



Research Data Sharing
without barriers

Access to **open** research data for environmental science: the **LifeWatch** experience in biodiversity

Advancing data-driven research through the Data Commons Session at RDA P6 PARIS



www.esi.eu



Presented by
Jesús Marco de Lucas
IFCA-CSIC, Spain



EGI-Engage is co-funded by the Horizon 2020 Framework Programme
of the European Union under grant number 654142



- **LifeWatch (lifewatch.eu) is an ESFRI** (*EU Research Infrastructure*)
 - Addressing Biodiversity & Ecosystems
 - An e-Infrastructure to build Virtual Research Environments (VRE)
 - Integrating **OPEN DATA** information
 - GBIF, LTER, GENBANK, SATELLITE IMAGES, TERRESTRIAL MAPS...
- **EGI-LifeWatch Competence Center**
 - Framework: EGI FedCloud
 - Dedicated Resources (~5000 cores + few PB, new node in Seville)
- **Support LW VRE**
 - Marine VRE (marine.lifewatch.eu)
 - Terrestrial + FreshWater VRE
- **Pilot projects**
 - Ecological Observatories Data Flow and “Big Data” analysis
 - Workflows: Galaxy and TRUFA; Network of Life
 - Citizen Science: Assisted Pattern Recognition

INITIAL REMARKS (from experience)

- BIODIVERSITY area has a rich biodiversity itself!
 - Large variety of actors in the community with different background
 - researchers, technicians, consultants, managers, etc.
 - Large variety of Use Cases
- Each Case Study requires a substantial effort
 - **We need to improve the tools to communicate ICT and Users**

Everything Should Be Made as Simple as Possible, But Not Simpler

- TWO Different Views on **OPEN** RESEARCH DATA
 - Institutions, Administration, Researchers, Companies
 - Institutions, Administration, Researchers, Companies
- **This is changing...slowly!**

Our customers are our researchers!

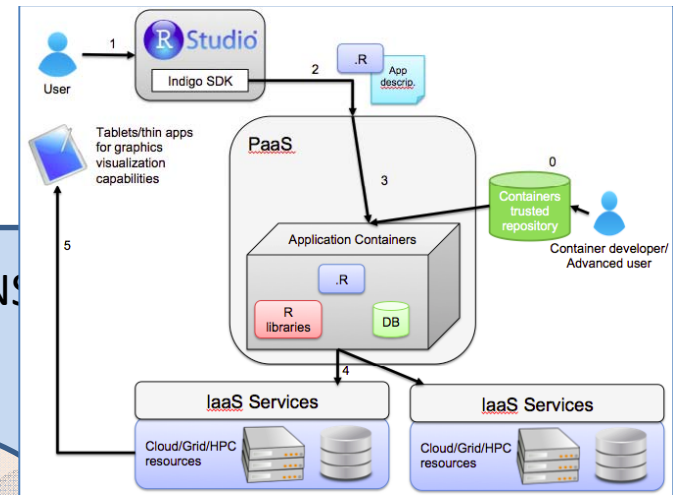
- Our researchers need:
 - ACCESS **INTEGRATE EXISTING OPEN RESEARCH DATA**
 - Catalogs, Links, Connect (WPS), Local Copies...
 - Internal/External Tools for **processing/analyzing**
 - **COMPUTING+STORAGE** RESOURCES AT O(+100)
 - Typically 100 cores, 100 TB
 - STORE **PRESERVE NEW ; OPEN ? RESEARCH DATA & ANALYSIS**
 - Acquisition, Collection, Intermediate Data, "Final" Data: FULL DATA LIFE CYCLE
 - **COMPUTING+STORAGE** RESOURCES AT O(+10)

Everything Should Be Made as Simple as Possible, But Not Simpler

*Example: give our researchers a 100 cores/ 100TB laptop
Is this a solution to their needs?*



Global Scheme



USER APPLICATIONS

Portals
Visualization
Liferay

Platform for Software as a Service

Our customers should not care!

Open Data & Preservation Platform (ODPL)
Orchestra

Distributed Control Platform

SOA/ Cloud Computing

Infrastructure as a Service (IaaS)
(integration in EGI FedCloud)

CO

e-INFRASTRUCTURE

RESOURCES

“External”
Data

“Internal”
Data

SITE

Network

Storage

Servers

SITE

...

