DiSSCo

Distributed System of Scientific Collections

CETAF ISTC meeting Copenhagen Feb 2018

Wouter Addink
DiSSCo Coordination team member

Natural Science collections - bio- & geo-diversity

Leading scientific facility

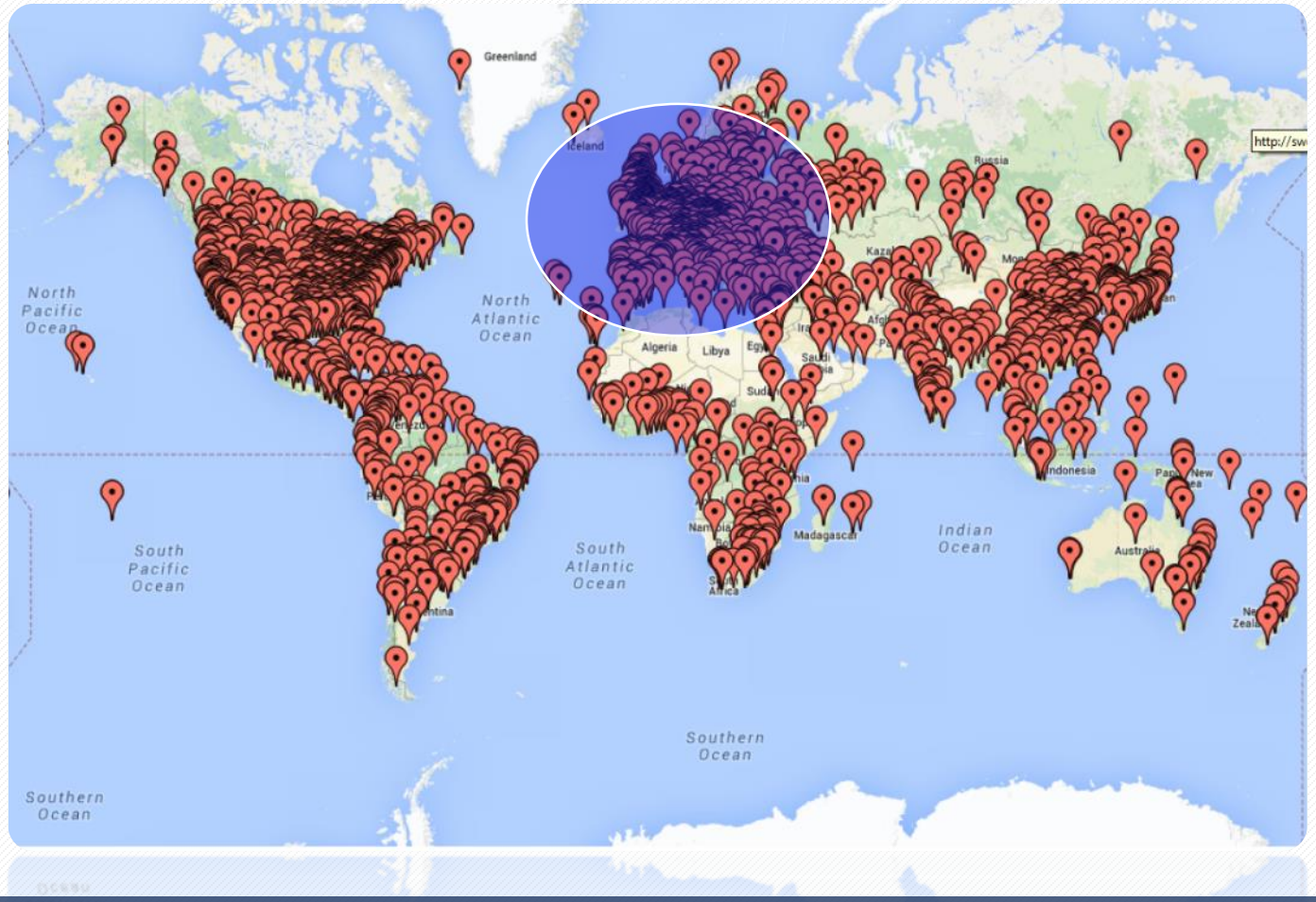
Discovering, describing and interpreting life on Earth

