# EGI & LifeWatch: first steps together

*EGI Community Forum 2014, HELSINKI*
*Session: Going beyond grid to enable life science data analysis*

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

# *What is LifeWatch???*

**LifeWatch is an ESFRI initiative in biodiversity and ecosystem research**

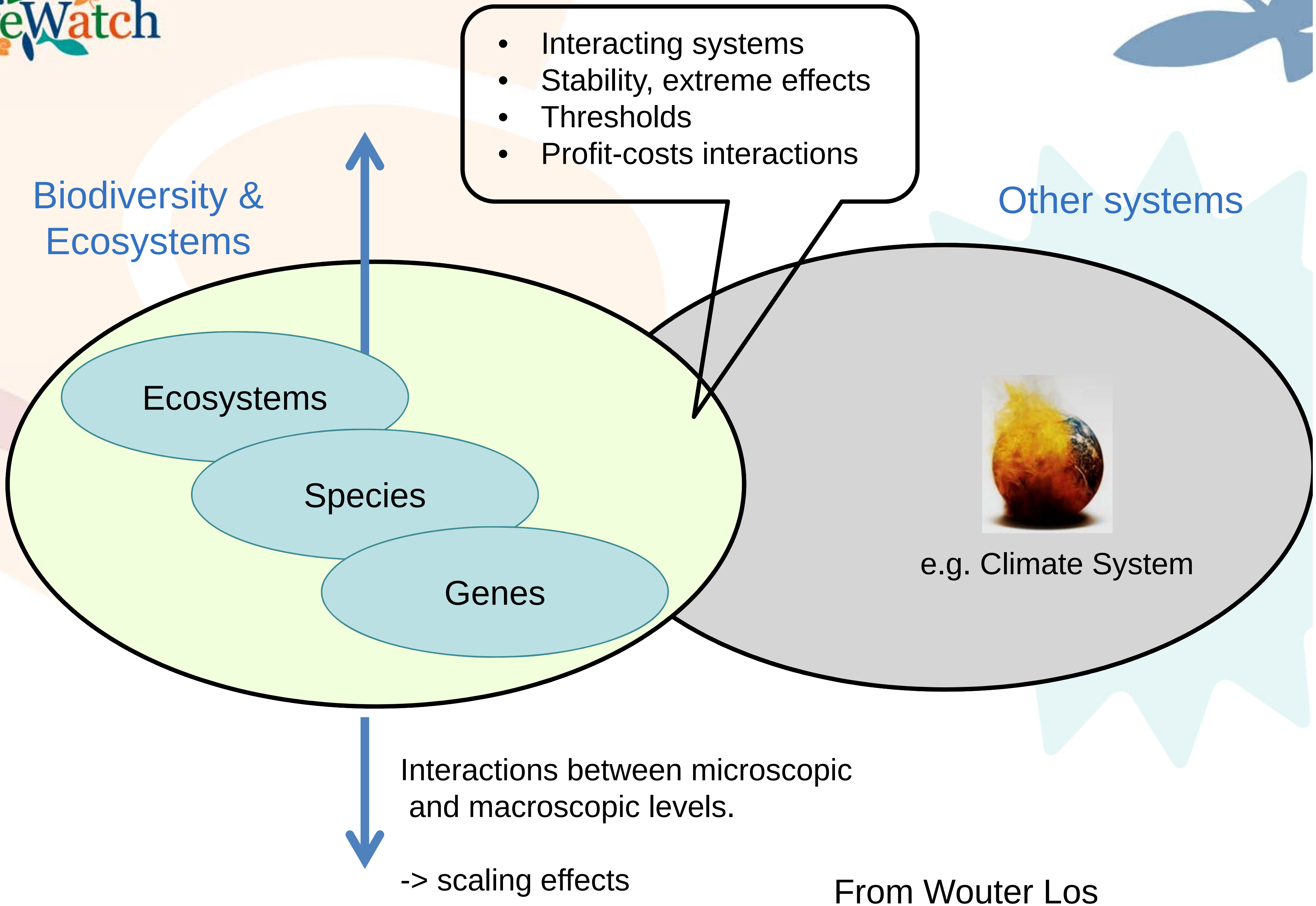
**An** *exploratory research environment that allows scientists to find data, combine resources, compose workflows, run analyses, develop models and visualize predictions.*

*User-friendly services to combined data, using established standards and unique identifiers are developed that can be further combined into work flows for different types of questions.*