TRUFA (Transcriptomes User-Friendly Analysis)

TRUFA is a **free** web service to **help you** perform RNA-seq analysis

– **INTEGRATE EXISTING OPEN RESEARCH DATA**

- LOCAL REPLICA OF PUBLIC OPEN DATABASES

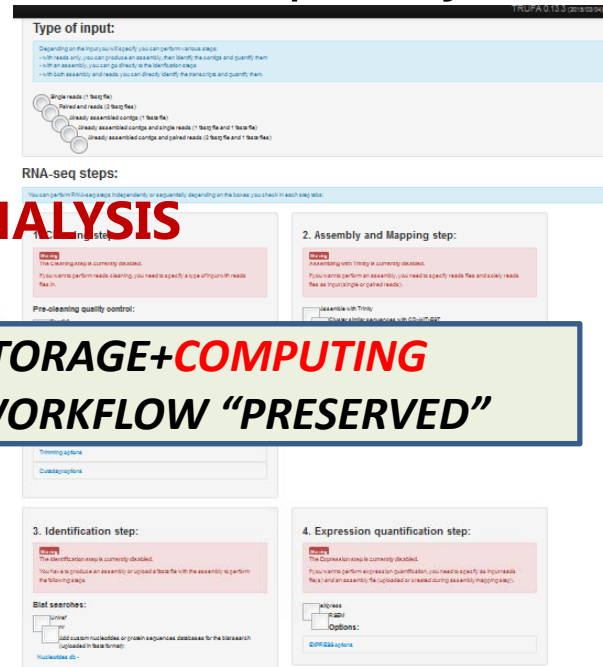
– **PRESERVE NEW OPEN RESEARCH DATA & ANALYSIS**

- UPLOAD USER DATA FILES (Large Files)
- **COMPUTING+STORAGE** RESOURCES AT O(+100)
 - NOW SUPERCOMPUTER WITH GPFS

STORAGE+COMPUTING WORKFLOW "PRESERVED"

LARGE SUCCESS, ALSO LARGE PROBLEM

- >150 USERS ALL OVER THE WORLD IN 3 MONTHS



736874	0_split4	genorama	PENDING	0:00	30:00	4	1	(Dependency)
736875	0_cblat_	genorama	PENDING	0:00	3-00:00:00	64	4	(Dependency)
736878	0_cblat_	genorama	PENDING	0:00	3-00:00:00	64	4	(Dependency)
736879	0_clean1	genorama	PENDING	0:00	30:00	1	1	(Dependency)
736880	0_fastqc	genorama	PENDING	0:00	3:00:00	4	1	(Dependency)
736881	0_split4	genorama	PENDING	0:00	1:00:00	4	1	(Dependency)
736882	0_trinit	genorama	PENDING	0:00	3-00:00:00	16	1	(Dependency)
736883	0_hmm_sc	genorama	PENDING	0:00	3-00:00:00	96	6	(Dependency)
736884	0_cdhit	genorama	PENDING	0:00	3-00:00:00	16	1	(Dependency)
736885	0_qc_bla	genorama	PENDING	0:00	1-00:00:00	16	1	(Dependency)
736886	0_clean1	genorama	PENDING	0:00	30:00	1	1	(Dependency)
736887	0_cegma	genorama	PENDING	0:00	1-00:00:00	16	1	(Dependency)
736888	0_blastp	genorama	PENDING	0:00	3-00:00:00	256	16	(Dependency)
736889	0_bowtie	genorama	PENDING	0:00	1-00:00:00	16	1	(Dependency)
736890	0_merge	genorama	PENDING	0:00	1:00:00	8	1	(Dependency)
736891	0_b2go_b	genorama	PENDING	0:00	3-00:00:00	16	1	(Dependency)
736892	0_bam_	genorama	PENDING	0:00	1-00:00:00	16	1	(Dependency)

Storage is limited, Local files (2 months, o(TB)) Faster Transfer needed

EGI-LIFEWATCH CC EXAMPLE 2

Support to Citizen Science (Assisted Image Recognition)

Citizens collect and upload geo-pos image observations of species using a mobile app (iNaturalist), the image is stored and an initial identification returned

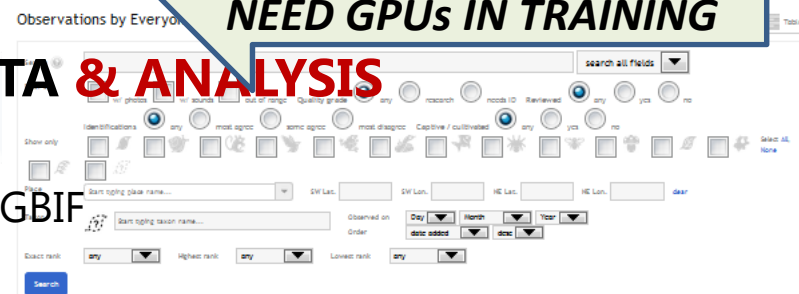
– **INTEGRATE EXISTING OPEN RESEARCH DATA**