Tackling societal challenges

Sustainable future

Europe: the global leader

55% of the world's assets with rich historical and global distribution



European Collection facilities:

> **1.5 billion** specimens

> **80%** of world's species

> **5,000** scientists employed

> **16,000** scientific visitors pa

> **10 million** public visitors pa

> **25 million** web visitors pa

- 16,000 researchers travel every year to physically access scientific collections
- 800k objects are packed and shipped (at an annual public cost of more than €70M)

DiSSCo: A new European infrastructure

114 National Facilities

21 Countries

- Largest ever formal agreement between natural science collection facilities
- Centralised governance model already in place
- Supporting network of working groups

- DiSSCo builds on top of a mature community of institutions
- Strategic collaboration already underpinned by sound governance and decision-making structures

Added value of DiSSCo

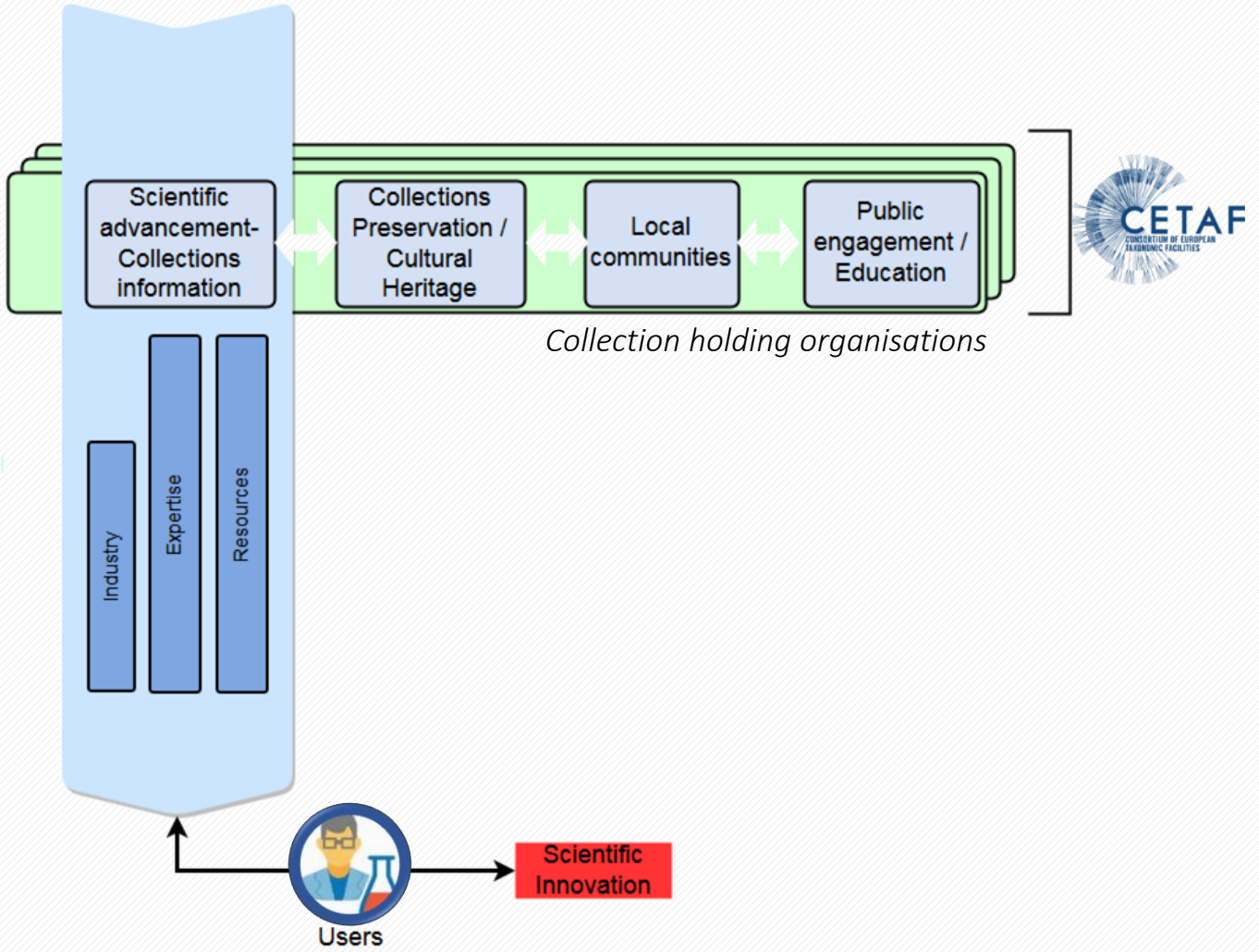
Without DiSSCo	With DiSSCo
Disconnected information sources	**Linked and open information** with semantic annotation
Slow and fragmented access	**Coordinated physical and virtual access** through a single entry point
Bio- and geo-diversity data invisible to other Ris	**Cross-disciplinarity** facilitated through RI systems interoperability
Provenance and quality difficult to ensure	**Provenance and quality assurance** embedded in services/processes
Big data science questions unresolvable	Bio- and geo-diversity **data brought to the big data pool**
Institutional based digitisation activities	**Coordinated digitisation** programmes: One EU collection
Disconnected efforts	**Coordinated investments** - Economies of Scale