**http://lifewatch.eu/**

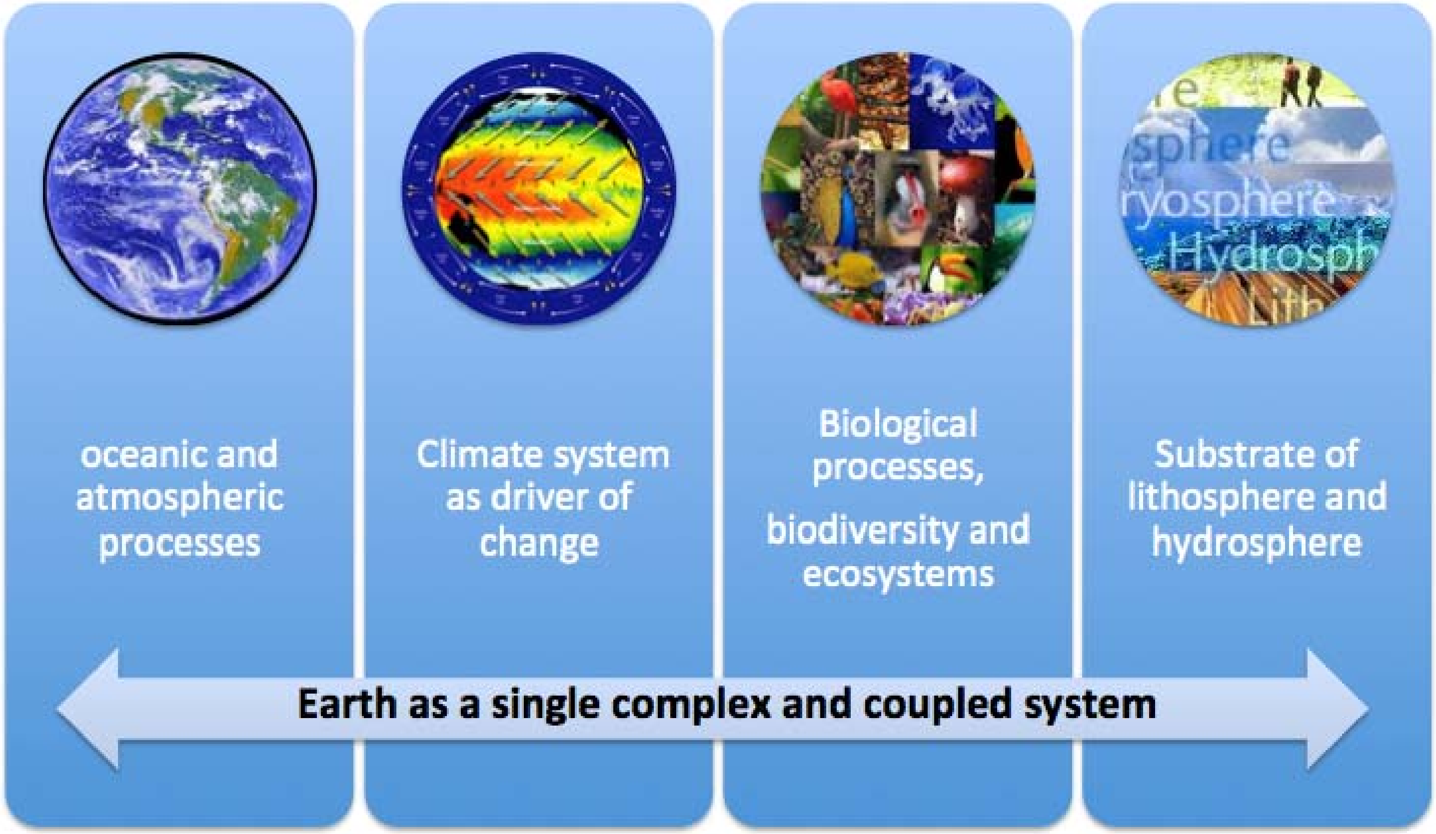
*Flexibility to be offered through grid- or cloud-based computational services. While the facility takes care of access and matching of resources 'under the hood', scientists and other users can focus on the analyses.*

LifeWatch




From Wouter Los presentation in Madrid

# Scientific challenges




From Wouter Los presentation in Madrid

AN EXAMPLE SELECTED ON PURPOSE:

# Monitoring Cyanobacterial Blooms

**ECOHYDROS SL** for **Confederación Hidrográfica del Duero (CHD)**
DORII (FP7 project with IFCA-CSIC), ROEM & ROEM+ (Avanza & LIFE projects with ITG)

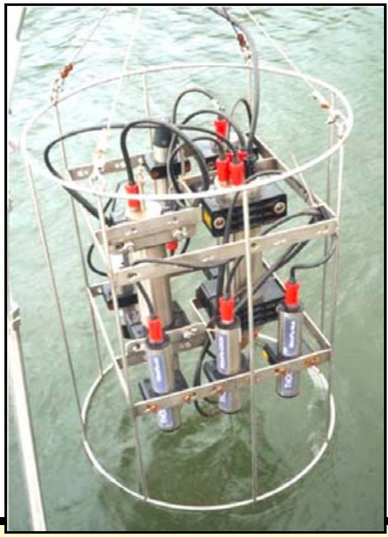
## New generation of smarter complex sensing platforms

**Advanced water quality measurement**

**Relevant environmental measurements**
- Atmospheric-meteorological variables
- Hyperspectral solar radiation in water
- Nutrients

**High sampling frequency** (minutes/hours) + data transmission to cloud services (replicated database, **on-line monitoring system**)

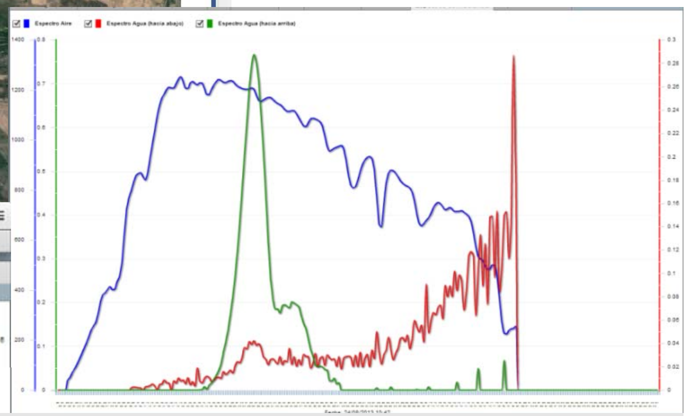
Mobility (GPS integrated) plus **Profiling** (wincher programmed/remote down to 50m)



(Coterillo, 2012).

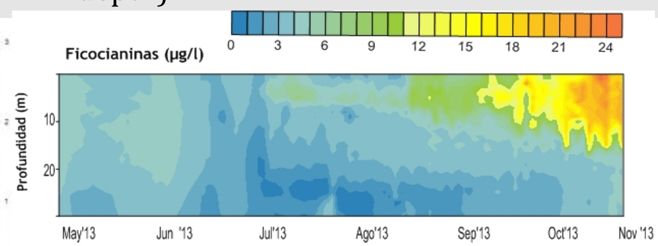
# ONLINE ACCESS (data, profiles, images) USING FEDCLOUD RESOURCES AT IFCA


**High res thermocline**
(5 sec sampling)



Radiation Spectrometers (4 nm)
(air=blue, underwater up=green, down=red at 12m. depth)


Cyano bloom in 2013, example of a profile (fluorometers, cyano=bars)




DxT graph showing cyanobacteriae concentration from profiler data

THE CHALLENGE:

# INTEGRATION

# of information and models

# using e-infrastructures

# for advance water reservoir management

**INFORMATION:**

**-DMP (Data Management Plan)** integrating internal and external information, key for preservation (PID).

**-GIS system**

**MODELS:**

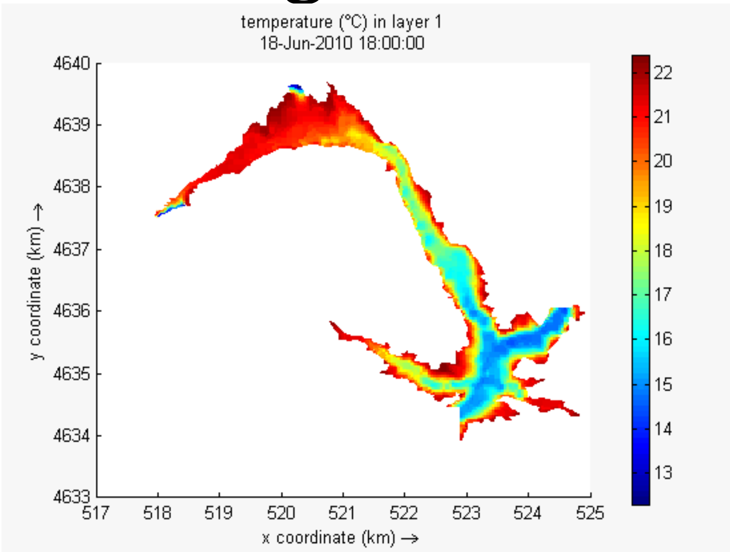
**-DELFT3D (open model)**

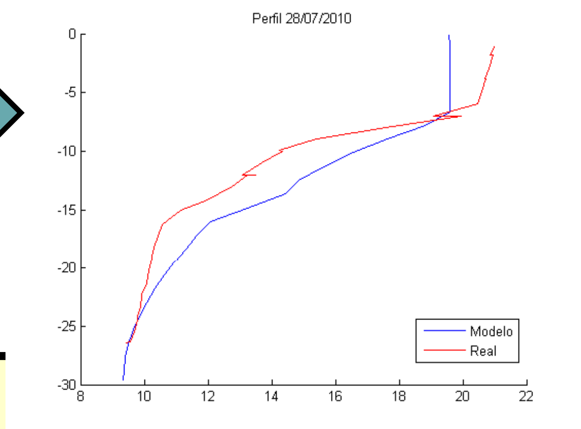
**Water reservoir modeling**

***Running: Cloud + Supercomputer***

**-physical (abiotic) VALIDATED !