Towards the LifeWatch Big Data Model construction

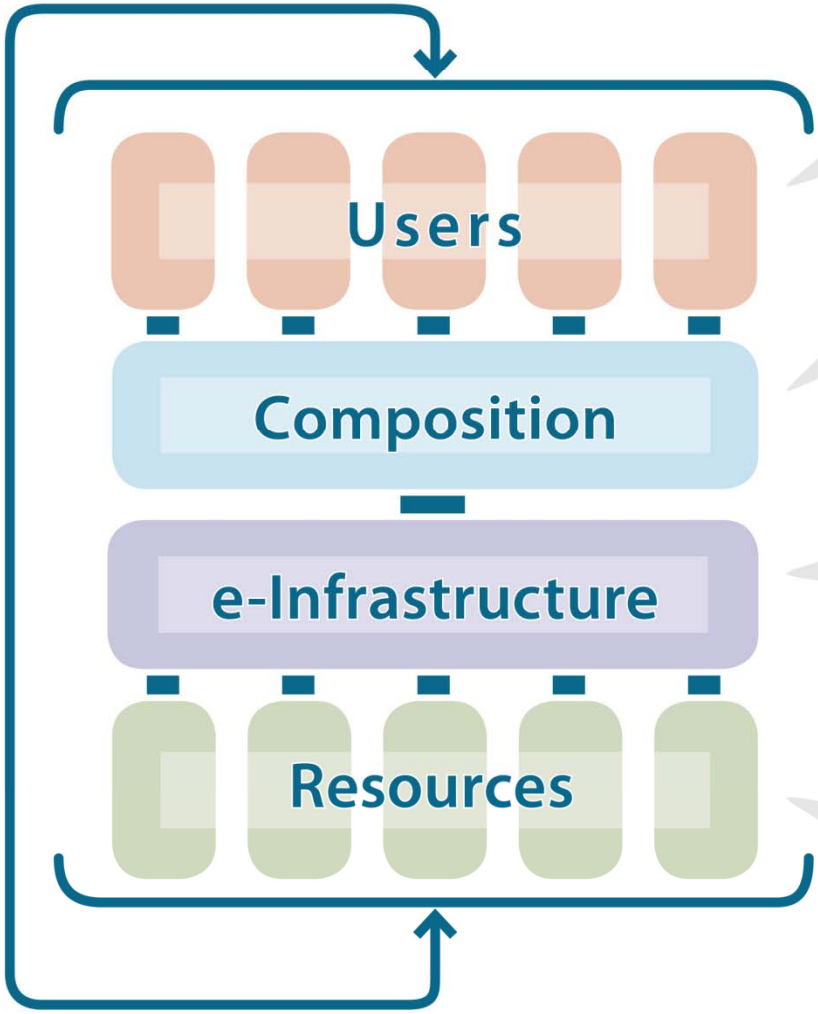
2nd LW e-Infrastructure Construction Operational Meeting
Session: ICT Core construction

prepared by Jesus Marco (marco@ifca.unican.es)
Instituto de Física de Cantabria (IFCA)
CSIC, NATIONAL RESEARCH COUNCIL
SANTANDER, SPAIN
With input from WL,AJ,JMG-A

ICT-Core Starting Tasks

ICT CORE	Start-up activities
Keep Reference Model up to date	Mechanism will be developed. Currently expansion done by ENVRI and EUDAT.
Analysis of requirements	Need requirements from distributed initiatives.
ICT-core technical unit project plan	Proposal will follow with lean organization with coordination and outsourcing capabilities.
Technical framework user portal	Priority for e-science users' portal. Cloud/Grid experiences will assist in drafting proposals.
IT release plan and annual work plans	Will follow (after tasks 26 and 27)
Core basic Application Services	Priorities of core basic application services for the initial years to be proposed.
Organize distributed construction/operations	A management tool will come into place to keep track of distributed activities and relations related to the distributed e-Infrastructure construction/operations. *A technical body will be created.*
Contribute to arrangements with data resources	Test cases to be addressed ((in cooperation with EUDAT, ENVRI, EUBON, LTER and GBIF).
Contribute to enabling data generation	Sensor enabled data generation is being addressed.