- DiSSCo provides clear added value to existing network activities
- DiSSCo required to deliver coordinated services, access and development

The DiSSCo approach

Distributed System of Scientific Collections

- **Transfer of authority** from facilities to central for all key operations
- Clear **decision making** mandate
- Focused **scientific scope**
- **Binding institutional commitments** (already in place)
- New **independent legal framework**



- **CETAF**: A 20-year network acting as a forum for discussion and advocacy
- **DiSSCo**: An integrated research infrastructure providing data, services and links to other infrastructures

DiSSCo science services

1 **e-Science services**

A one-stop shop for services providing unified discovery, access, interpretation and analysis of complex linked data

2 **Physical and remote access services**

A universal harmonised physical access service and digitisation on demand service

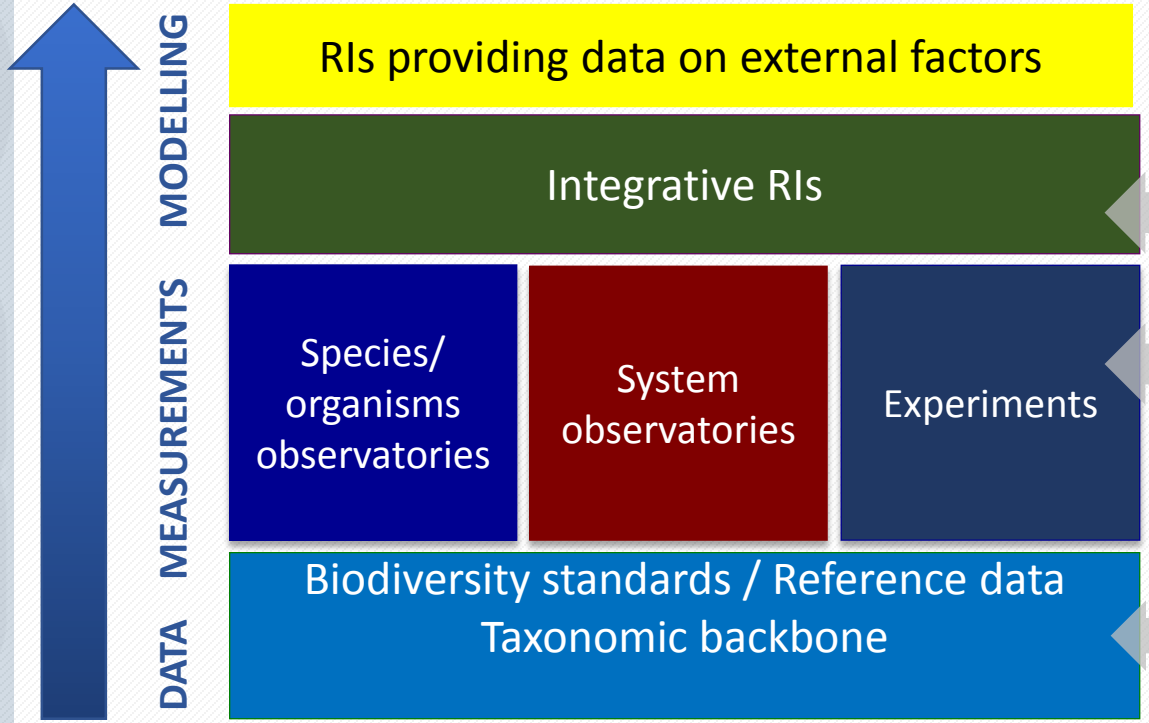
3 **Support & Training services**

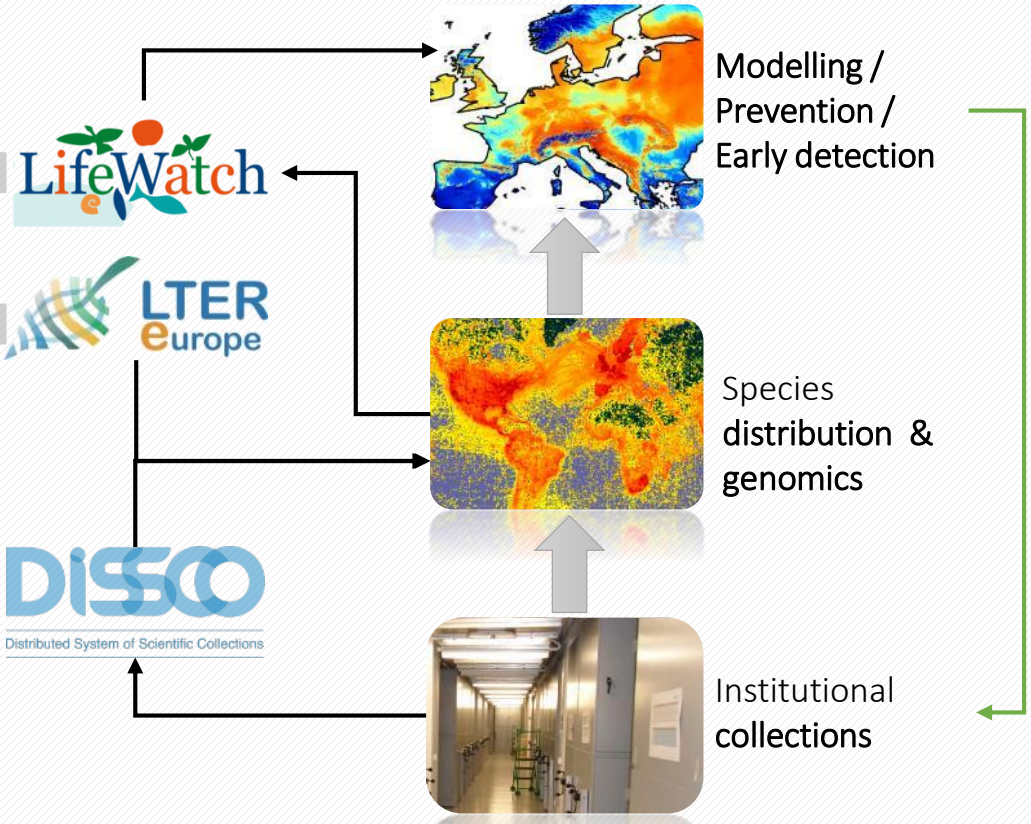
Integrated user support desk and implementation of multi-modal training programmes to enhance data skills

DiSSCo provides three distinct classes of service to the scientific community

DiSSCo services to other infrastructures

ENVRI plus, 2017





Alien Invasive species use case

Need and Urgency of DiSSCo identified by:

Env Science WG (2010)
OECD SciColl (2013)
ESFRI roadmap (2016)

- DiSSCo delivers services on bio- and geo- diversity reference data to other RIs (especially in the Environmental domain).
- DiSSCo occupies a foundational layer in the Research Infratsructure landscape

Catalogue of Life in DiSSCo

Required for DiSSCo to fill the identified Taxonomic Gap in the RI landscape (ESFRI landscape analysis 2016)