**

**-water quality (biotic) Working on it**

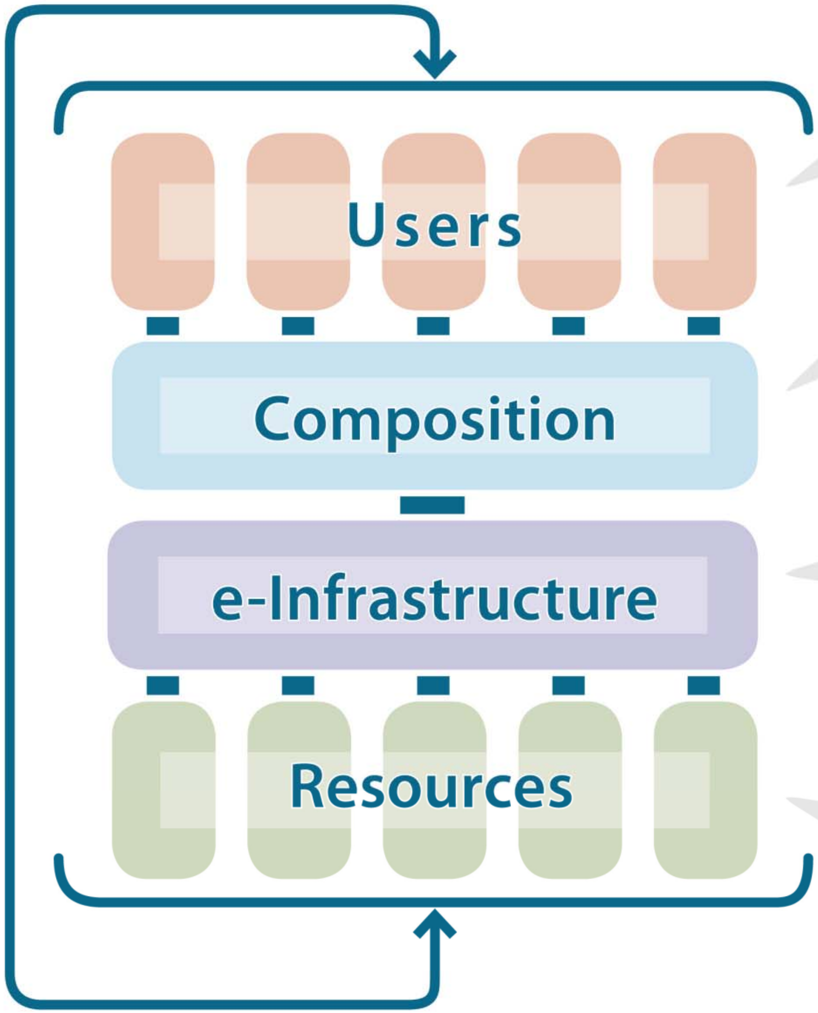



**DECISION/MANAGEMENT TOOLS**

**NEXT STEP!**

# “Simple” IT Reference Model




**Collaboration**
- Common Exploratory Environment
- Collaborative Virtual Organisations

**Workflow development**
- Semantic Matching
- Visualisation

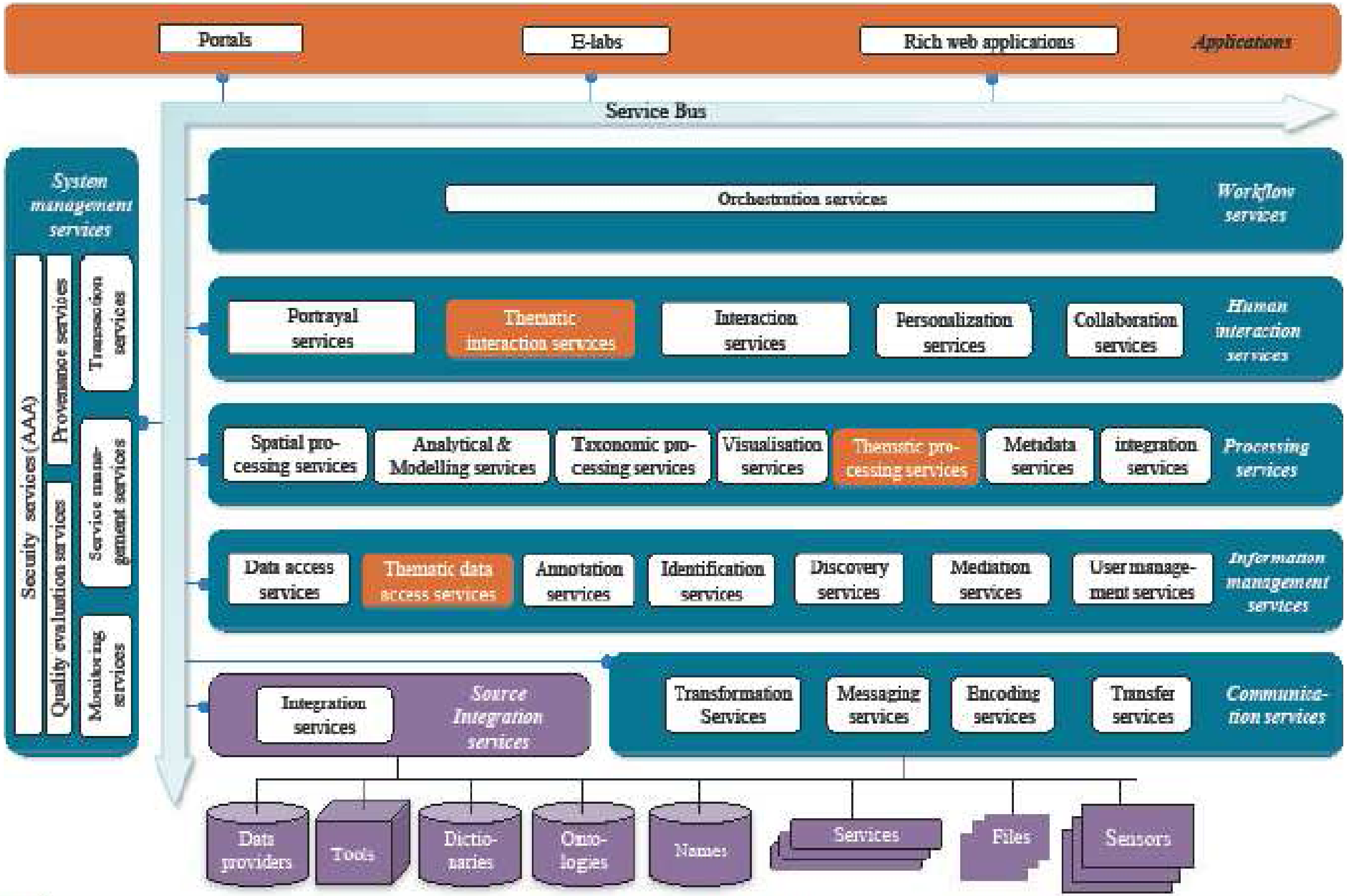
**Analysis & Processing**
- Integration of resources
- Quality controls
- Grid computation

