Digital research data in the Sigma2 prospective

NARMA Forskningsdata seminar

30. Januar 2018

Maria Francesca Iozzi, PhD, UNINETT/Sigma2

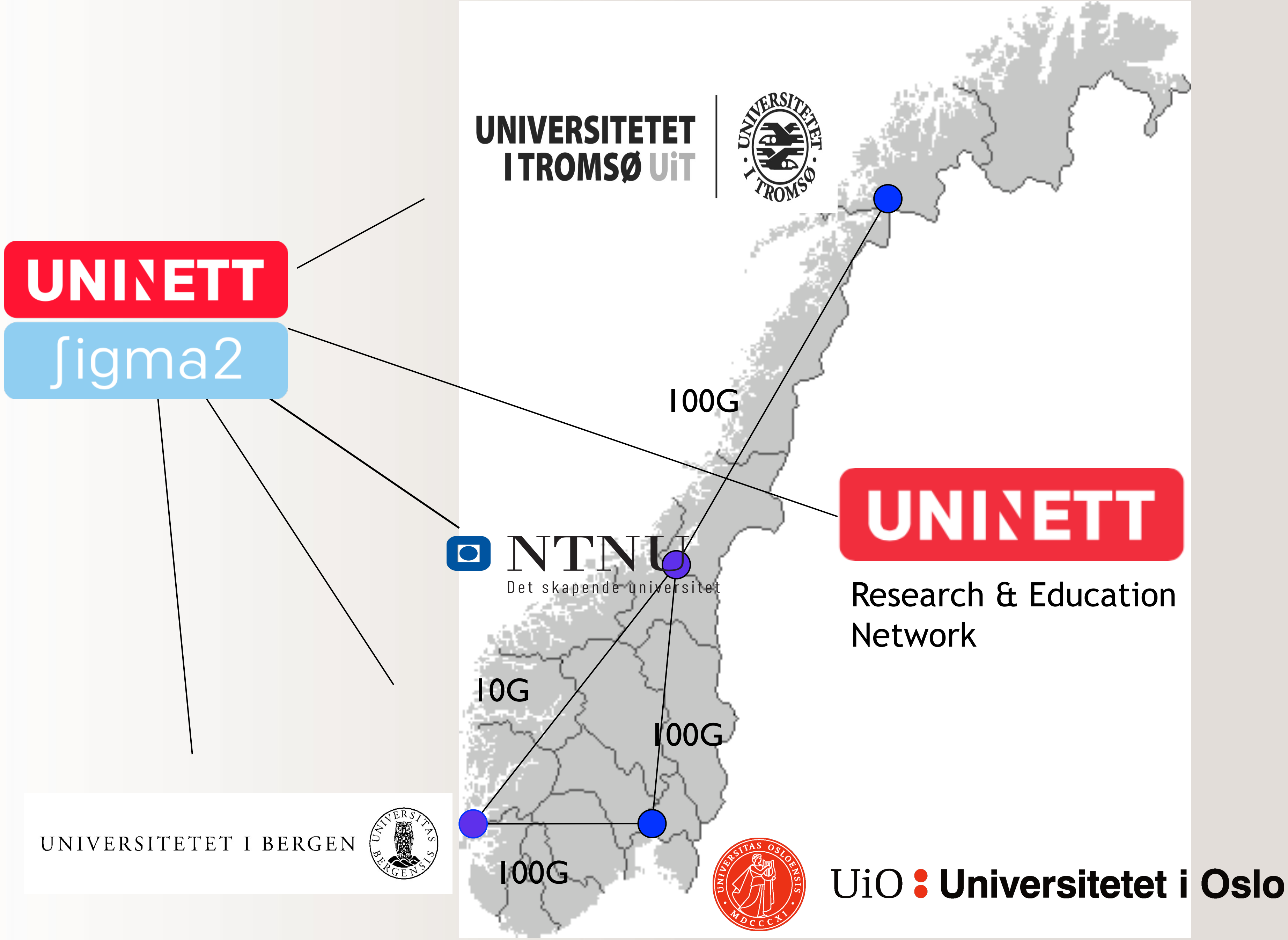
Hans A. Eide, PhD, UNINETT/Sigma

Agenda

- About UNINETT Sigma2
- Research data
- Sigma2 e-Infrastructure Services:
    - DMP
    - Storage
    - Analysis and Computing
    - Archiving
    - Advanced user suppport
- Get on board!

National e-infrastructure - a very brief history

- From the beginning, it was always recognized that e-infrastructure, just like other research infrastructure, should be shared.
- Early on, research institutions competed for basically the same funding and established disconnected e-infrastructure resources.
- In the early 2000's, the need for coordination and sharing lead to the establishment of UNINETT Sigma and the Metacenter. Universities still competed for the same funding and had their own hardware resources, no common strategy.
- **In December 2014, the 4 major universities (UiB, UiO, UiT, NTNU) and the Research Council of Norway (RCN) decided to establish UNINETT Sigma2 and collectively operate the national e-infrastructure.**



UNIVERSITETET I TROMSØ UiT

UNINETT

Sigma2

NTNU
Det skapende universitet

UNINETT
Research & Education Network

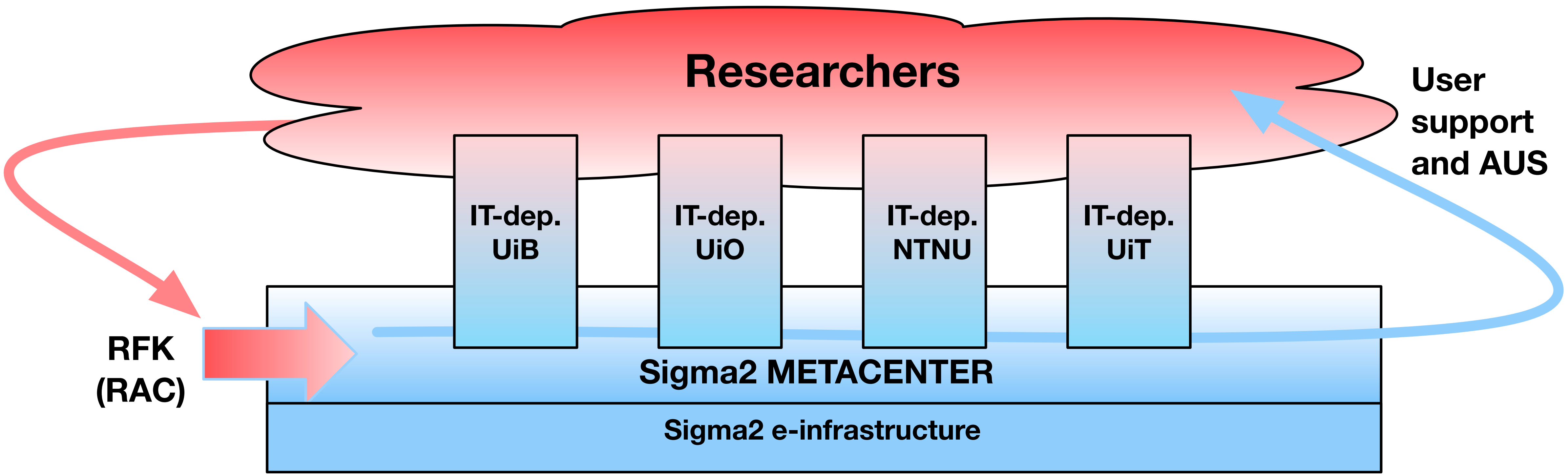
UNIVERSITETET I BERGEN

UiO : Universitetet i Oslo

Sigma2’s high level objectives

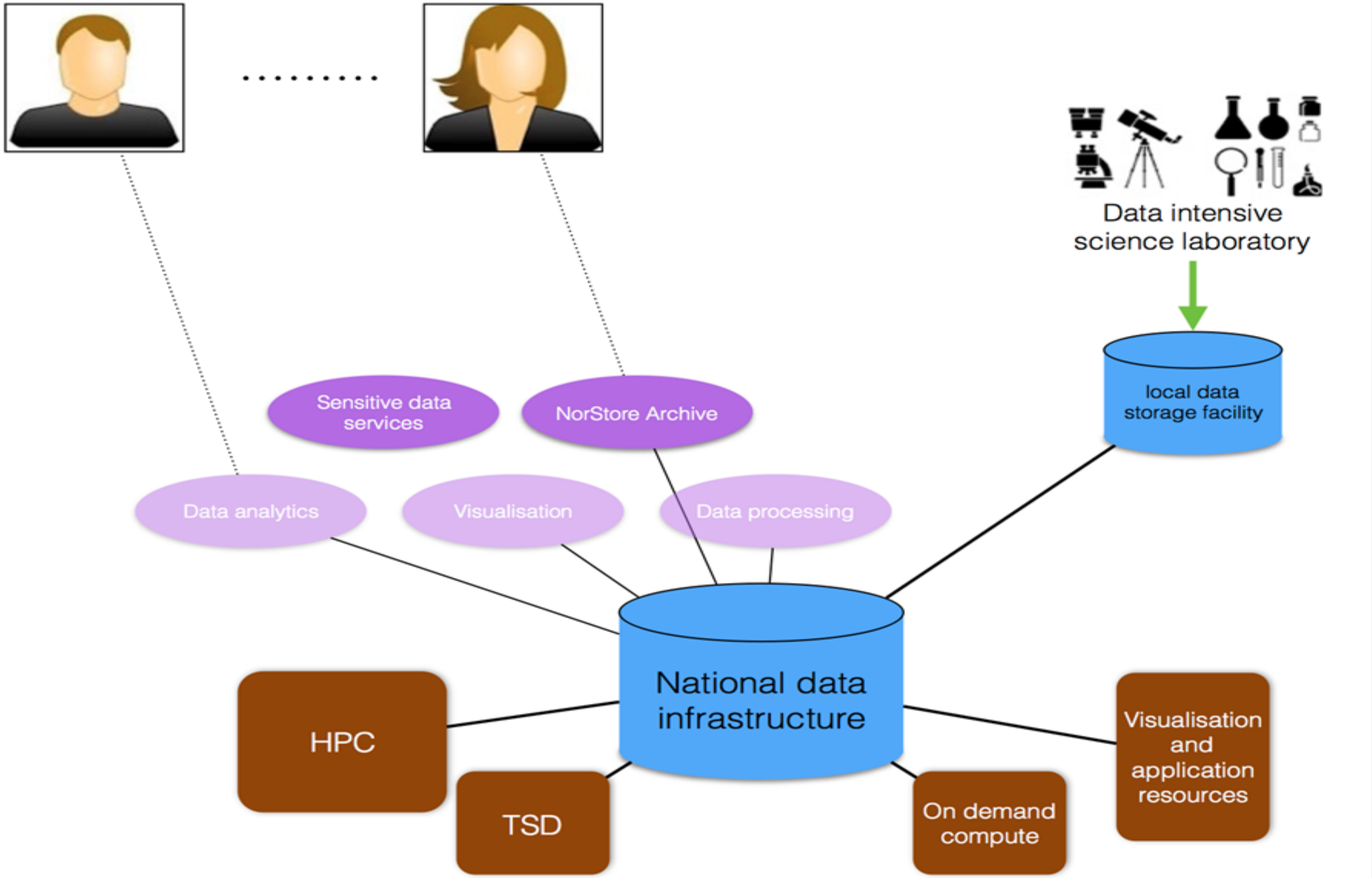
- Procure, operate and develop a critical national e-infrastructure
- Promote e-infrastructure to new research communities
- Lead and coordinate participation in international cooperation for e-infrastructure
- Provide an attractive and sustainable e-infrastructure for all research communities, with the following characteristics:
  - High reliability and availability
  - Cost effectiveness
  - Predictable access
  - Interoperability within the national e-infrastructure and between national and international infrastructures (e.g. PRACE, EUDAT)
- Provide services for data analytics of large datasets (Big Data)

The Metacenter



- National coordination and shared, consolidated resources have cost and efficiency advantages but creates a “distance” to the end-users (researchers)
- This is countered by keeping the support staff and competence near where the research is going on, at the universities
- Combined with a **data-centric** architecture for the e-infrastructure, this model combines the advantages of the centralized model and the local model