- Integration as Taxonomic Backbone in DiSSCo Community Services
- DiSSCo Thematic Services to (re-)connect taxonomic research with specimen data
- Quality control and enhanced linkages
- Contribution of taxonomic expertise through the Col Plus Clearinghouse

Joint development of a practical, community-based approach to rapid completion of the GBIF Taxonomic backbone and other existing digital catalogues of published scientific names

DiSSCo Design: Accelerating scientific research

Design Study

Proven research requirements

- Information highly dispersed
- Require integrated data and services
- Coordinated strategy

Need

Response

DiSSCo
Distributed System of Scientific Collections

Accelerate, scale up, coordinate strategy

- Accelerate generation and linking of information into research data objects
- Ensure provenance and quality
- Provide reliable, unified, certified services and harmonised policies
- Provide services to other Research Infrastructures

Outputs

Open science Broad user base

- Enable big data & open science
- Expanded scientific constituency
- Focus on societal challenges
- Enable multi-disciplinary science

Impact

UN Sustainable Development Goals

- DiSSCo is a response to the need for high quality, aggregated data and services to tackle societal challenges
- DiSSCo will broaden user base, by lowering the entry barrier to big data science

Case study – Invasive Alien Species

UN Sustainable Development Goals (Target 15.8)

EXAMPLE: Alligator Weed
(Alternanthera philoxeroides)
Negative impact on native species, ecosystem services and infrastructure

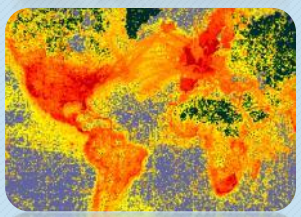
Economic costs of IAS for EU
€20 Billion / year
Kettunen et al. 2009

Urgent challenge

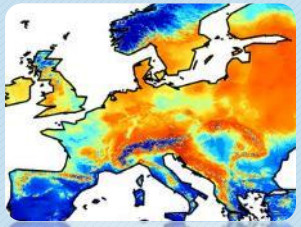
DiSSCo
Distributed System of Scientific Collections

Institutional **collections**

Species **distribution & genomics**



Modelling / Prevention / Early detection



Facilities & information

Linked Data

Analysis / Interpretation Services

- *Climate data*
- *Ecological monitoring data*
- *Genomic information*

Other Research Infrastructures

- Linking dispersed information is imperative for providing solutions at the scale, urgency and quality required
- DiSSCo provides the links required to tackle key scientific and societal challenges

Global data standards

DiSSCo works with global bodies to address key technical challenges

Produces (meta-)data Standards for biodiversity information for the last 35 years

DiSSCo is working within TDWG (Interest and Task Groups) to drive further development of Standards (incl. controlled vocabularies). Will reuse existing Standards when possible (e.g. DarwinCore) [Formal agreement for future collaboration in place: MoU]

DiSSCo leads on the work of several Interest and Working Groups including the Disciplinary Collaboration Framework IG towards a transdisciplinary interoperability framework.

DiSSCo works also in the context of other IT and domain standardisation bodies to ensure better alignment

- DiSSCo works with key stakeholders (incl. TDWG & RDA) to develop data standards and guidelines.
- TDWG and DiSSCo have already signed collaboration agreements

One European collection

A new business model

DiSSCo
Distributed System of Scientific Collections

Synchronising 114 facilities:

- Act under the premise:
  One European Collection of scientific assets
- Understanding scientific needs for developments of collections at European level rather than institutional
- Creating Economies of scope and scale (pooling resources, expertise & experience)
- Deploy monitoring mechanisms to measure impact of collections at institutional and national level
- Enabling national institutions to invest in specialisation strategies (e.g. in alignment with national priorities)

Institutional collections will not only benefit from maximising the scientific impact of their assets, but also they will be able to inform **strategic collections development decisions** based on global needs. DiSSCo mandates a new business model in the operation of its distributed facilities

Decision making: DiSSCo Stakeholder Forum

Data, workflow and systems integrity

FAIR principles

EOSC

- EUDAT/EGI participate in Stakeholder forum and Technical Advisory Board
- Pilots: Herbadrop with EUDAT, and AAI implementation with EGI