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

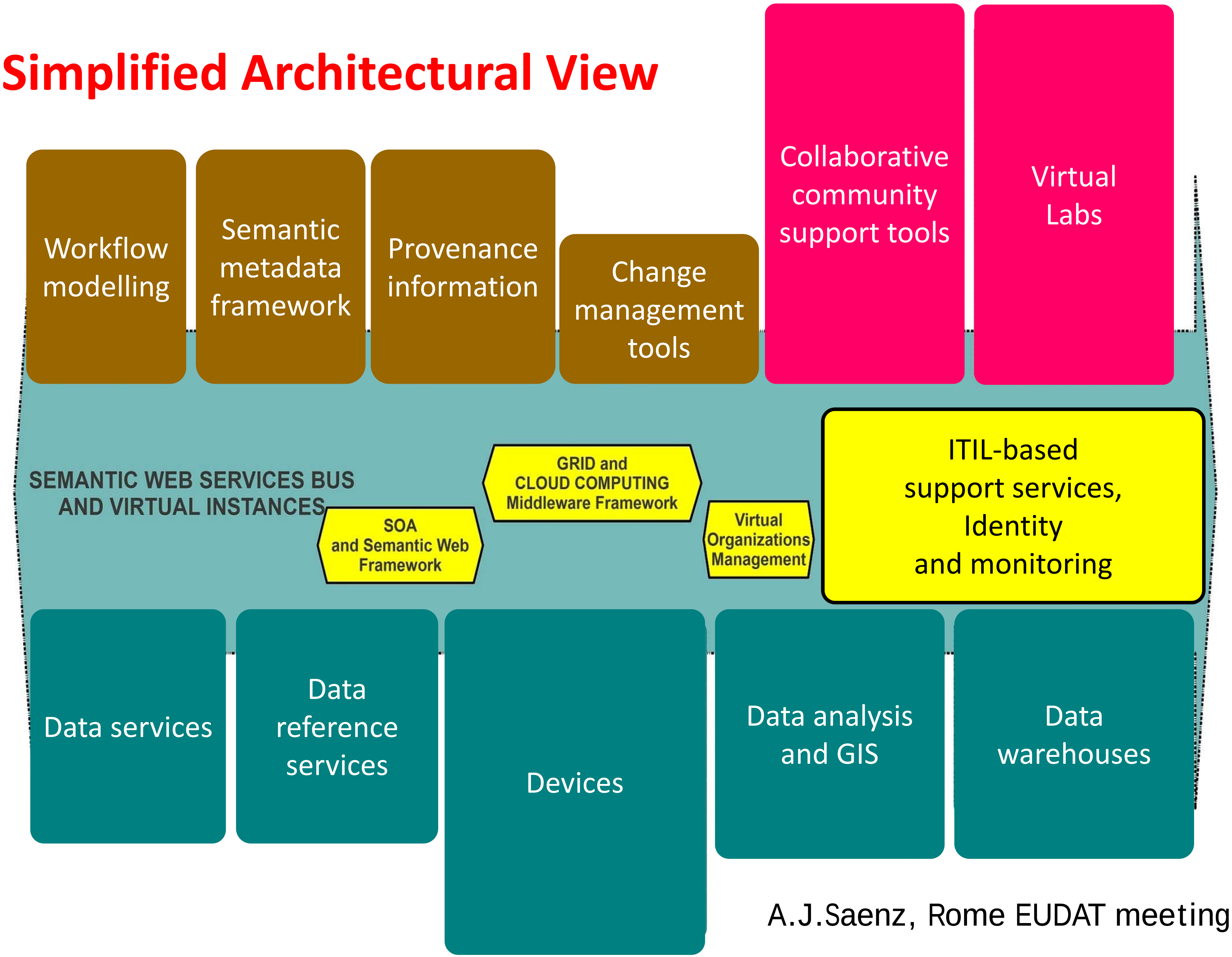
# Core ICT (e-)Infrastructure

- 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

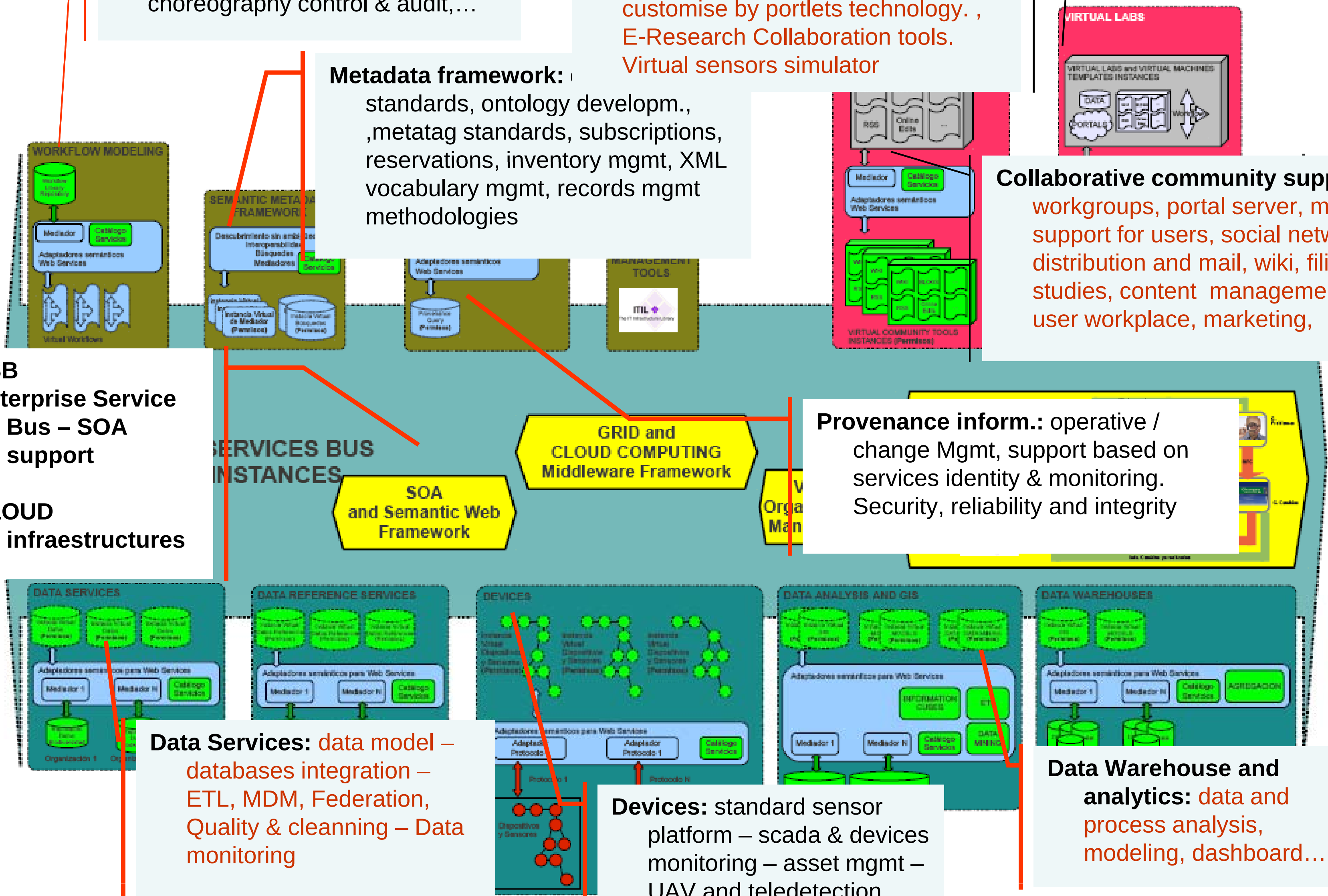
# Solution for HETEROGENEITY: An SOA approach



# Simplified Architectural View




A.J.Saenz, Rome EUDAT meeting




**Workflow modeling:** process modeling, BPM, process trigger, choreography control & audit,…

**Virtual Laboratory:** resources for data analysis, modeling and case management. Virtual machines and sharing resources. Portal Server customise by portlets technology. , E-Research Collaboration tools. Virtual sensors simulator

**Metadata framework:** standards, ontology developm., ,metatag standards, subscriptions, reservations, inventory mgmt, XML vocabulary mgmt, records mgmt methodologies

**Collaborative community support:** workgroups, portal server, multichannel support for users, social networks, distribution and mail, wiki, filing, case studies, content management, CMS, user workplace, marketing, …