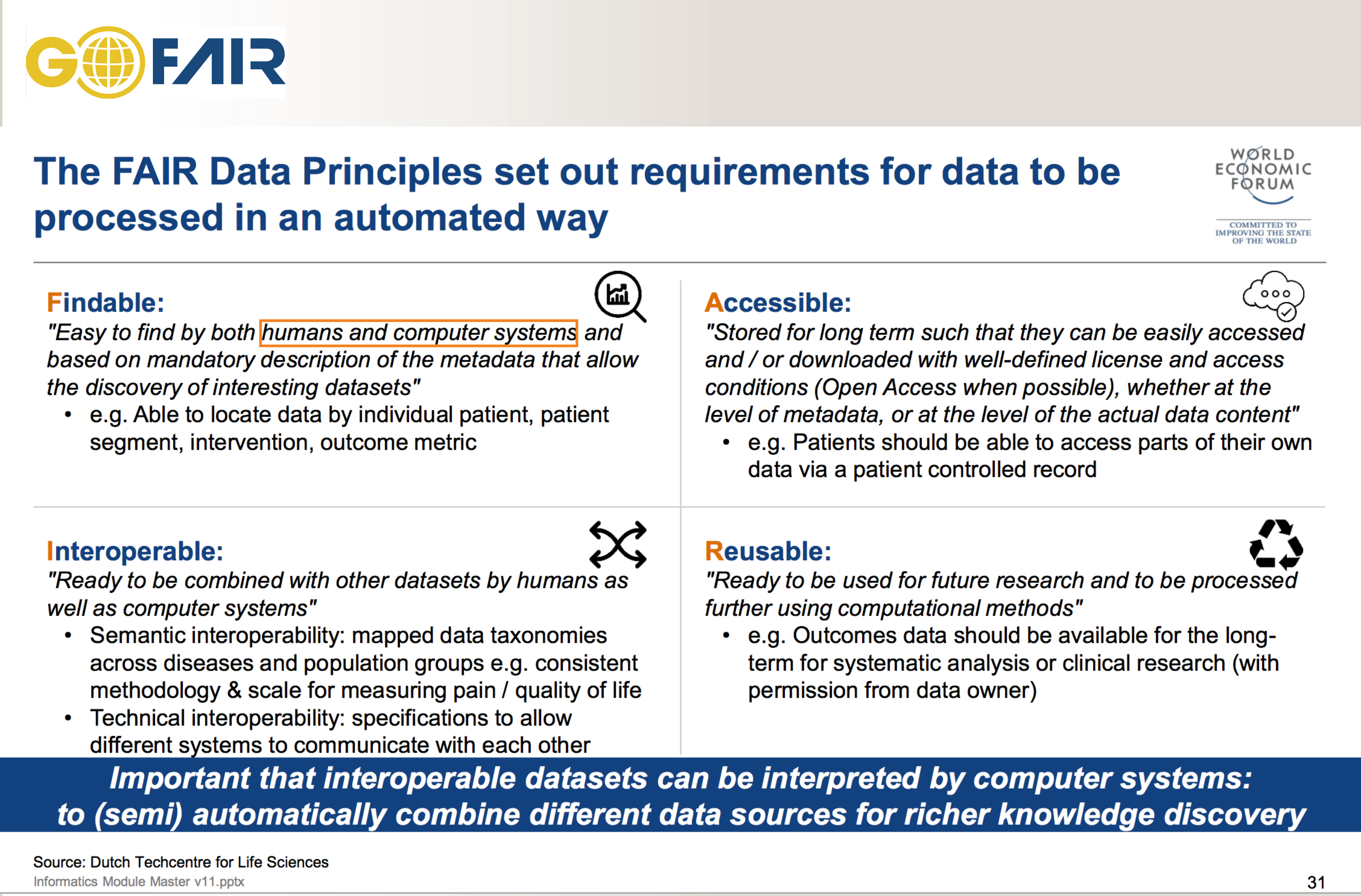
Data-centric architecture


Data intensive science laboratory
local data storage facility
Sensitive data services
NorStore Archive
Data analytics
Visualisation
Data processing
National data infrastructure
HPC
TSD
On demand compute
Visualisation and application resources

In summary

The core mission of UNINETT Sigma2 is to provide services that researchers need today, e.g. advanced user support, training, data services such as storage, archive, data management tool, data analytics (Big Data) and high performance computing (HPC), that all together facilitate research, FAIR use of data and the collaboration among research communities.

Research data



The FAIR Data Principles set out requirements for data to be processed in an automated way

Findable:

"Easy to find by both humans and computer systems and based on mandatory description of the metadata that allow the discovery of interesting datasets"

- e.g. Able to locate data by individual patient, patient segment, intervention, outcome metric

Accessible:

"Stored for long term such that they can be easily accessed and / or downloaded with well-defined license and access conditions (Open Access when possible), whether at the level of metadata, or at the level of the actual data content"

- e.g. Patients should be able to access parts of their own data via a patient controlled record

Interoperable:

"Ready to be combined with other datasets by humans as well as computer systems"

- Semantic interoperability: mapped data taxonomies across diseases and population groups e.g. consistent methodology & scale for measuring pain / quality of life
- Technical interoperability: specifications to allow different systems to communicate with each other

Reusable:

"Ready to be used for future research and to be processed further using computational methods"

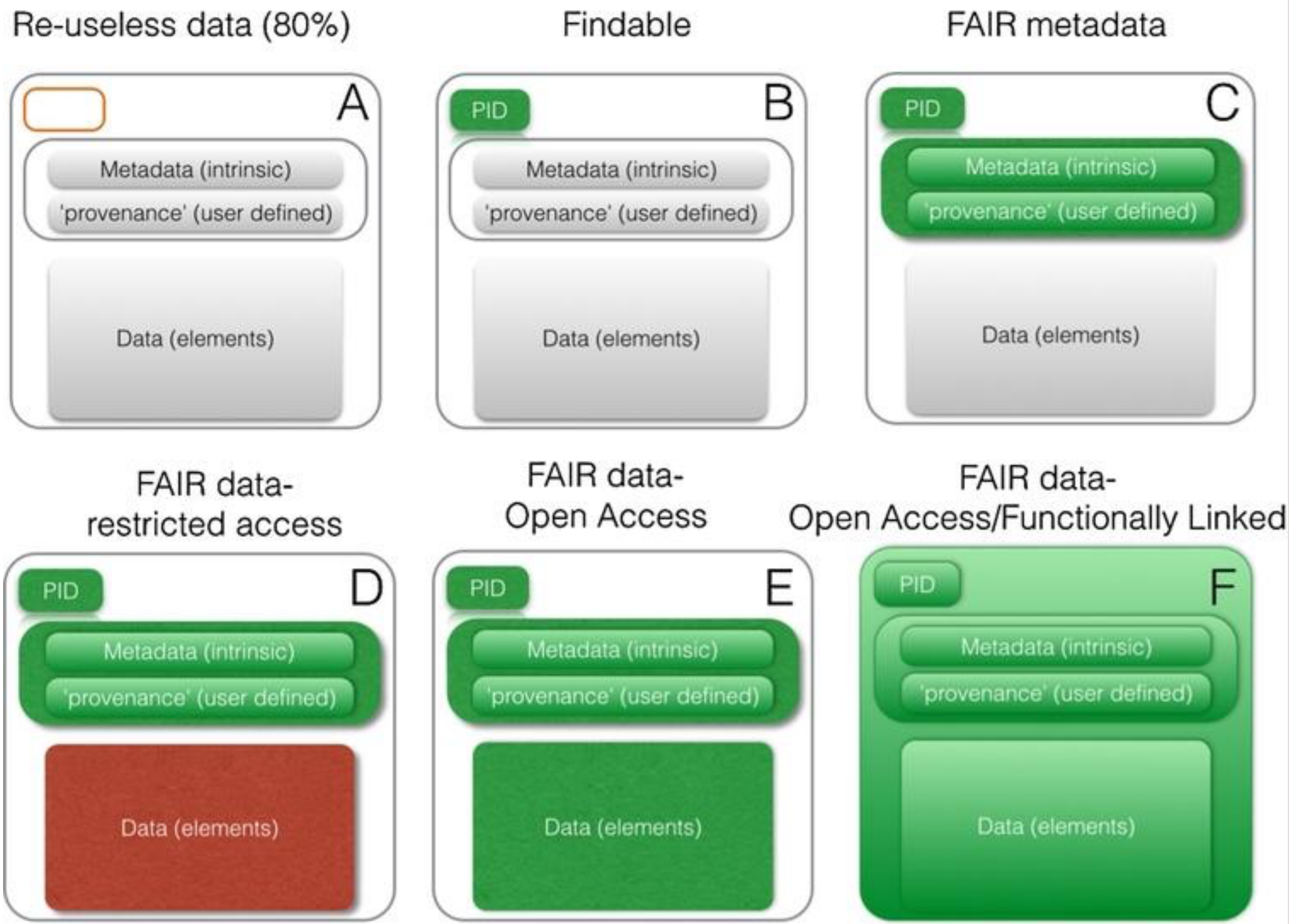
- e.g. Outcomes data should be available for the long-term for systematic analysis or clinical research (with permission from data owner)

Important that interoperable datasets can be interpreted by computer systems: to (semi) automatically combine different data sources for richer knowledge discovery