- USE EXISTING IMAGE DATABASES TO TRAIN NN

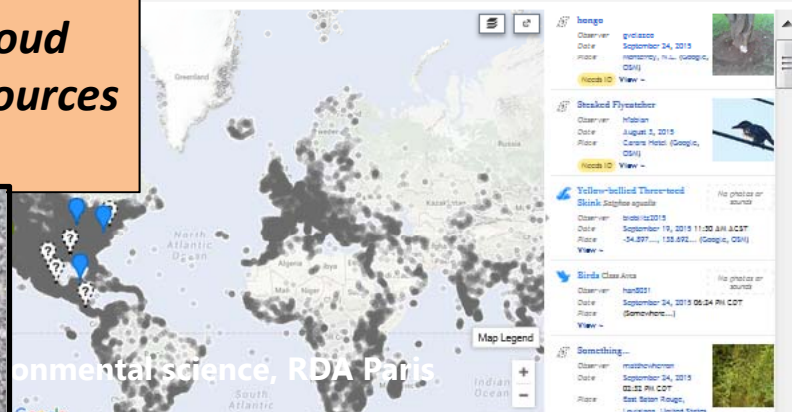
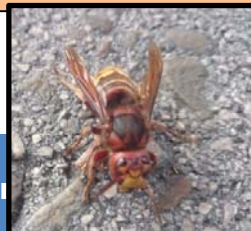
**STORAGE+COMPUTING
NEED GPUs IN TRAINING**

– **PRESERVE NEW OPEN RESEARCH DATA & ANALYSIS**

- UPLOADED IMAGES (SIGNIFICANT STORAGE)
- IDEALLY, OBSERVATION IS FUTURE INPUT TO GBIF



**GBIF.es services/storage already running in FedCloud
Additional Storage requires “elastic” increase of resources
Connect/Integrate with GPU**



Monitoring & Modeling ALGAE BLOOM in a Water Reservoir

LIFE+ Project lead by a SME, collecting monitoring data (environmental station+ water quality and chloro-cyano profiler), and modeling hydro+bio

– **INTEGRATE EXISTING OPEN RESEARCH DATA**

- USE **METEO**, TERRAIN, BATHIMETRY, LAND USE
- **HYDROLOGICAL** INPUT

**STORAGE+COMPUTING
NEED HPC FOR DELFT-3D**

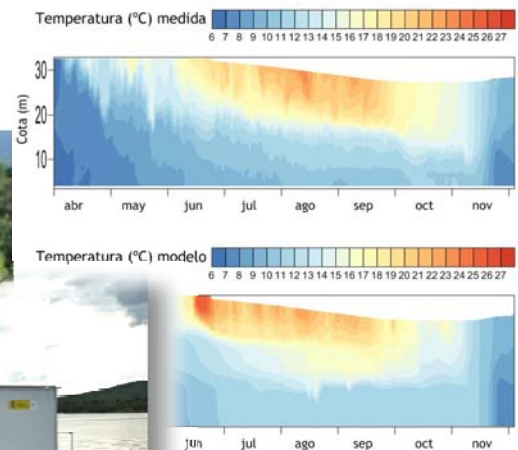
– **PRESERVE NEW OPEN RESEARCH DATA & ANALYSIS**

- REMOTELY COLLECTED DATA INTO REPLICATED DB
- COMPLEX MODEL OUTPUTS
- MULTIPARAMETRIC ANALYSIS

Model already running in FedCloud

Adapt Multiparametric scan

Preserve Thermoclines analysis (in R)