**ESB Enterprise Service Bus – SOA support – CLOUD infraestructures**

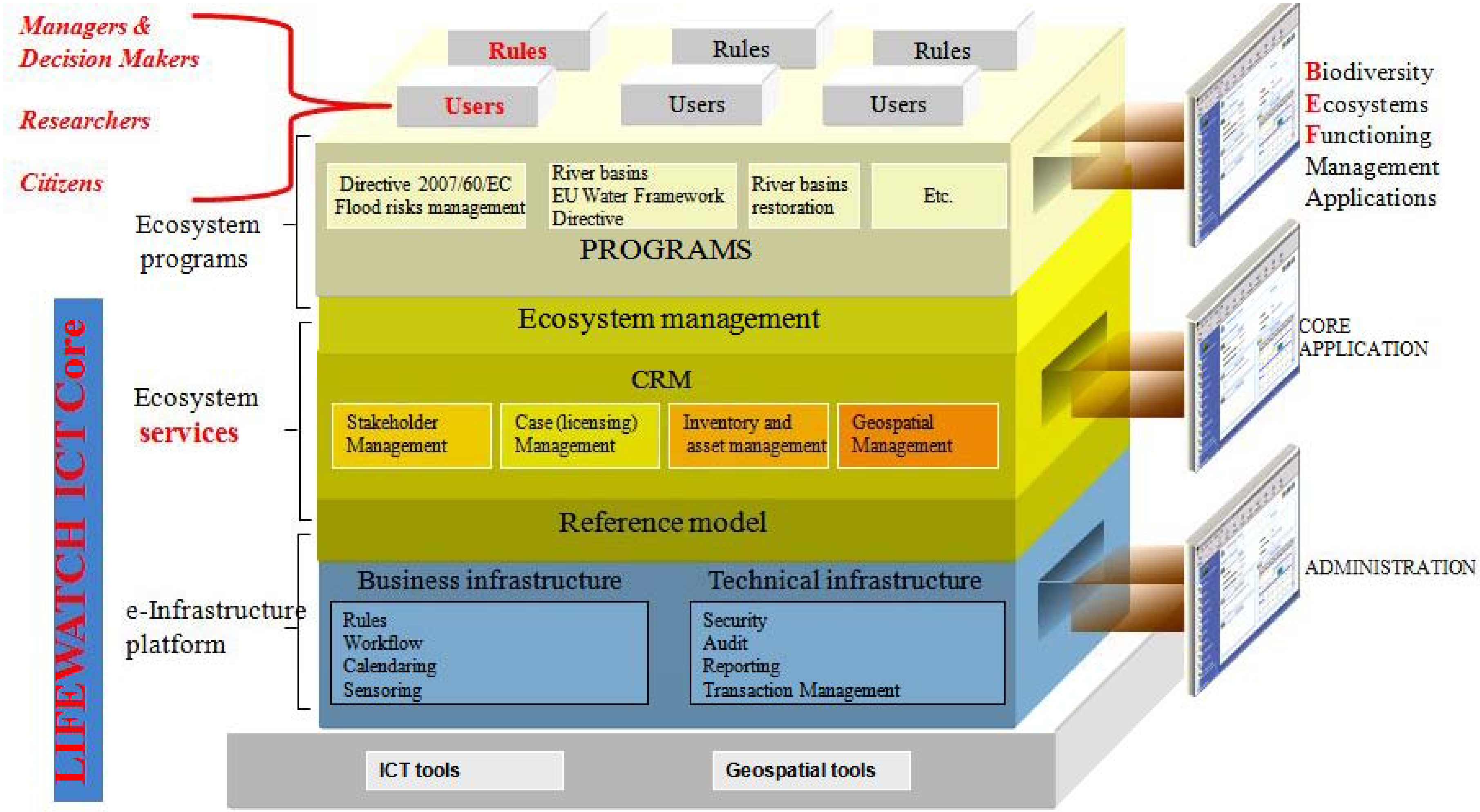
**Provenance inform.:** operative / change Mgmt, support based on services identity & monitoring. Security, reliability and integrity

**Data Services:** data model – databases integration – ETL, MDM, Federation, Quality & cleanning – Data monitoring

**Devices:** standard sensor platform – scada & devices monitoring – asset mgmt – UAV and teledetection support

**Data Warehouse and analytics:** data and process analysis, modeling, dashboard….

# Our scheme must take into account our stakeholders: researchers, citizens & **managers**



# *How to explore the LW Core-ICT Implementation*

As presented at LW 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

# 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. |

# *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?*

# *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
  - But LW needs additional services…

# Additional services

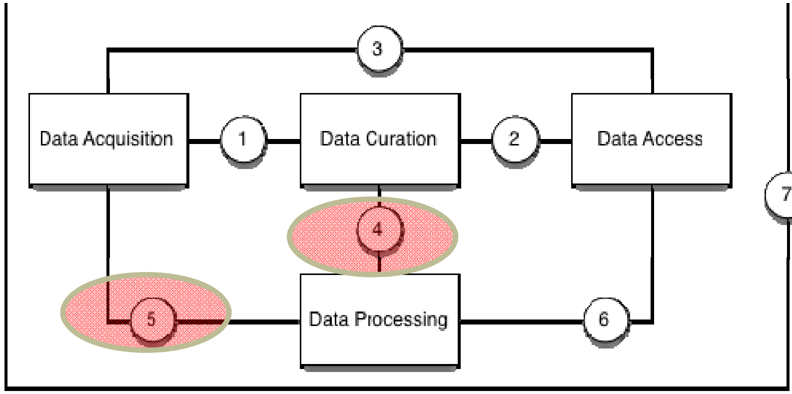
- Additional services (related to EGI) are being studied:
  - Considering output of ongoing projects
    - EUDAT, ENVRI, BIOVEL, COOPEUS, iMARINE, CREATIVE-B
  - Some of them may be implemented with EGI support/collaboration
    - Identity federation for researchers, educators and students
    - Digital 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