Source: Dutch Techcentre for Life Sciences

Courtesy of Barend Mons, GoFAIR

Data as increasingly FAIR Digital Objects

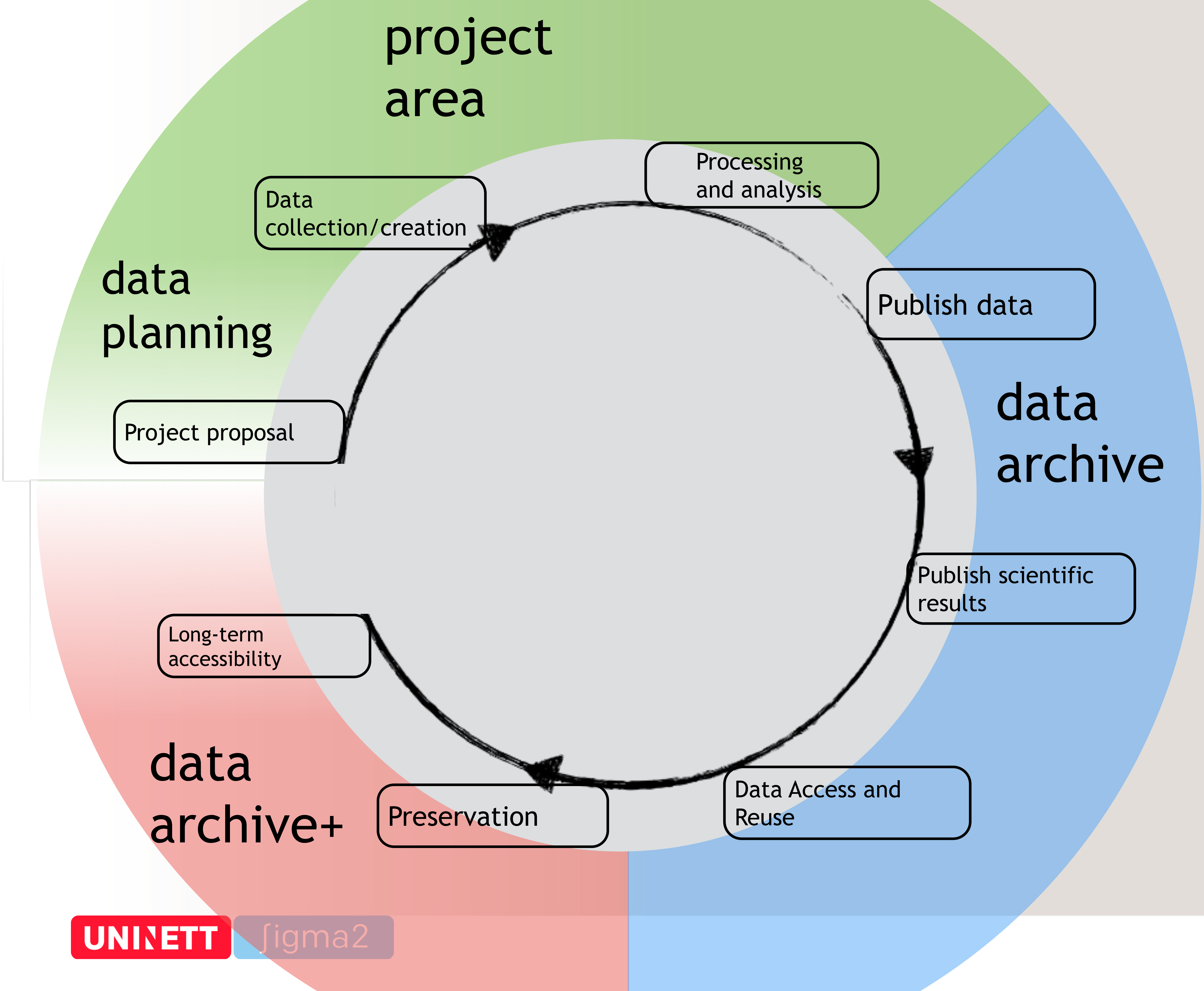


Courtesy of Barend Mons, GoFAIR

Metadata – essence for research data

- A must-have for credible research data

nometadata.org



project area

data planning

data archive

data archive+

Data collection/creation

Processing and analysis

Publish data

Publish scientific results

Data Access and Reuse

Preservation

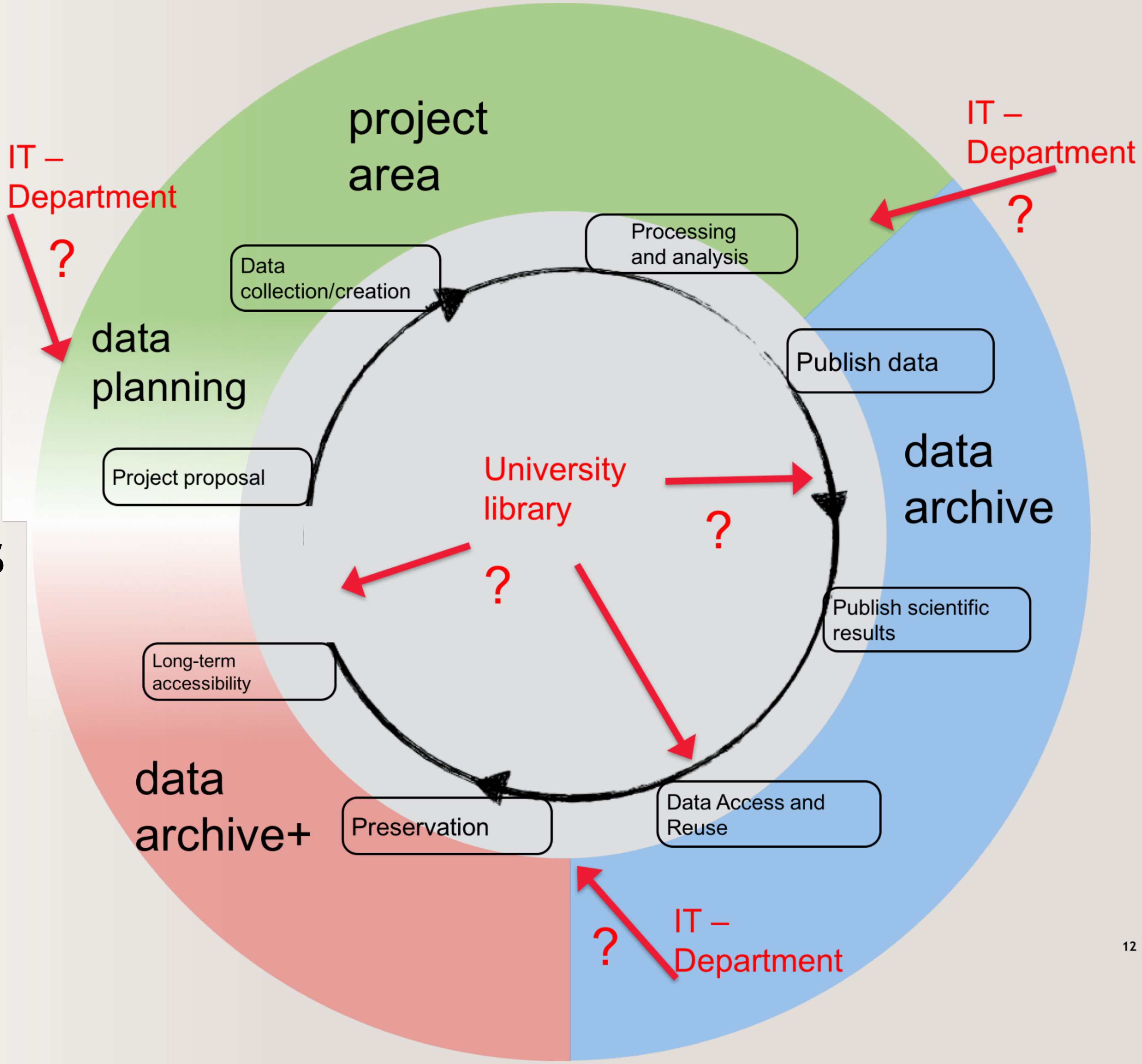
Long-term accessibility

Project proposal

Different actors: Who does what?

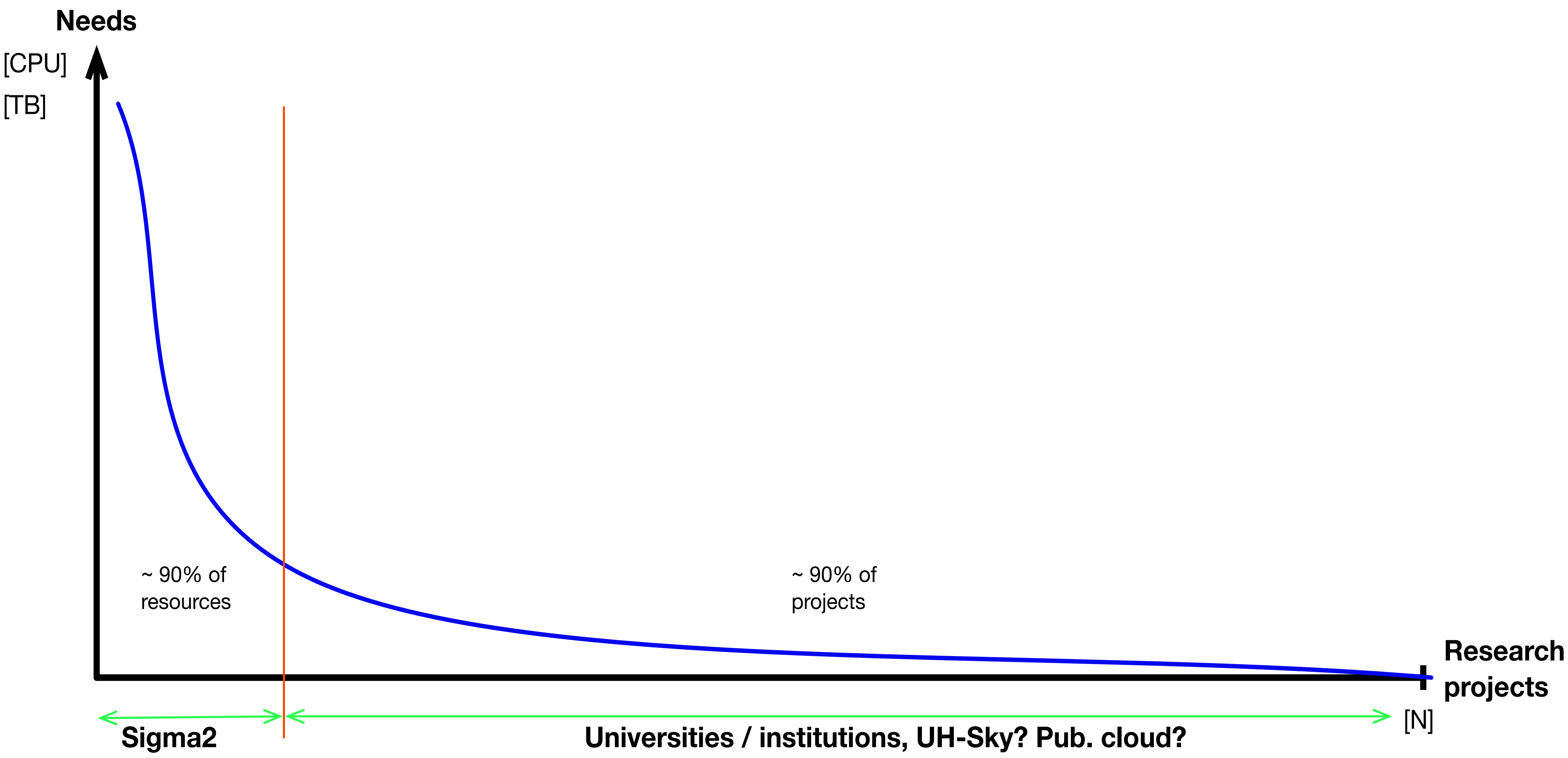
- International organizations
- Governmental organizations
- National organizations
- Universities/Institutions
- Departments/Research Groups

- And commercial actors?