"Simple" IT Reference Model



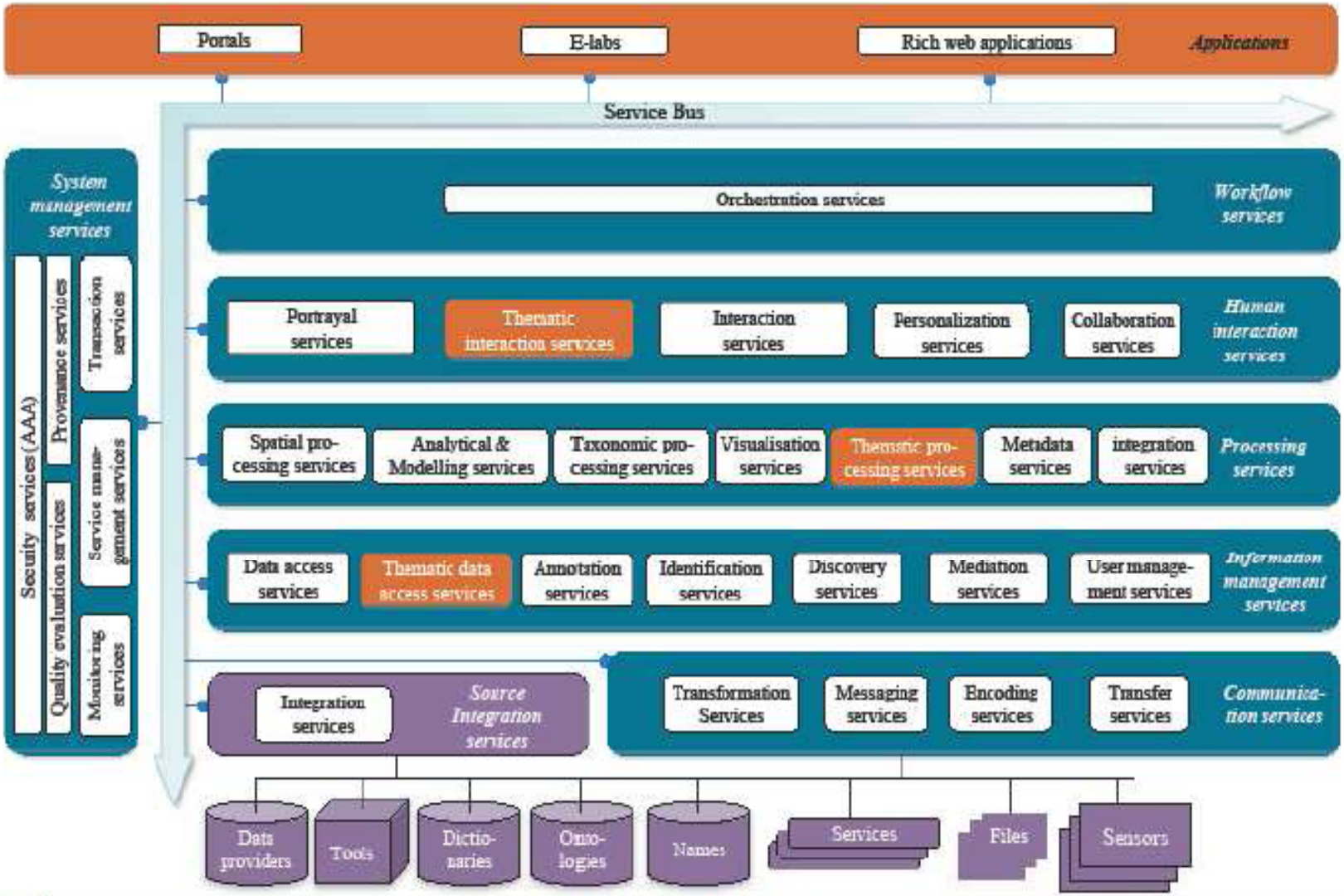
Collaboration
- Common Exploratory Environment
- Collaborative Virtual Organisations

Workflow development
- Semantic Matching
- Visualisation

Analysis & Processing
- Integration of resources
- Quality controls
- Grid computation

Data
- Existing measurements & observations
- Real-time sensor networks (earth based and remote)
- Other infrastructures

Solution for HETEROGENEITY: An SOA approach



How to explore the LW Core-ICT Implementation

As presented at Interministerial in Seville (July 2013)

A SUGGESTED PATH:

- Revise Key Components and Actors
  - Learn from Preparatory Phase and from on-going projects
  - Learn from other Research Infrastructures
  - Interact with all partners in LW
    - Learn, collaborate, **build relationships**
  - **IN ORDER TO CONTRIBUTE TO A REVISED TASK LIST (END 2013)**

- A pilot project to understand the global framework:

Adaptation and improvement of the e-Infrastructure ICTS-EBD (Estacion Biologica de Doñana)