Future-proofing data & digitisation

Co-funded by the Horizon 2020
Framework Programme of the European Union

*DiSSCo EC-funded project (**ICEDIG.eu**)*
(scored: 14.5/15)

- Quality assurance and fit for purpose by applying common standards and procedures
- Stable to future technologies through robust metadata schemes
- Advanced digitisation techniques will be driven by research requirements

Established technology
2D image
1 picture with sub-micron resolution
= 20 MB / specimen

Emerging Technology
3D model
raw data and reconstructed images
= 200 GB / specimen

- Value of current data will persist – new information will be added
- DiSSCo will drive technological innovation to support future research
- Societal challenges require urgent mobilisation of information

Re-use and decommissioning

User-oriented service layer

Persistent data layer

E-infrastructure layer

- 5- and 10- year evaluation based on scientific impact
- **Decommissioning possible**

- **Data availability ensured even beyond DiSSCo operation**
- Deposition of data and scientific workflows in certified public data repositories (national/institutional)

- **Service availability** and quality through **Service Level Agreements**

- Separation of service and data layers allows decommission/rebuild of user facing services.
- Data will be stored in certified repositories at national/institutional level.

User Strategy and access plans

Services

Virtual access to data and services

Physical and remote access

Training and support

User strategy & Access plans

- Open access (by default) to data
- Controlled access to added value services
- Single entry point

- Physical access and digitisation on demand excellence/innovation driven (drawing on 13y of SYNTHESYS experience)
- Priority to requests that cannot be served through virtual access

- Central user support desk
- Coordination of training

Q9

- During the preparatory phase we will develop the exact user strategy in alignment with the DiSSCo business model
- H2020 funding for this is already secured through the ICEDIG project

Financial Resourcing: Sources of support

European Level	Additional financing through European Investment Bank and other financial instruments
National Level	Industrial scale digitisation funded through national digitisation programmes and supported by national data infrastructure
Institutional Level	Curation, collection management and expertise for data enrichment, on-demand digitisation

- DiSSCo requires resources at European, national and institutional levels
- National and institutional commitments are already in place for the preparatory phase (14 countries)