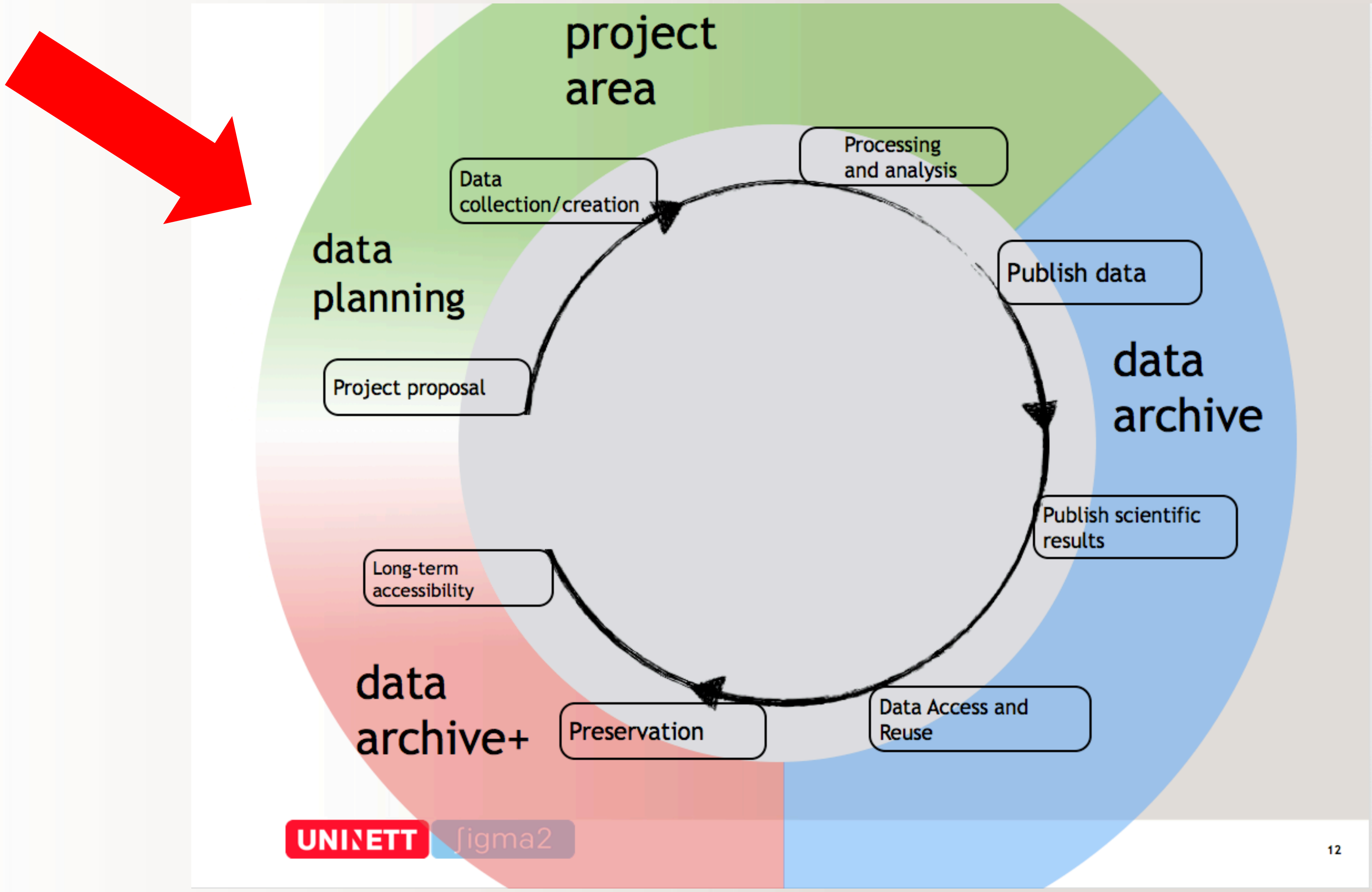
UNINETT ʃigma2

Local vs national e-infrastructures



Sigma2 e-infrastructure services

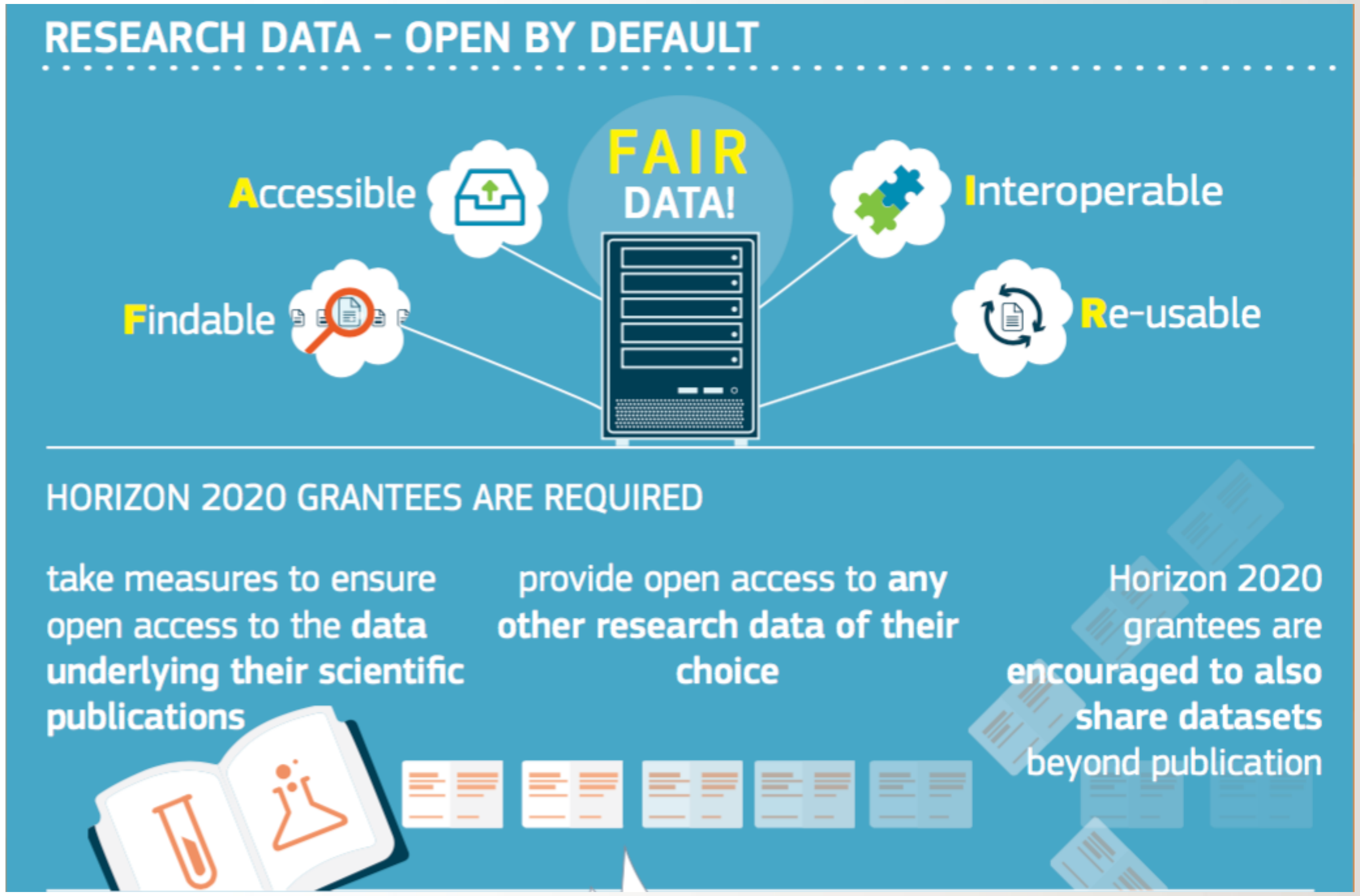
Data Management Stewardship



The Future of FAIR Data Stewardship

European Commission

OPEN RESEARCH DATA
IN HORIZON 2020



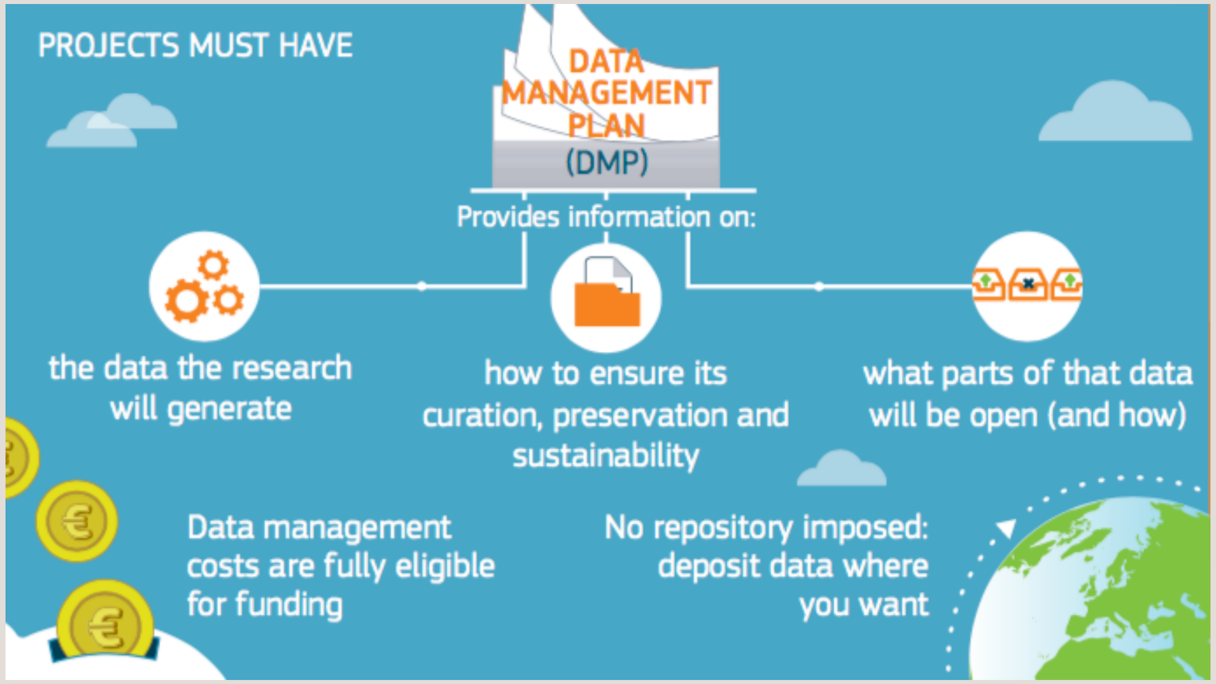
RESEARCH DATA – OPEN BY DEFAULT

HORIZON 2020 GRANTEES ARE REQUIRED

take measures to ensure open access to the **data underlying their scientific publications**

provide open access to **any other research data of their choice**

Horizon 2020 grantees are **encouraged to also share datasets** beyond publication



PROJECTS MUST HAVE

DATA MANAGEMENT PLAN (DMP)

Provides information on:

- the data the research will generate
- how to ensure its curation, preservation and sustainability
- what parts of that data will be open (and how)

Data management costs are fully eligible for funding

No repository imposed: deposit data where you want

5%

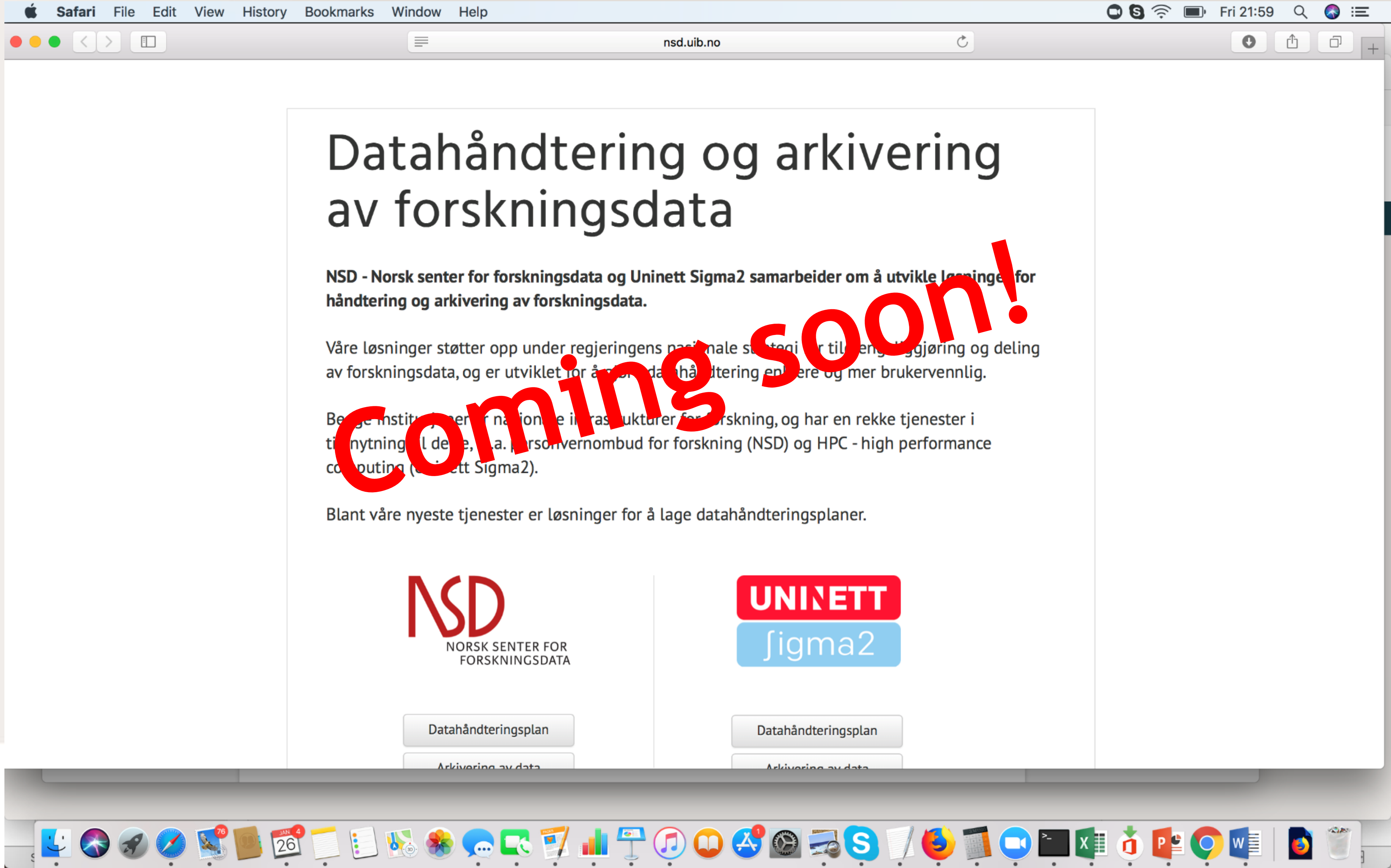
FAIR

UNINETT Sigma2

Courtesy of Barend Mons, GoFAIR

DMP tools in Norway

- Tools to facilitate the creation of the DMP
- Two DPM tools in Norway, one provided by NSD and one provided by Sigma2
- A common webpage as entry point to guide the researchers in the process of choosing the best tools for their needs:



Datahåndtering og arkivering av forskningsdata

NSD - Norsk senter for forskningsdata og Uninett Sigma2 samarbeider om å utvikle løsninger for håndtering og arkivering av forskningsdata.