  - Funded by MINECO (CSIC to be commisioned to execute it starting in ~~2013~~) **2014**
    - Setup an operational framework supporting from basic services to advanced data processing and collaborative work
    - Improve the sensor monitoring network at Doñana
    - MATCH & INTEGRATE ICT Services CAPABILITIES IN ANDALUCIA

My naive view of the process to publish a research paper or complete a report in biodiversity

- Data Taking A: Biodiversity specimens observation/collection
  - Manual or Automated (Instrumentation?)
  - Professional or Amateur/Citizens for Museums/Research centers/Organizations
  - Occasional/Campaign or Systematic/Monitoring
  - Integration of annotations/previous references
- **Data Taking A': Genomic information**
- Data Taking B: Environment
  - From external remote monitoring (satellites, radar, LIDAR, etc.)
  - From in-place monitoring (basic to complex sensors and probes, cameras, spectrometers, etc.)
- Data Integration
  - Collections, Papers
  - Databases, Maps
- Data Curation: Identification/Classification, Taxonomy, Integration in GEObase
- **Model**
  - **Specimens evolution, niches, interaction, etc.**
  - **Impact of changes (eg. Environmental, human activities)**
- *Validation, Publication/Report and Design of new experiments*

LifeWatch

NASA, ESA

LTER

GBIF

LW ICT-core "Ecosystem"

- LIfeWatch & National LW Initiatives
- LTER-Europe, LTSER (supported by ALTER-Net)
- GBIF, TDWG
- RDA
- IPBES (Intergovernmental Platform on Biodiversity & Ecosystem Services)
- FAO interest for fisheries and agriculture, AG-Infra, i-Marine, FLOD,
- GEO Ecosystems
- GEOBON genomic layer
- Biosos Earth Observation / EBONE; NATURA2000 sites
- General Ecosystem Models (Predicts, BioVel)
- Ecological Observatories & Genomic Observatories
- Biocode / BiSciCol: VertNet/Genbank
- Microbiome project
- Local Ecological Footprint Tool, Connectivity:www.groms.de
- Ecological Index, BICT, Vibrant
- Catalogue of Life
- Traits: integration of pheno and genotypic data; Phenotype Ontology Research Coordination Network
- BiodiversityDataJournal
- Integrating information using OCR / Vibrant
- Service Networks, Service Sets (deployed on e-Infra) and Biodiversity Catalogue: Integrated Virtual Environment (IVE) for Biodiversity Science (Creative-B)
- Workflows and provenance (Wf4ever, SCAPE)
- Virtual Research Environments (i-Marine, D4Science, gCube)
- Scratchpads (websites for taxonomists)
- OpenAgrid / Agrovoc; data.fao.org
- EnvEurope (semantics and data)
- COMPSs: programming framework for distributed infrastructure
- EUBrazilOpenBio Ecological Niche Modeling Service
- EUBrazilCloudConnect
- New tools for environmental monitoring (Acoustic, Trackers…)
- AAA solutions
- Long Term Preservation (Rebind)
- Ocean Sampling Day
- GeoBroker & A Broker Framework for Next Generation Geoscience (BCube)
- FreshWaterBiodiversity (Mobilizing data and constructing data networks)
- pro-iBiosphere
- PESI
- EUBON
- GN (Global Names)

Reflection on our context

- LW "global" funding is limited
  - Focus on coordination + selected global services
- 1- National initiatives/results must be integrated
- 2- Coordinate with EU/Global initiatives with resources:
  - **E-Infrastructures: EGI, EUDAT, PRACE**
  - Data: GBIF, LTER
  - OTHER ESFRI Initiatives
- 3- Exploit previous/ongoing results from EU projects
- **4- Consider new H2020 opportunities**
- *5- Can we engage SMEs/Industry?*
- *6- What about Public Managers?*
- *7- Can we support Citizen Science?*

Core ICT (e-)Infrastructure

- Essential 'central' components
  - Single portal access for all users
  - Datasets & services / tools catalogues
  - Access to computational resources
  - Security (AAA)
  - Provenance and citation tracking
  - Annotations
  - Virtual Collaborative Environments / VO / BTCN
  - Workflow composition, execution and management
- Data & tool resources
  - New data resources to be 'admitted'
  - Statistical, analytical & modelling tools
- Innovation Lab
- Intellectual property management

EGI services for LW:

- MODEL: LW brings users & resources together!
  - LW core-ICT (Spain) will operate an e-infrastructure in 2014
    - **LW core-ICT could/will integrate grid/cloud infrastructure in EGI**
    - **LW VOMS will be supported by LW core-ICT**
    - **LW core-ICT will rely on IberGrid for this integration in EGI**
  - LW national initiatives will be integrated
    - LW core-ICT will support integration at different levels (NGI role?)
  - LW will explore successful examples in EGI FedCloud:
    - EUBrazilOpenBio Ecological Niche Modeling Service
    - EUBrazilCloudConnect
    - **New challenge for phenology with LTER/Univ.Granada**
- So, LW will use existing EGI services

EUDAT services for LW?