Financial Resourcing: Aligned Projects

SAP

Strategic Alignment of Projects

ICEDIG

Funded by the Horizon 2020 Framework Programme of the European Union

€3M | 2018-2020

CoL+

GBIF

€0.5M | 2017 - 2020

DiSSCo Design Study

€10M | 2014 - 2017

SYNTHESYS+

€10M | 2019 - 2021

DiSSCo Deploy

€2M | 2024 - 2025

MOBILISE

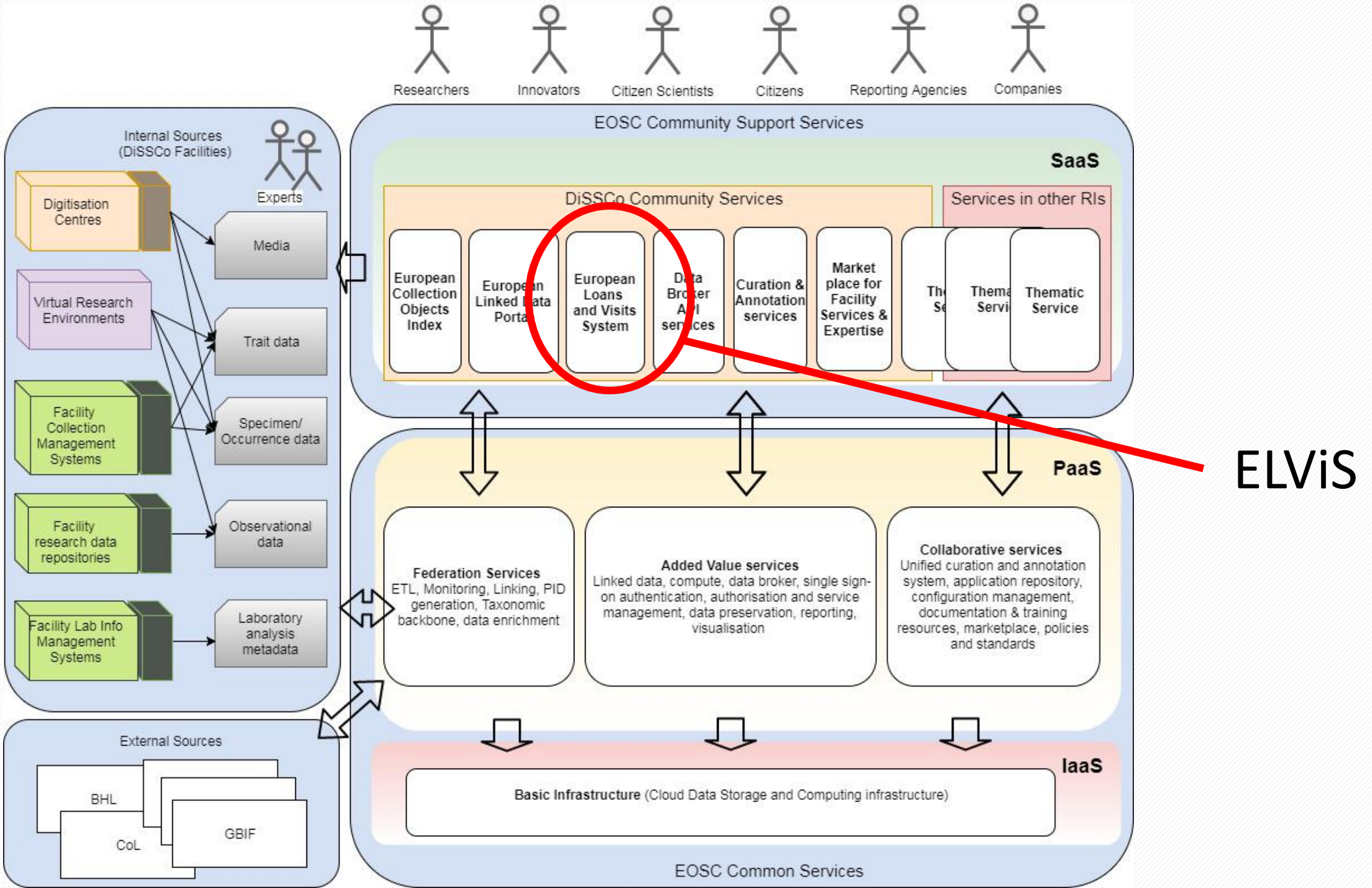
€0.5M | 2018 - 2022

DiSSCo Construct

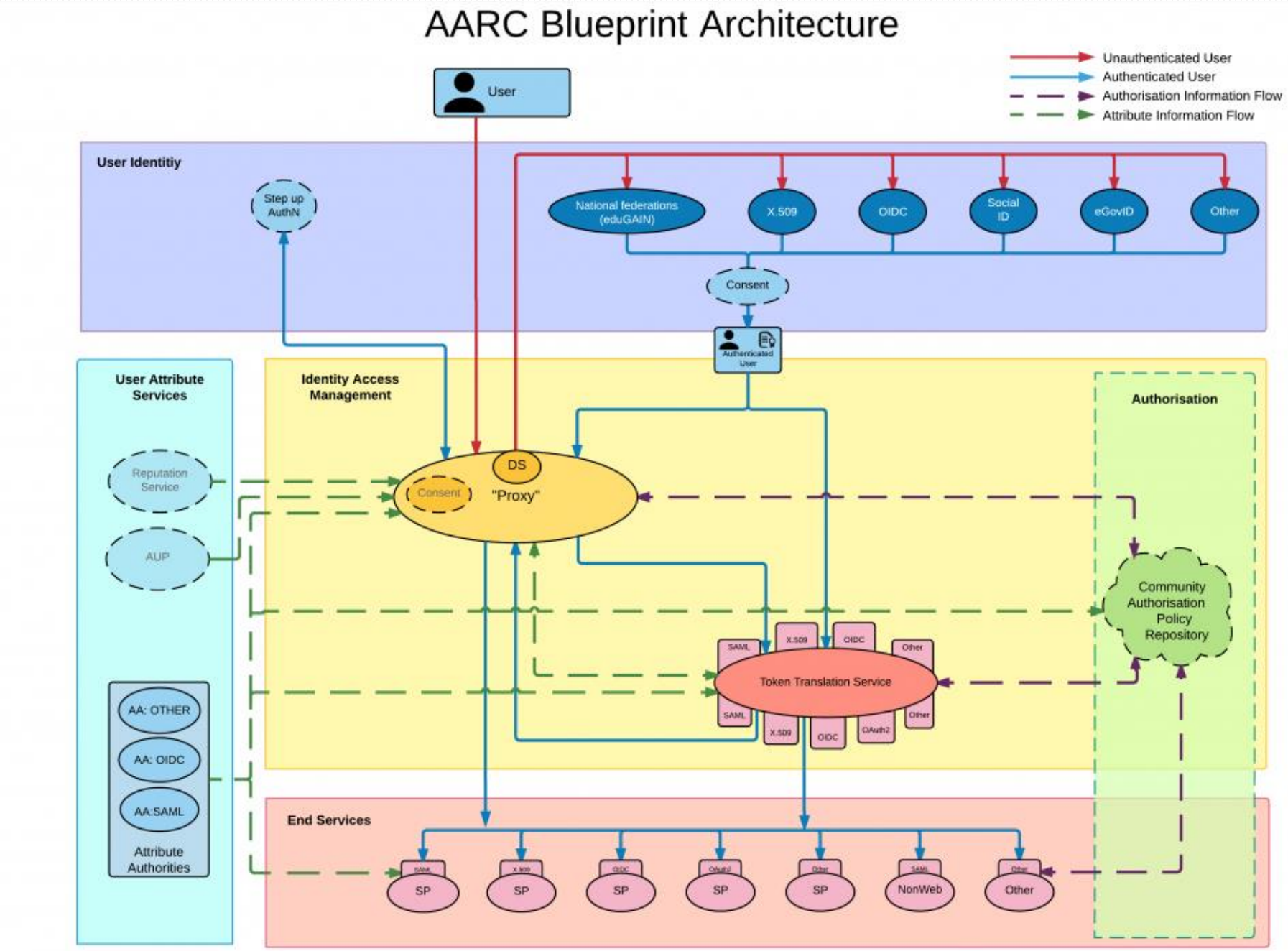
€53M | 2021 - 2024

DiSSCo Prepare

€20M | 2019 - 2023


Researchers
Innovators
Citizen Scientists
Citizens
Reporting Agencies
Companies
Internal Sources
(DiSSCo Facilities)
Experts
Digitisation Centres
Virtual Research Environments
Facility Collection Management Systems
Facility research data repositories
Facility Lab Info Management Systems
Media
Trait data
Specimen/ Occurrence data
Observational data
Laboratory analysis metadata
External Sources
BHL
CoL
GBIF
EOSC Community Support Services
SaaS
DiSSCo Community Services
Services in other RIs
European Collection Objects Index
European Linked Data Porta
European Loans and Visits System
Data Broker API services
Curation & Annotation services
Market place for Facility Services & Expertise
Thematic Service
PaaS
Federation Services
ETL, Monitoring, Linking, PID generation, Taxonomic backbone, data enrichment
Added Value services
Linked data, compute, data broker, single sign-on authentication, authorisation and service management, data preservation, reporting, visualisation
Collaborative services
Unified curation and annotation system, application repository, configuration management, documentation & training resources, marketplace, policies and standards
IaaS
Basic Infrastructure (Cloud Data Storage and Computing infrastructure)
EOSC Common Services
ELViS

AARC: Authentication and Authorisation for Research and Collaboration



DiSSCo
Distributed System of Scientific Collections

A mature and urgently needed initiative

- Europe has the opportunity for scientific leadership at a global level
- Direct response to identified needs in the European and international RI landscape
- Lowers the barrier for big, open science practices across tens of thousands of users
- A super-mature community with 114 self-sustainable facilities (nodes) across 21 countries
- Agreed transfer of authority to a central Hub – governance agreed for all phases

A key missing Research Infrastructure for national and global sustainable development goals

DiSSCo

Distributed System of Scientific Collections