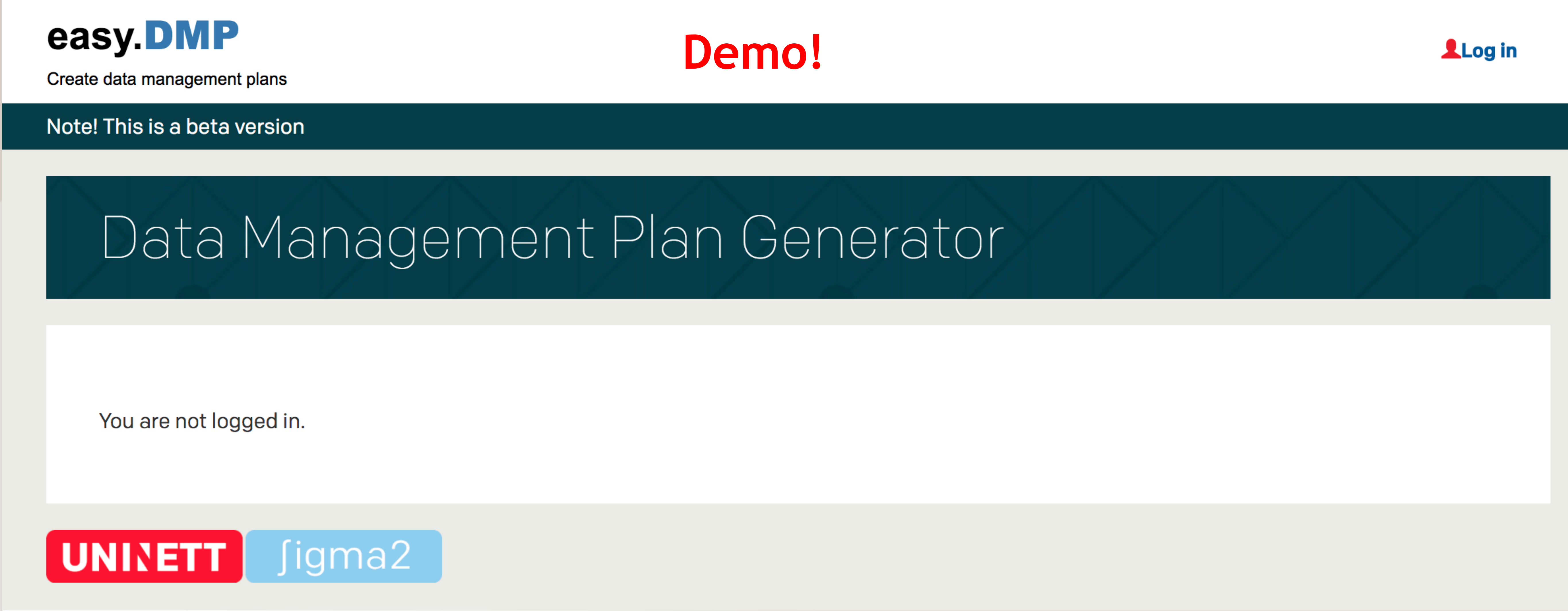
Våre løsninger støtter opp under regjeringens nasjonale strategi for tilgjengeliggjøring og deling av forskningsdata, og er utviklet for å gjøre datahåndtering enklere og mer brukervennlig.

Begge institusjonene er nasjonale infrastrukturer for forskning, og har en rekke tjenester i tilknytning til dette, bl.a. personvernombud for forskning (NSD) og HPC - high performance computing (Uninett Sigma2).

Blant våre nyeste tjenester er løsninger for å lage datahåndteringsplaner.

Coming soon!

Demo!



easy.DMP

Create data management plans

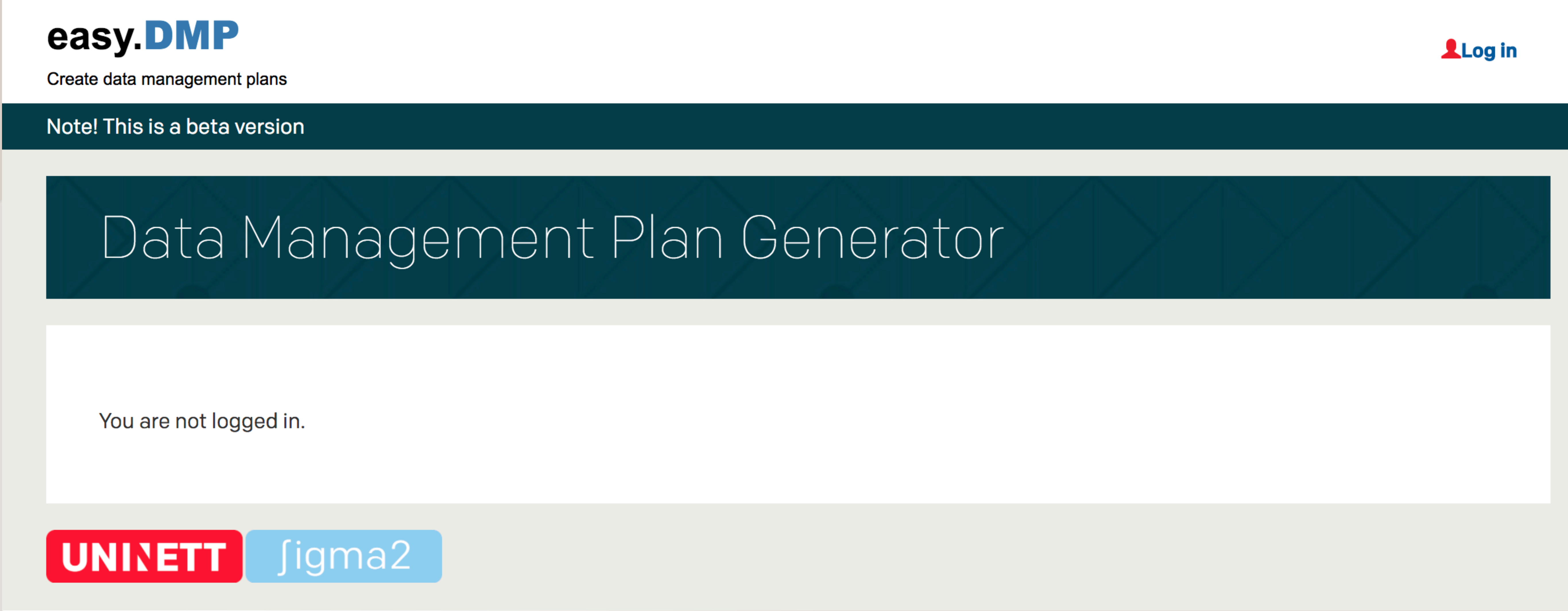
Log in

Note! This is a beta version

Data Management Plan Generator

You are not logged in.

https://easydmp.paas2.uninett.no/ (beta version!!)



easy.DMP
Create data management plans

Log in

Note! This is a beta version

Data Management Plan Generator

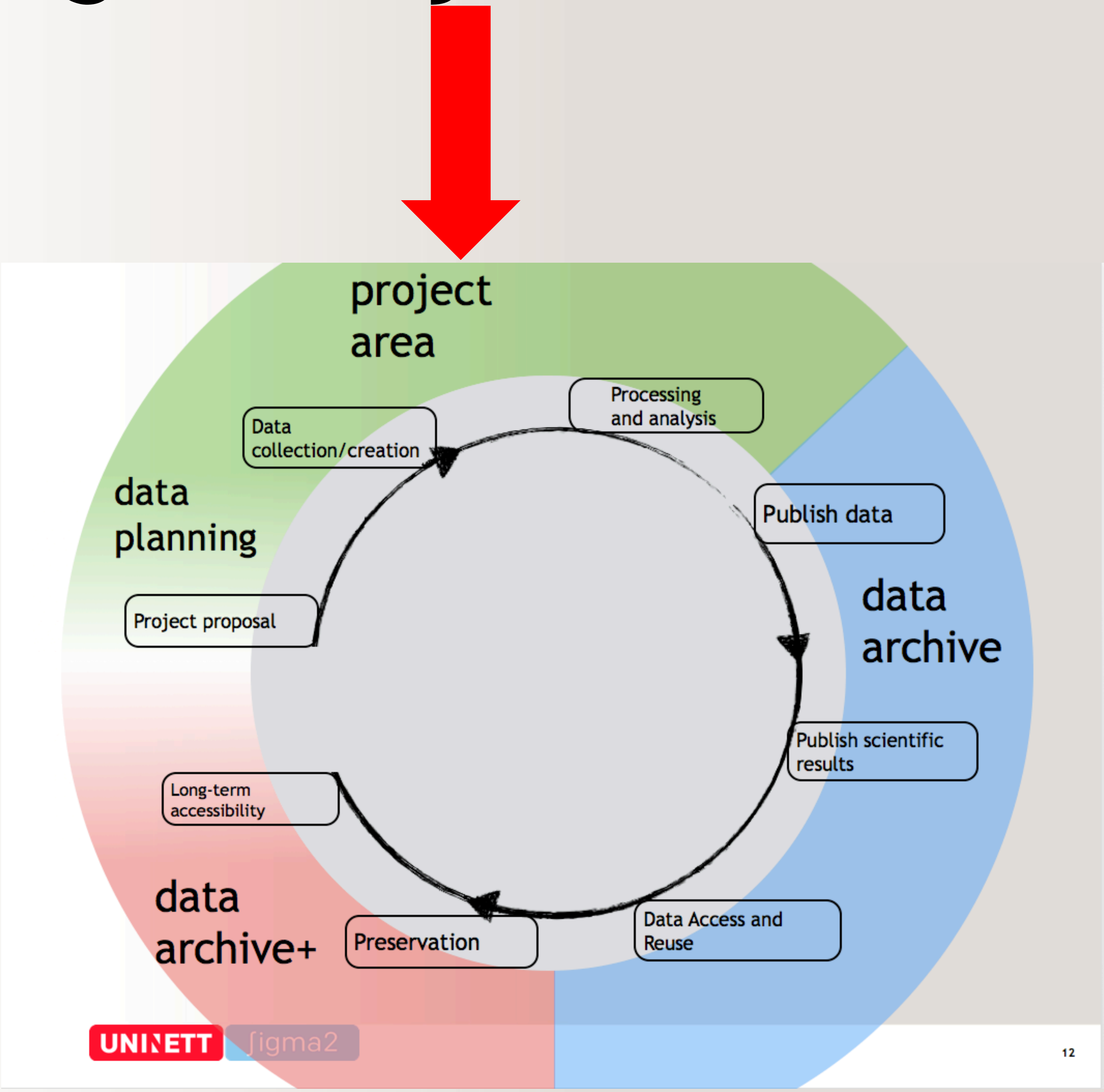
You are not logged in.

