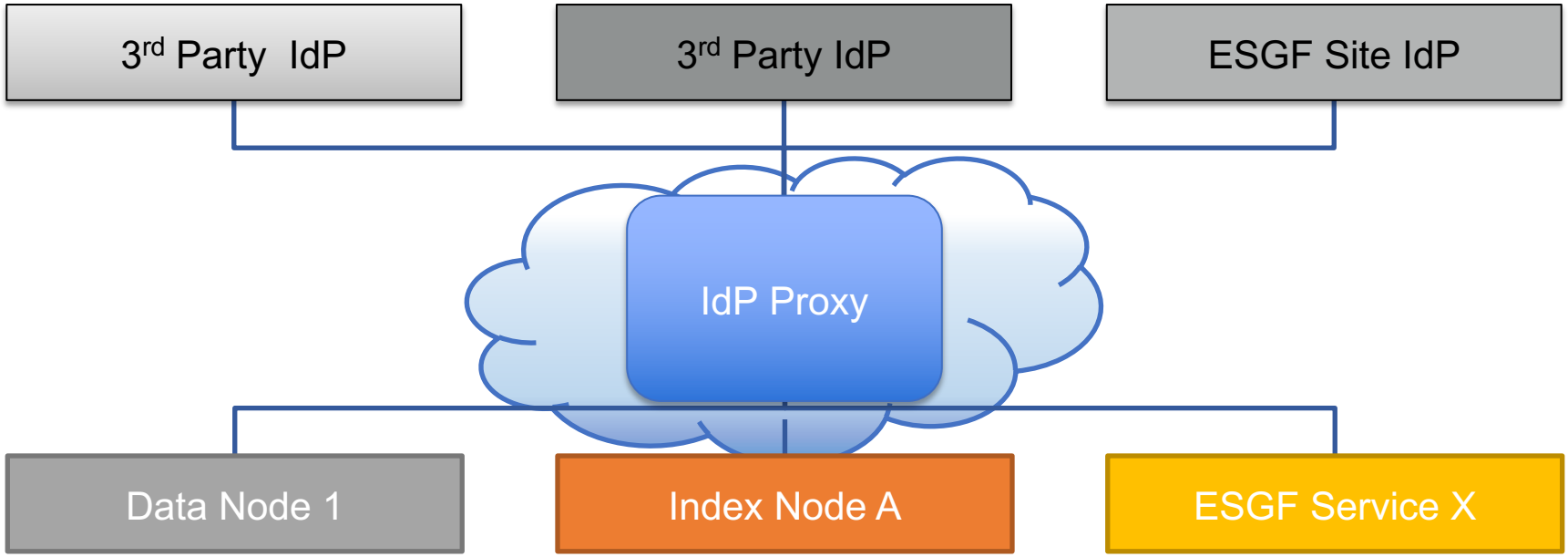
The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Centralised Search Index



The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Identity Provider (IdP) Proxy



The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Prioritisation

- Strong consensus to progress installation and deployment architecture
  - Concentrate on making a robust core to the system
  - Importance of modular approach: interfaces critical to achieve this

- Identity management and access control system in urgent need of update
  - e.g. OpenID 2.0 was deprecated a long time ago!

The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Near-term Roadmap: what practical steps, when?

- What:
  - Container-based deployment
  - New data node with Nginx for file serving, TDS reserved for OPeNDAP only
  - Simple static catalogue for TDS
  - OpenID Connect with Keycloak for ID management
  - Search and publishing changes for later release

- When:
  - Prototype deployment by March(!)
  - Initial production version ready by June
  - Further features in follow-up releases

The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

More information

- Future Architecture Report:
- https://bit.ly/esgf-fut-arch-report

The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084