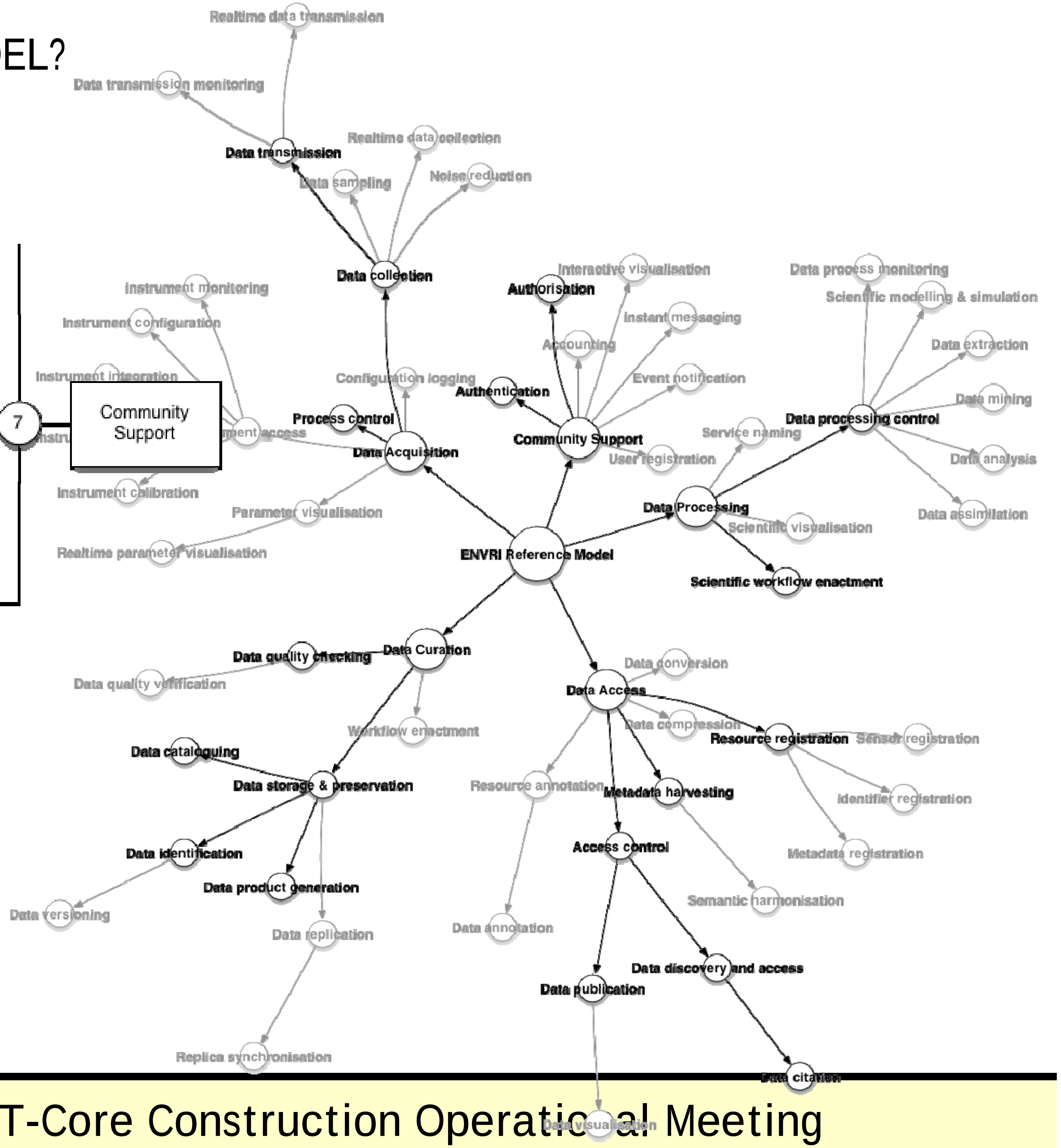
BIG DATA topics

- Along 2014 we will work to complete a VRE proposal

# *A Reference Model?*

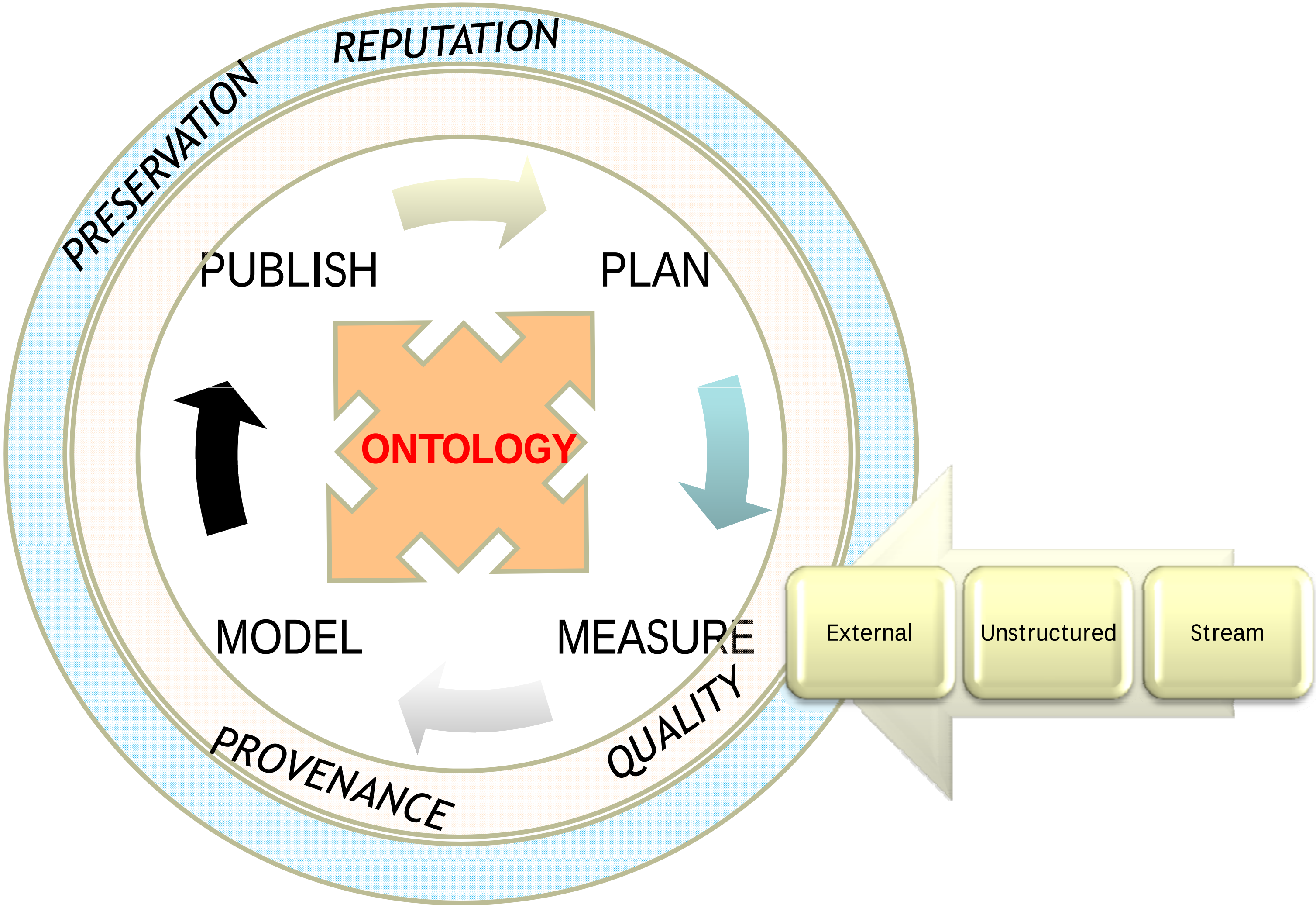
- ORGANIZE AROUND A DATA MODEL?
  - **ENVRI** example
  - Identify key processes