UNINETT Sigma2

- Support metadata repositories (in collaboration with **OpenAIRE ***)
- Developed in partnership with EUDAT2020
- Support H2020 schema, and any other schemas (universities, research communities specific…)
- Machine Readable output

) **OpenAIRE is a network of Open Access repositories, archives and journals that support Open Access policies.*

NIRD Storage – Project Area



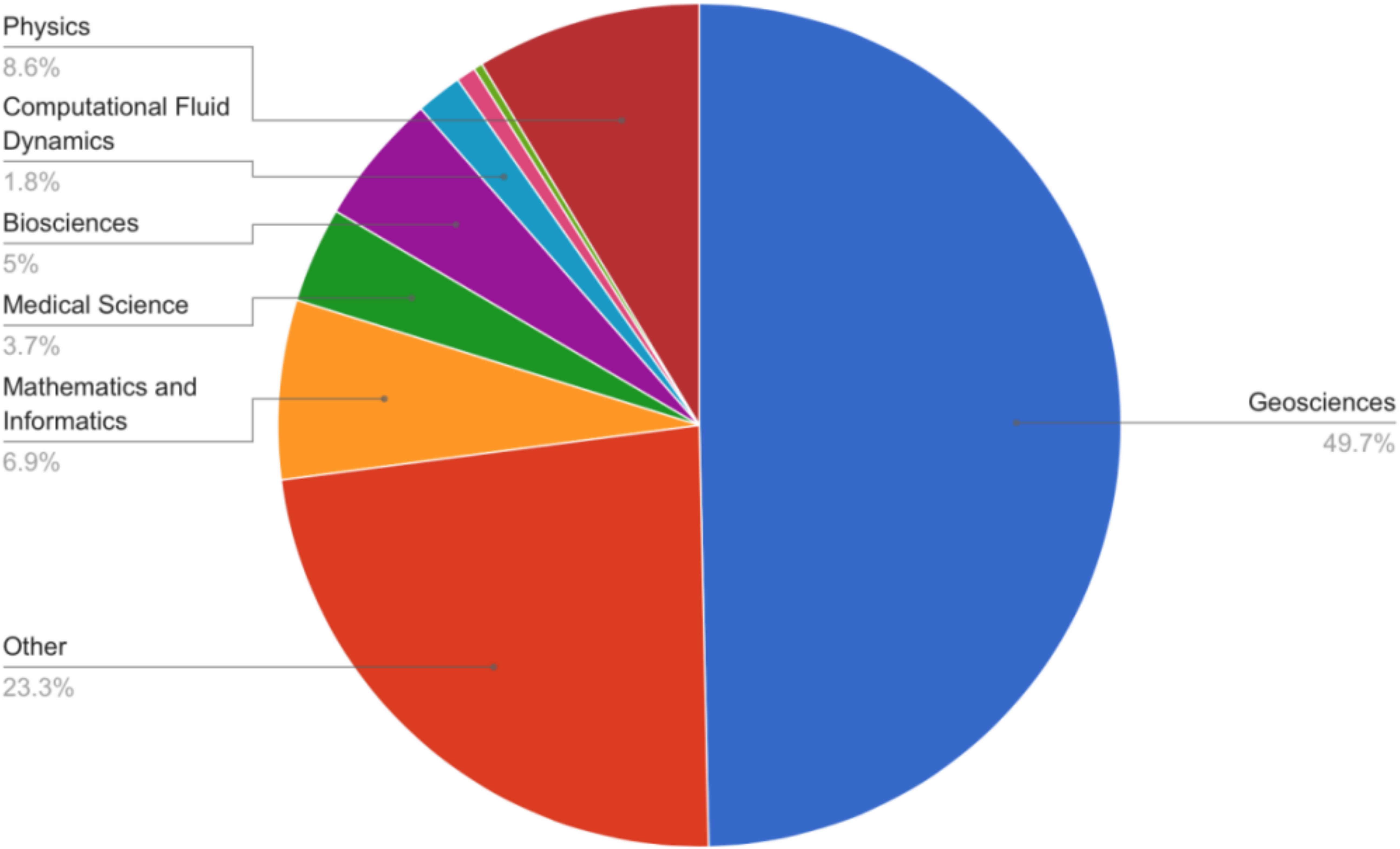
NIRD Storage infrastructure

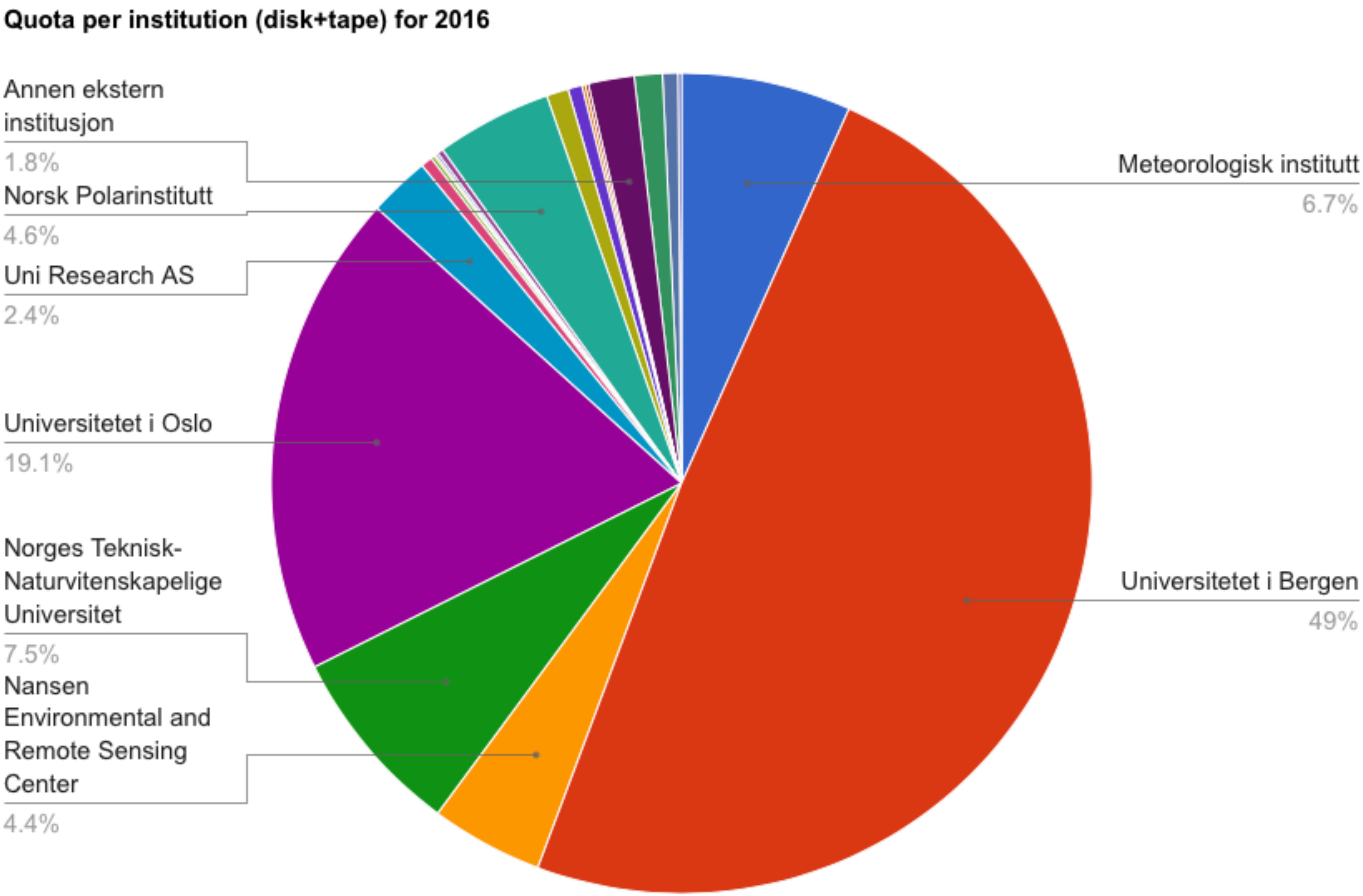
- Project storage (minimum 10 TB)

- Norstore is replaced by NIRD – National Infrastructure for Research Data

System	Capacity [PB]	Deployed	Location
Norstore	3.7	1/2013	Oslo (+Tromsø)
NIRD	5.6	9/2017	Tromsø + Trondheim
(NIRD exp.)	~10?	(2/2018)	

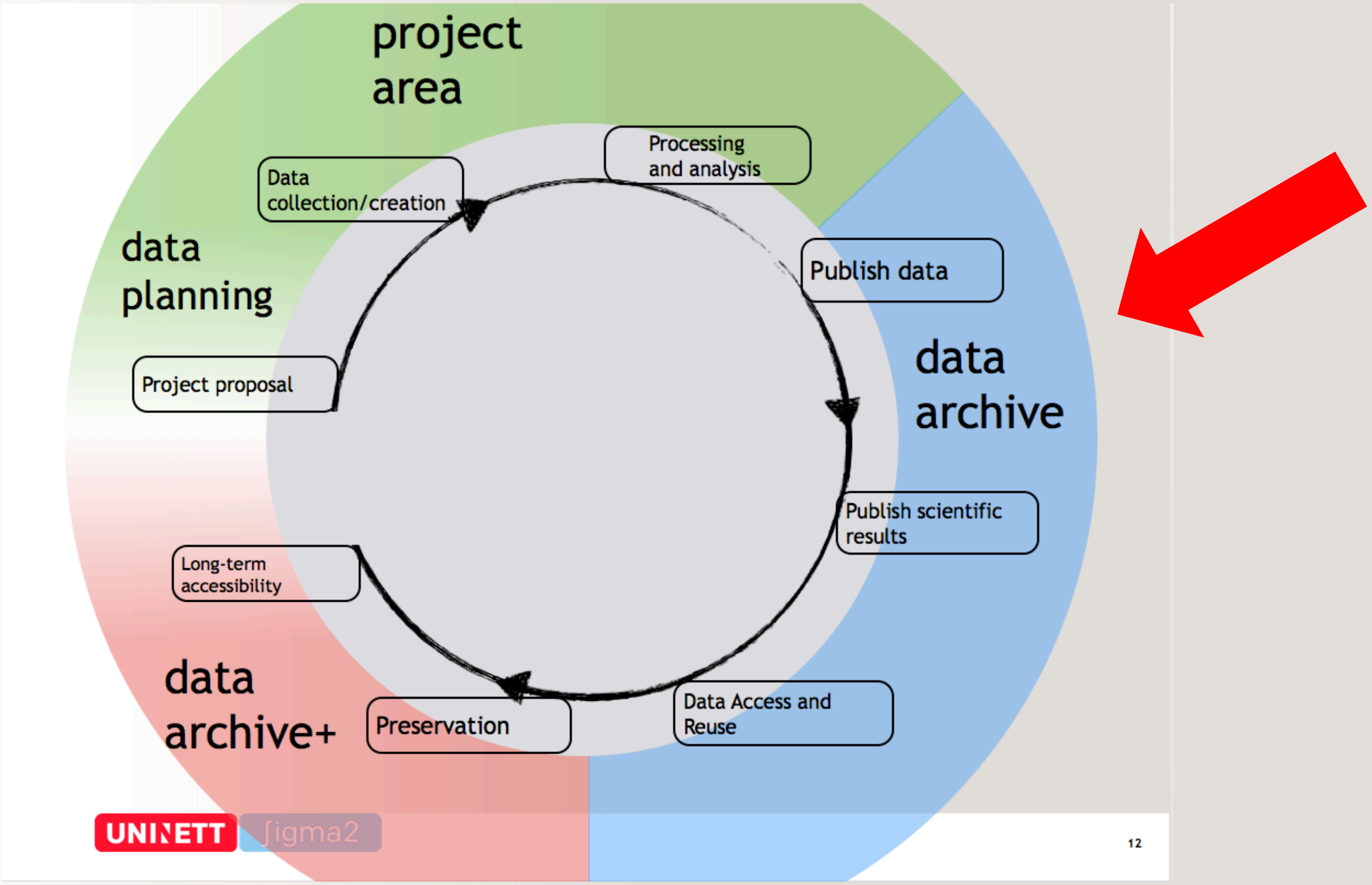
Quota per discipline (disk+tape) for 2016