SOLUTIONS BEING EXPLORED

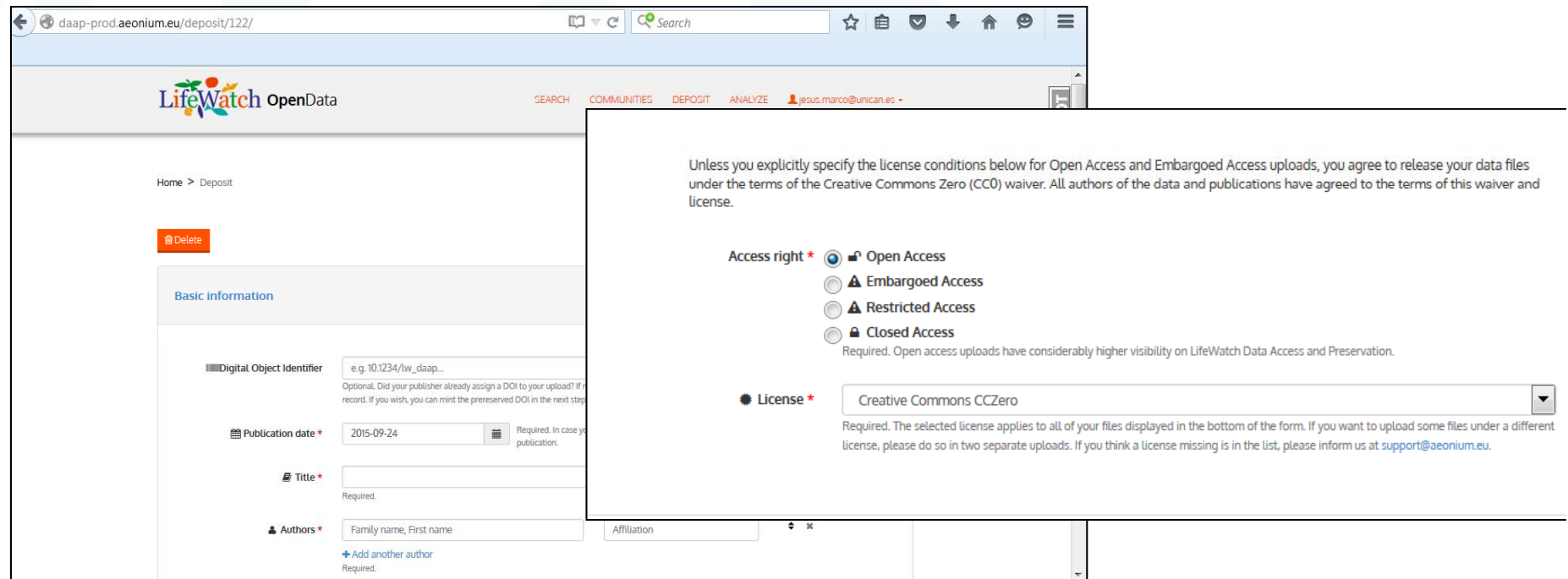
- Support external resources (data, tools): **VRE**
- Enable a “/lifewatch/home” for each researcher/each community, accessible with ID via a **preservation portal**
- Users will define the “openness” of their
 - DATA (private/**embargo**/open/published-DOI)
 - ANALYSIS (R/python, via github)
 - WORKFLOWS at SaaS level (R,python)
- Support it with a global (federated) distributed storage
 - OneData (Data Commons basic component)
- Integrated also with FedCloud computing resources
 - We will rely on INDIGO project developments to optimize!
- Enforce DMP (Data Management Plan)



*If it needs to be preserved => **DMP** & **OPEN** (after embargo)*

SOLUTIONS BEING EXPLORED

- Support external resources (data, tools): **VRE**
- Enable a “/lifewatch/home” for each researcher/each community, accessible with ID via a **portal**



The screenshot shows a web browser window with the URL `daap-prod.aeonium.eu/deposit/122/`. The page header includes the LifeWatch OpenData logo and navigation links for SEARCH, COMMUNITIES, DEPOSIT, and ANALYZE. A user profile for `jesus.marco@unican.es` is visible. The main content area shows a deposit form with a 'Delete' button and a 'Basic information' section. The form fields include: Digital Object Identifier (with a hint: 'e.g. 10.1234/lw_daap...'), Publication date (set to 2015-09-24), Title, and Authors (with a field for 'Family name, First name' and an 'Add another author' link). A modal dialog box is open, displaying a license selection interface. The dialog text states: 'Unless you explicitly specify the license conditions below for Open Access and Embargoed Access uploads, you agree to release your data files under the terms of the Creative Commons Zero (CC0) waiver. All authors of the data and publications have agreed to the terms of this waiver and license.' The 'Access right' section has radio buttons for 'Open Access' (selected), 'Embargoed Access', 'Restricted Access', and 'Closed Access'. The 'License' section has a dropdown menu set to 'Creative Commons CCZero'. A note below the license selection states: 'Required. The selected license applies to all of your files displayed in the bottom of the form. If you want to upload some files under a different license, please do so in two separate uploads. If you think a license missing is in the list, please inform us at support@aeonium.eu.'

- Enforce DMP (Data Management Plan)

*If it needs to be preserved => **DMP** & **OPEN** (after embargo)*

Thank you for your attention.

Questions?

Tourists in the ARTIC seeing how glaciers melt

Next EGI-LW CC meeting in
EGI Conference in BARI
11/11 11h



www.egi.eu



elroto.elpais@gmail.com

This work by Parties of the EGI-Engage Consortium is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