- Functions map

# Closing the Knowledge Loop



# Can we address a challenge?

- Grand Challenge: **Predictive Modeling of Biosphere**
  - **Global Carbon cycle** **DISCUSSED 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

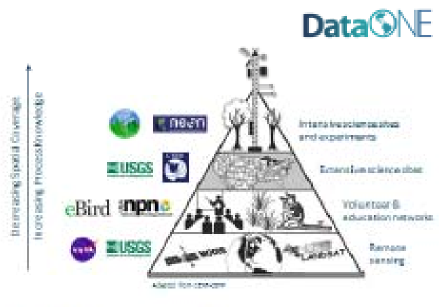
# Phenology and the land surface

### General questions

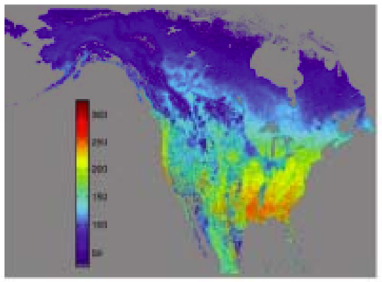
- How does the sensitivity of phenology to climate change vary across biomes?
- What are the vegetation-climate feedbacks that are mediated by phenology?
- What is the optimal way to observe phenological impacts on albedo, surface roughness, canopy conductance, and its impacts on fluxes of carbon, water and energy?
- How will these feedbacks be impacted by shifts in phenology caused by climate change?

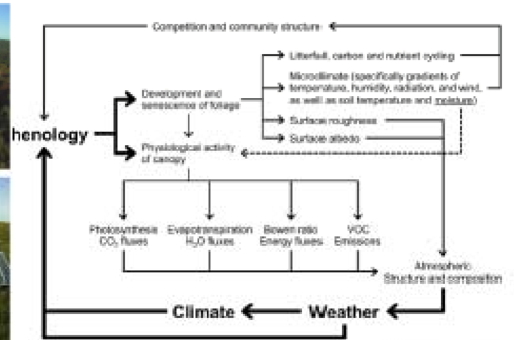
### Challenges/Opportunities

- In many ecosystems we still need better understanding of how environmental drivers control phenology – the role of photoperiod is still not well understood
- Much work has concentrated on start of growing season, more data are required on the end of growing season
- How can conceptual understanding of feedbacks on climate be translated into quantitative estimates of the impact of phenological shifts?
- What data are required to optimally improve phenology modeling?
- How can information from ground-based and space borne observing platforms be optimally combined?



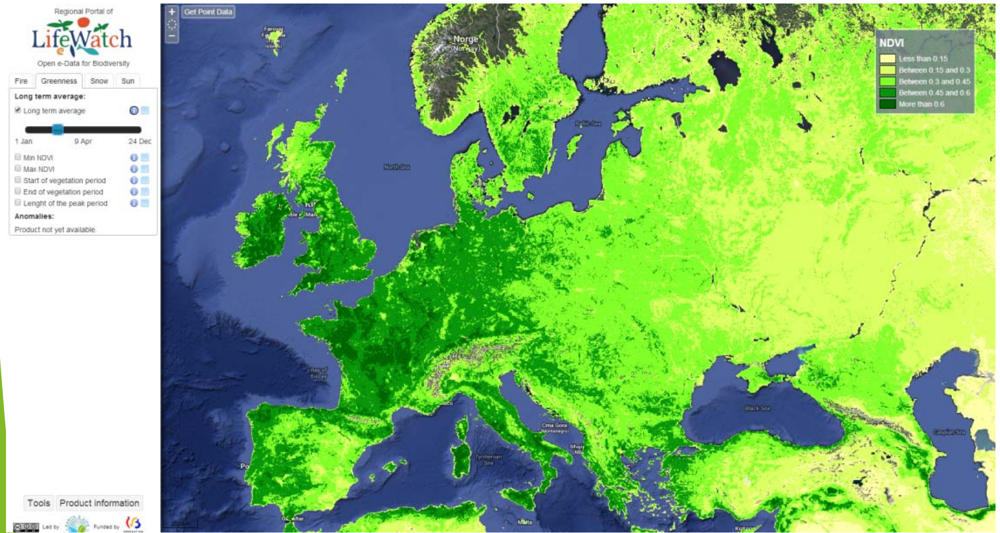

Bob Cook NORTH AMERICAN CARBON PROGRAM



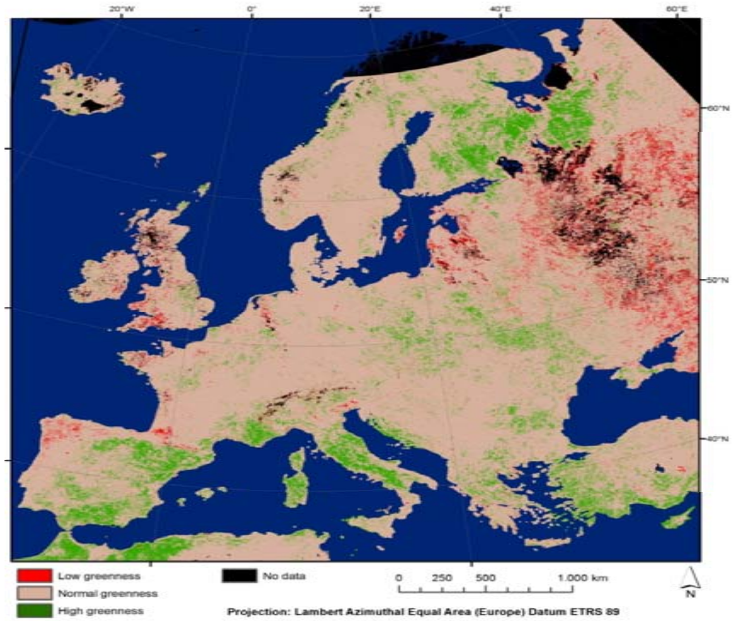



Richardson et al., 2013

# Overview of the growing season at European scale




Long term average




Anomalies

# *A final message from some old friends…*


**We don't like this new home!**

# *Backup slide*


From EOL through GBIF

**Our old friends were… eels!**