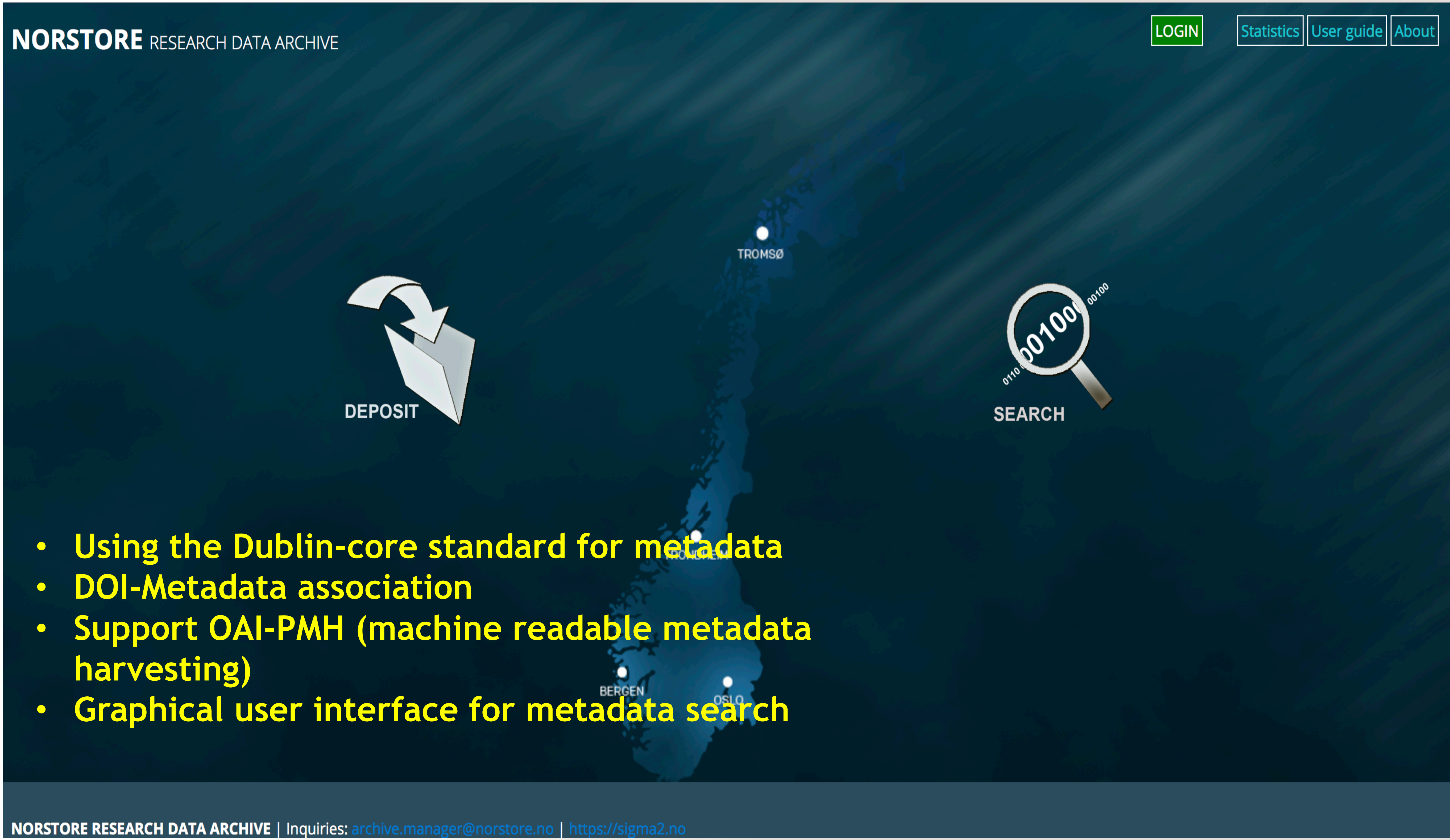


UNINETT sigma2

Archive, publish data and data reuse



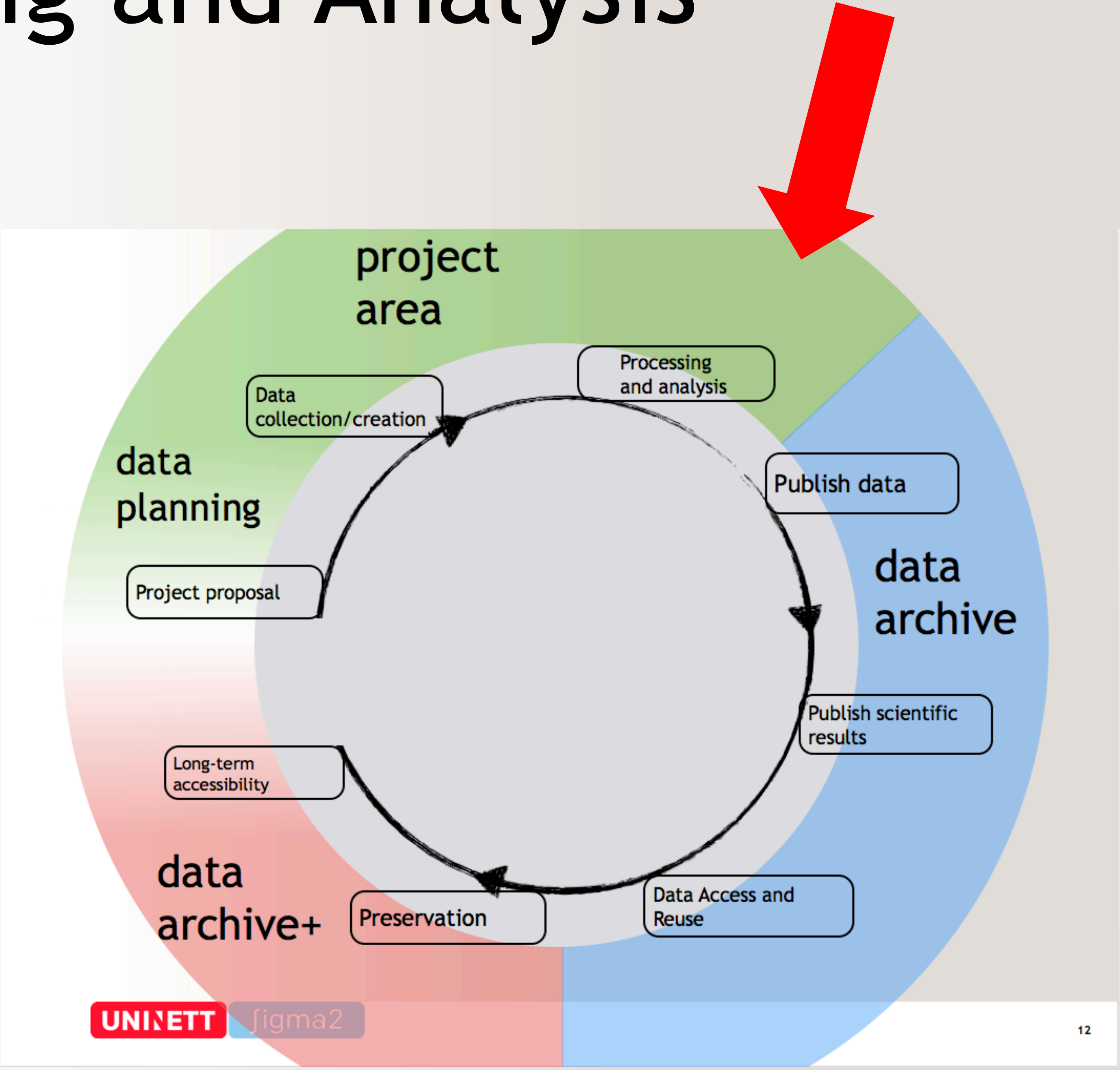
NIRD Archive



- Using the Dublin-core standard for metadata
- DOI-Metadata association
- Support OAI-PMH (machine readable metadata harvesting)
- Graphical user interface for metadata search

NORSTORE RESEARCH DATA ARCHIVE | Inquiries: archive.manager@norstore.no | https://sigma2.no

Processing and Analysis



project area

Processing and analysis

Data collection/creation

data planning

Project proposal

Publish data

data archive

Publish scientific results

Long-term accessibility

data archive+

Preservation

Data Access and Reuse

12

The NIRD Service Platform

- Bring compute to the data, not the other way around (data-centric architecture, sits “on top of” NIRD)
- Powerful compute nodes and virtualization technology (Kubernetes, Docker containers) for on-demand tasks and fast service deployment
- Designed for close integration with commercial cloud services.

Strength of the Service Platform (SP)

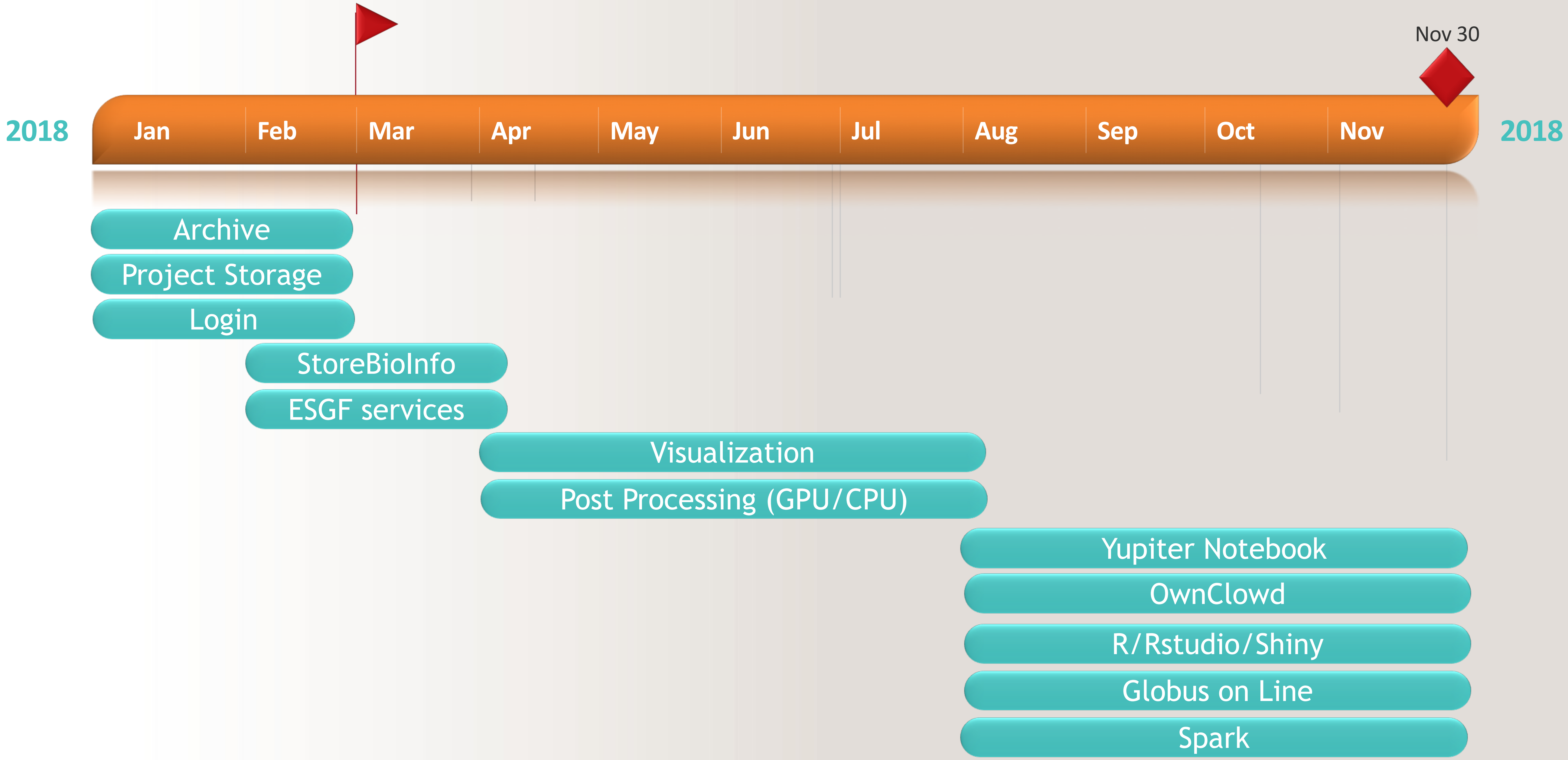
- **Flexible and versatile:** SP can host any dockerized service

- **Cost-effective:** SP computing resources can be use to dockerized jobs or tradictional HPC jobs (single threaded or OpenMP jobs)

- **Customizable:** researchers can run their own service (web service, computing workflows etc…) provided that it is dockerized

- **GPUs for visualization** and **GPU/CPU computing** (data analytics, machine learning, artificial intelligence)

Services Deployment Roadmap



Nov 30

2018 Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov 2018

Archive

Project Storage

Login

StoreBioInfo

ESGF services

Visualization

Post Processing (GPU/CPU)

Yupiter Notebook

OwnClowd

R/Rstudio/Shiny

Globus on Line

Spark

UNINETT ʃigma2

Services for sensitive research data

- Data that can be related to human subjects is by law/nature sensitive[*], and the importance and prevalence of this type of data in research is rapidly increasing as it relates to health and other societal issues of high impact and visibility.