- EUDAT has two key sides for LW:
  - Knowledge about DATA management
  - New services:
    - B2 SHARE
    - B2 STAGE
    - B2 FIND
- LW can/should explore them!
  Contribute to Proposals Call (befor 26th Feb!)
- Select topic(s):
  - Real Time
  - Semantic Mapping
  - Workflow Execution
  - Data Lifecycle

PRACE services for LW?

- Supercomputing framework requires
  - Large computing challenge
  - Preparation, well in advance, of proposals
- LW should identify challenges, and promote proposals!
  - Proposals can be pre-tested at national/regional resources!
- LW could also contribute to projects integrating HPC
  - Integrating large data repositories with cloud/grid to HPC
  - Workflow experiences

Additional services

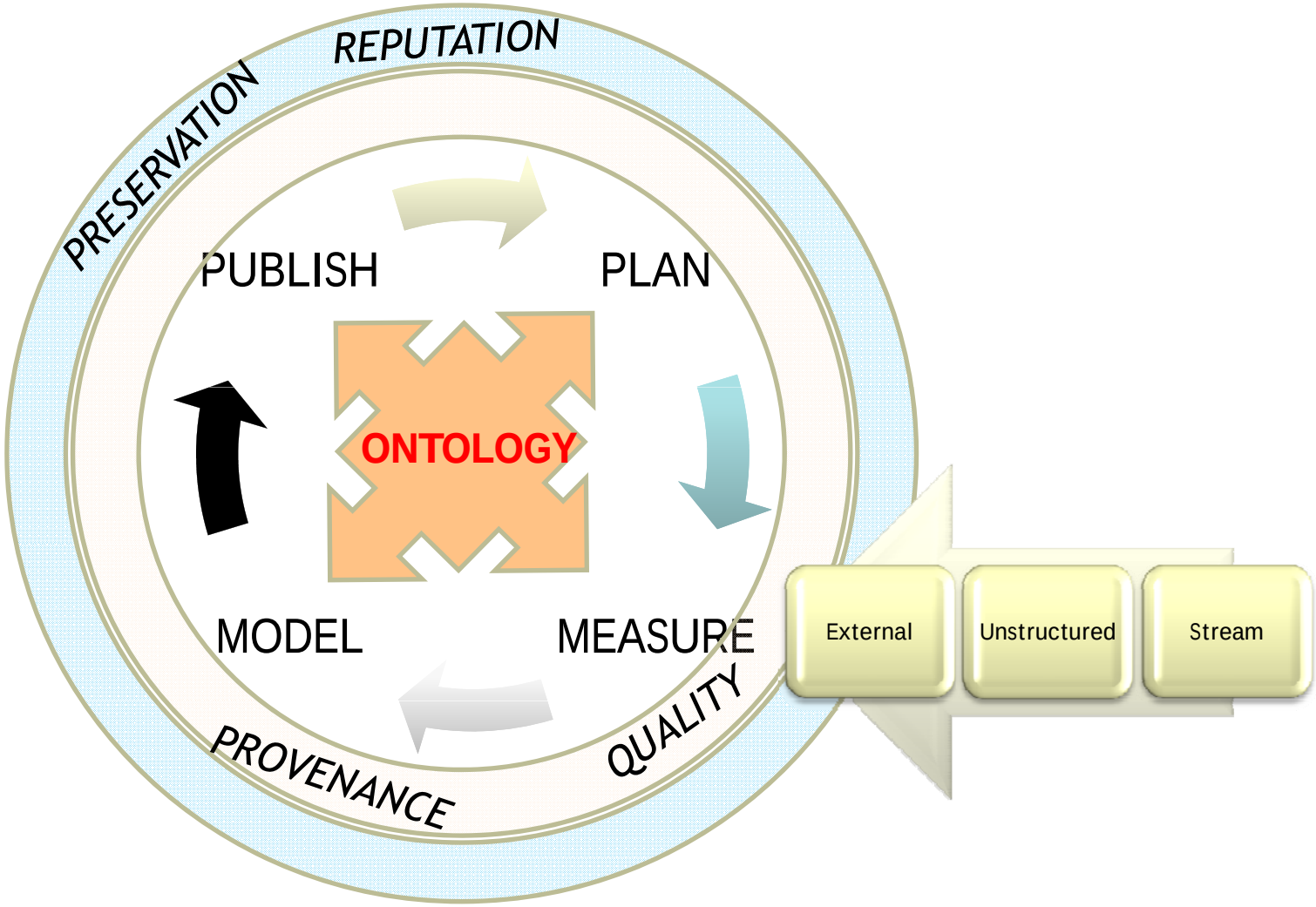
- Additional services are being studied:
  - Considering also output of ongoing projects
    - ENVRI, BIOVEL, COOPEUS, iMARINE, CREATIVE-B
  - Some of them:
    - Identity federation for researchers, educators and students
    - Digital Identifier e-Infrastructure for digital objects (and PID issues)
    - Simple Storage/File System + Medium/Large DBMS cloud/grid instances
    - Large, persistent DBMS, GIS systems in cloud/grid framework
    - Parallel (multithread?) datamining (in phyton and R) cloud/grid instance
    - Systems to handle & process real time streams
    - Access to large databases/directories common to other research areas
    - Workflows connecting to HPC resources (o(10^2-10^3) processes, 1-100 TB)
    - Support to virtual eLaboratory
    - Data discovery and access
- Along 2014 we need to work to complete a VRE proposal