- Our ability to do research involving sensitive data is dependent on e-infrastructure that can protect the data according to laws and regulations while at the same time providing access and resources according to the needs of the researchers.

- UiO/USIT, together with Sigma2 and others, have collaborated on establishing a secure e-infrastructure to provide services for sensitive data. The resulting ”TSD” is a **national** platform for all types of research involving sensitive data.

(*) PERSONAL DATA REVEALING INFORMATION REGARDING RACIAL OR ETHNIC ORIGIN, POLITICAL OPINIONS, RELIGIOUS OR PHILOSOPHICAL BELIEFS, TRADE-UNION MEMBERSHIP, DATA CONCERNING HEALTH, SEX LIFE.

High Performance Computing (HPC)

- Transiting from one HPC system at each of the four universities, to a shared model with two systems, with 2-year leap-frogged installation across a 4-year lifetime for each (two tracks).
- From 1 October ‘17 compute load serviced by Abel, Stallo and **Fram**. From early ‘19 Fram + the next system, “B1”.
- **Shared and distributed operations between the four universities coordinated by Sigma2.**
- Access to compute time on Colossus (TSD) for sensitive data available also from Sigma2.
- Accelerators, GPUs and Xeon Phis, currently available on Abel, soon also on the NIRD Service Platform (nVidia P80 or P100).
- The HPC resources, TSD and the NIRD Service Platform to complement each other in a data-centric “echosystem”.

High Performance Computing (HPC) resources

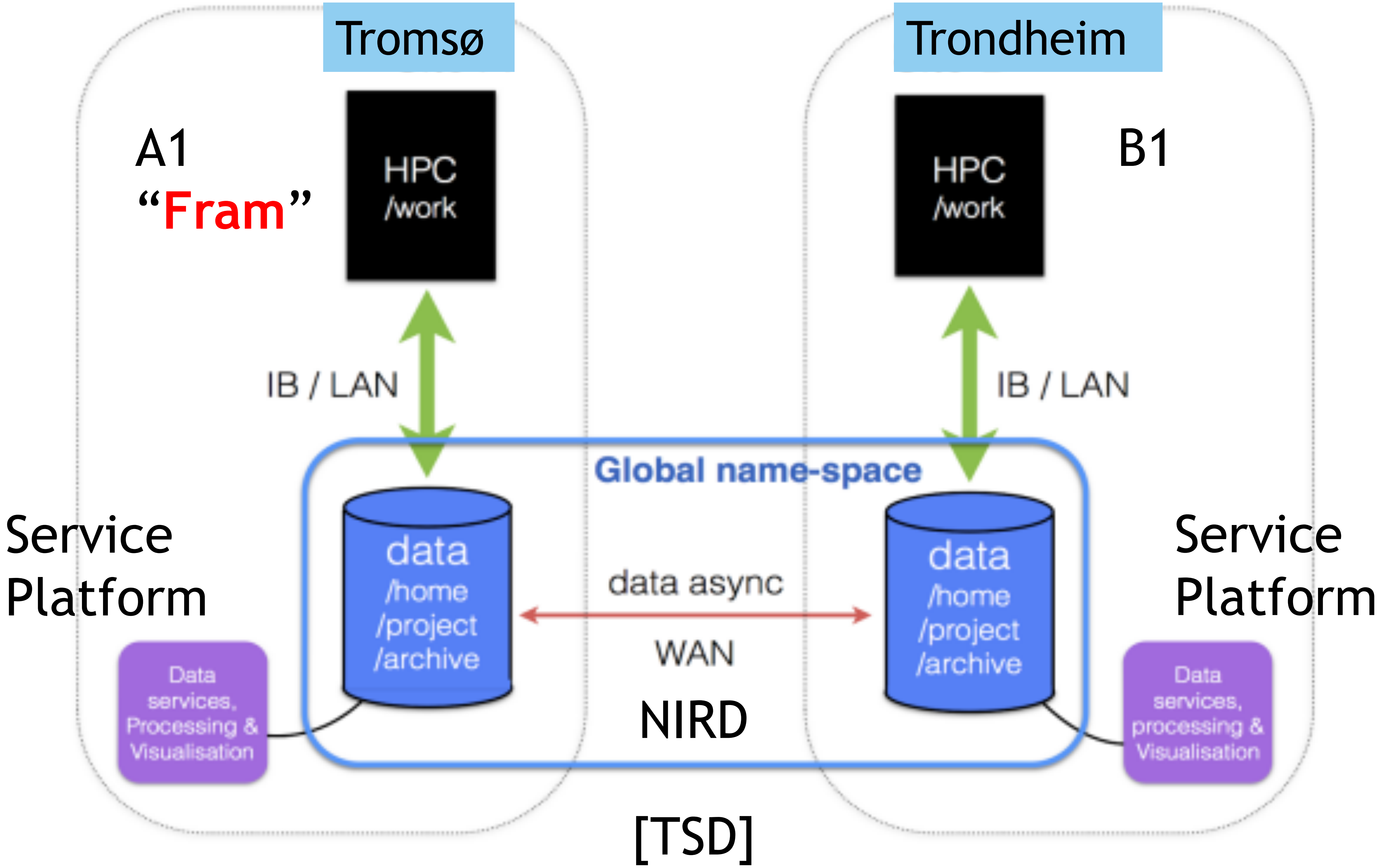
System	Sigma2 capacity (MCPUhrs/yr)	Tot. performance (TFLOP/s)	Deployed
Hexagon	102.8	109	4/2012
Abel	75.9	182	10/2012
Vilje	113.0	312	10/2012
Stallo	120.4	~291	10/2012 (+ utv.)
Colossus*	<13	~30	4/2014
Sum	322.1	894	
Fram	279.2	1071	10/2017
”B1”	?	?	(4Q/2018)
“HTC** platform”	?	?	(2H2018)

(*) For sensitive data, part of TSD

(**) HTC = High Throughput Computing / cloud platform

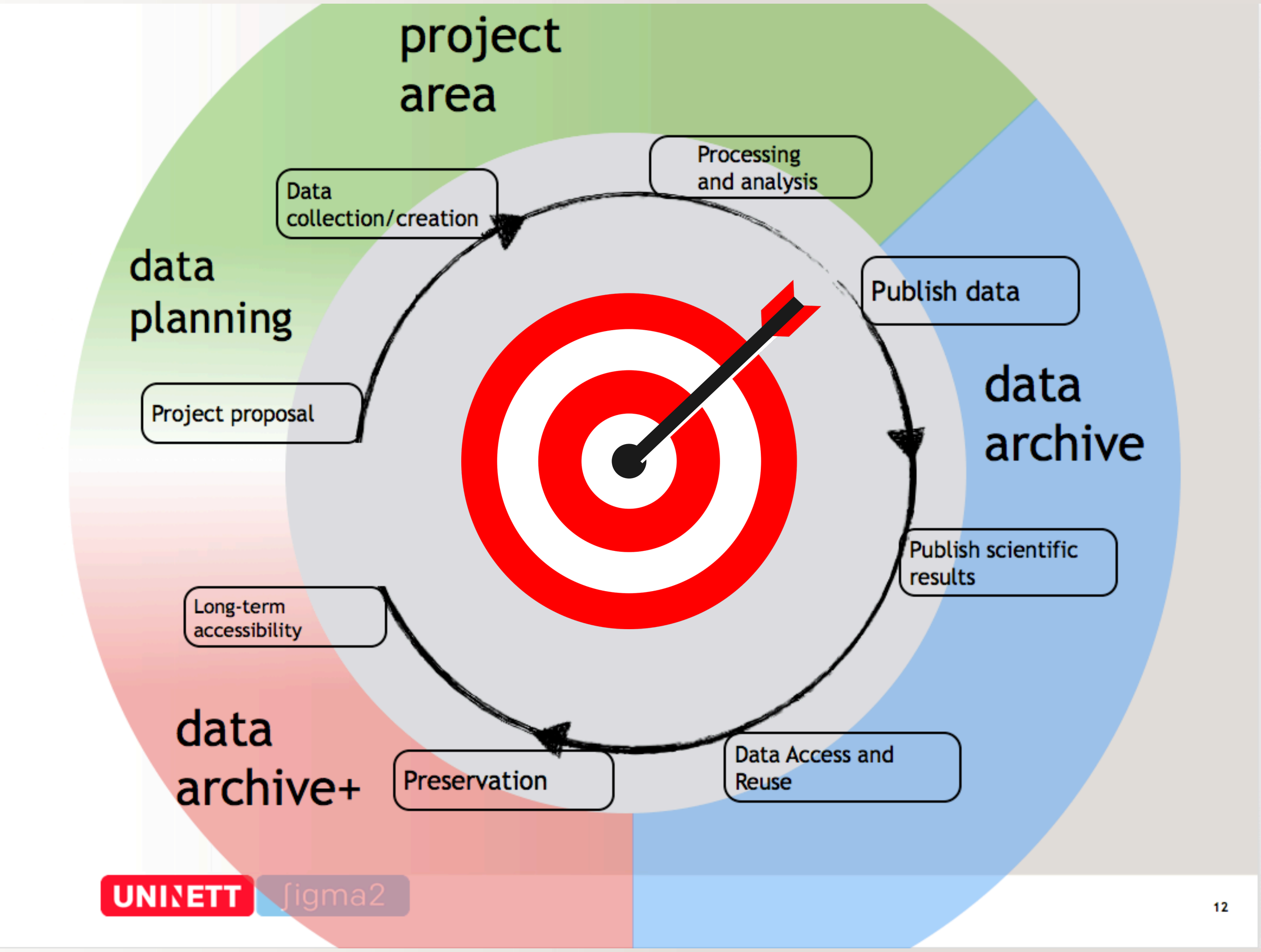
UNINETT ʃigma2

Implementing the data-centric architecture



UNINETT ʃigma2

Advanced User Support (AUS)



Advanced User Support (AUS)

- 1) Project based AUS:
  - Can be the sole initiative of a researcher or a science area
  - Granted by RFK with 2-3 PMs spent over a maximum of 6 months, continuous applications
- 2) Discipline specific AUS
  - Initiated by Sigma2 in cooperation with a science discipline
  - Can have allocations of more than 12 PMs spent over a maximum for 2 years
  - Joint funding

Advanced User Support (AUS)

For the HPC services, project based advanced user support aims at helping scientists to improve or extend the performance and capabilities of their applications. This can be in a number of ways, including:

- code parallelization
- code porting
- code profiling, optimization, benchmarking
- improving user-interfaces
- software development

For the storage services, project based advanced user support aims at:

- assist researchers to create data plans
- implementing best practices for collecting and handling data
- identifying or defining meta-data schema
- identifying suitable storage formats
- identifying dedicated or specialised tools to help access or visualize data, utilise the facilities better

Advanced User Support (AUS)

- How to apply for AUS:
  - At any time, contact sigma2@uninett.no or start from https://www.sigma2.no/content/advanced-user-support-0
  - Small AUS projects might be granted within a week, larger projects (e.g. discipline specific AUS) might need longer time

Getting access to the national e-infrastructure

Getting access to the national e-infrastructure

By application

- Calls twice a year (Jan/Feb, Aug/Sep):
  - **https://www.metacenter.no/mas/application/project/**

Right away

- Small and exploratory needs (e.g. on Fram)
  - https://www.metacenter.no/mas/application/project/
  - If in doubt: sigma2@uninett.no

- See https://www.sigma2.no/content/apply-e-infrastructure-resources

Resource allocation

- Resources made available to all research carried out under the auspices of Norwegian research institutions
- Decided by the Resource Allocation Committee (RFK)
- Applications are assessed on the basis of the project's scientific quality
- Two calls every year for major applications (continuous calls for minor applications and advanced user support)

Help!

Technical support

- User documentation:
  - **https://www.sigma2.no/content/support-e-infrastructure-users**

- All support requests: **support@metacenter.no**
  - Applications for compute and storage resources go to sigma2@uninett.no

www.sigma2.no