BIG DATA topics

H2020 opportunities

- EINFRA-2014-1: e-Infrastructure for Open Access
  - Community Knowledge Tool
- EINFRA-2014-2:
  - Manage/Preserve Big Research Data
  - RDA
  - HPC
- INFRAIA-2014-2015:
  - support new communities (ex. LTER sites? NETLAKES?…)
- INFRAIA-2015-1: New skills
- INFRASUPP-2014-2: International Collaboration
- INFRADEV-1-2014: ESFRI Clusters
- INFRADEV-3-2015: ESFRI projects operation
- ICT-2014-1: Cloud
- ICT-2014-2: 5G Network

Closing the Knowledge Loop



Can we address a challenge?

- Grand Challenge: **Predictive Modeling of Biosphere**
  - **Global Carbon cycle** **JOIN us at AGU meeting (29 April, Vienna)**
  - Essential Biodiversity Variables (EBV) for IPBES
    - IPBES=Intergovernmental Platform on Biodiversity & Ecosystem Services (cf. IPCC)
    - EBV= a measurement required for study, reporting, and management of biodiversity change.
  - Examples of candidate EBV:
    - Species populations: Abundances and distributions (inc. invasive alien)

EXAMPLES OF CANDIDATE ESSENTIAL BIODIVERSITY VARIABLES

EBV class	EBV examples	Measurement and scalability	Temporal sensitivity	Feasibility	Relevance for CBD targets and indicators (*1,9*)
Genetic composition	Allelic diversity	Genotypes of selected species (e.g., endangered, domesticated) at representative locations.	Generation time	Data available for many species and for several locations, but little global systematic sampling.	Targets: 12, 13. Indicators: Trends in genetic diversity of selected species and of domesticated animals and cultivated plants; RLI.
Species populations	Abundances and distributions	Counts or presence surveys for groups of species easy to monitor or important for ES, over an extensive network of sites, complemented with incidental data.	1 to >10 years	Standardized counts under way for some taxa but geographically restricted. Presence data collected for more taxa. Ongoing data integration efforts (Global Biodiversity Information Facility, Map of Life).	Targets: 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15. Indicators: LPI; WBI; RLI; population and extinction risk trends of target species, forest specialists in forests under restoration, and species that provide ES; trends in invasive alien species; trends in climatic impacts on populations.
Species traits	Phenology	Timing of leaf coloration by RS, with in situ validation.	1 year	Several ongoing initiatives (Phenological Eyes Network, PhenoCam, etc.)	Targets: 10, 15. Indicators: Trends in extent and rate of shifts of boundaries of vulnerable ecosystems.
Community	Taxonomic	Consistent multitaxa surveys and	5 to >10	Ongoing at intensive monitoring sites	Targets: 8, 10, 14.

Pereira et al., Science 2013

How to move?

- Next presentations will show more information needed/discussion
- Tomorrow discussion
- Plan for basic start setup
  - Minimal central services
  - Integrating existing national services
    - Examples: LW Sweden, LW Belgium
    - …
  - Coordination with
    - Italy Service Center, Netherlands
    - EGI, EUDAT, PRACE
    - GBIF, LTER
    - Other ESFRI
    - International Collaborations
- Organization:
  - Task teams? IC3?…
- Knowledge map
  - **FINAL USERS**
  - **EXISTING WORK AND EXPERTISE AT ICT-BIODIVERSITY LEVEL**

THANKS!