Integration of WORSICA's thematic service in EOSC - challenges and achievements

LNEC: ***Ricardo Martins***, Alberto Azevedo, Anabela Oliveira
LIP: Samuel Bernardo, Jorge Gomes, Mário David, João Pina
DANS-KNAW: Slava Tykhonov
IFCA: Pablo Orviz

Summary

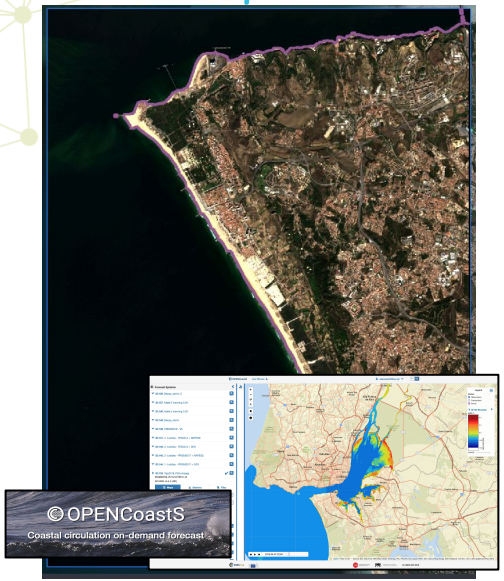
- What is WORSICA?
- Challenges
- Service Architecture (Past/Present)
- Technical Description
- Conclusions
  - Achievements
  - Future work

What is WORSICA?

- **WORSICA (Water mOnitoRing SentInel Cloud plAtform)**
  - a web service that aims at integrating remote sensing and in-situ data for the determination of water presence in coastal and inland areas, applicable to a range of purposes from the determination of flooded areas (from rainfall, storms, hurricanes or tsunamis) to the detection of large water leaks in major water irrigation networks.

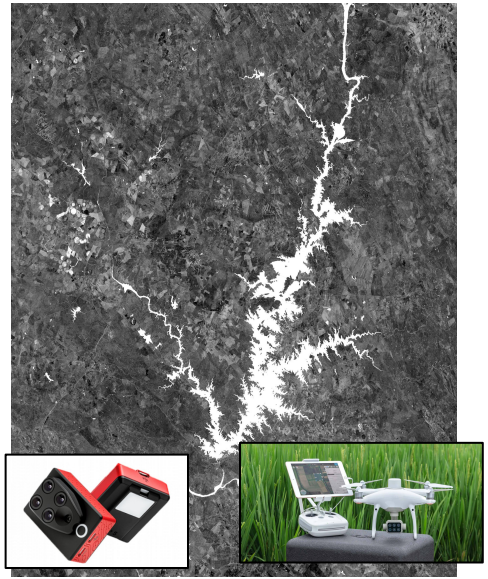
- **WORSICA will reuse and connect some work from other projects/services from LNEC:**
  - OPENCoastS (https://opencoasts.ncg.ingrid.pt)
  - WADI (https://www.waditech.eu)
  - Mosaic.pt (http://mosaic.lnec.pt)

WORSICA
Main products



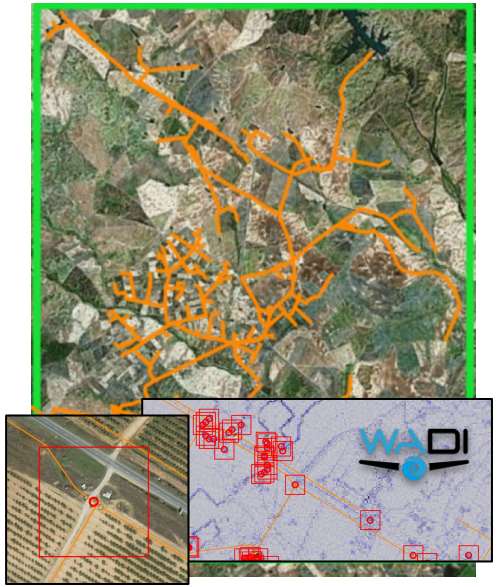
Coastline detection

Use of remote sensing (Sentinel-2, Pleiades and multispectral drone imagery) for the detection of water-land interface and possible calculation of the Digital Elevation Model for each line using the EOSC-hub OPENCoastS service.



Water bodies detection

Determination of water indexes to detect water bodies in inland areas (lagoons, reservoirs, etc.), using satellite and drone-based imagery.

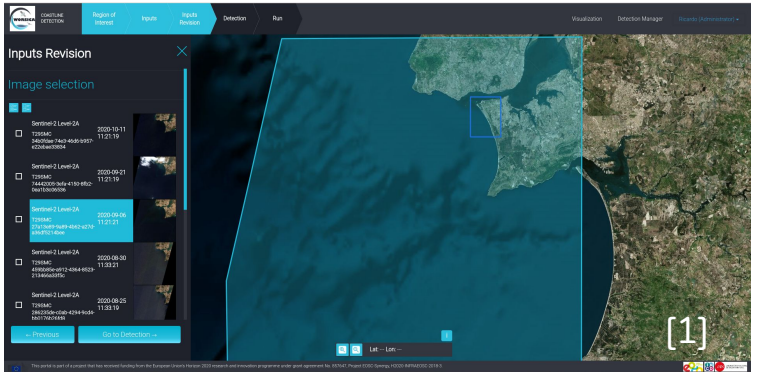


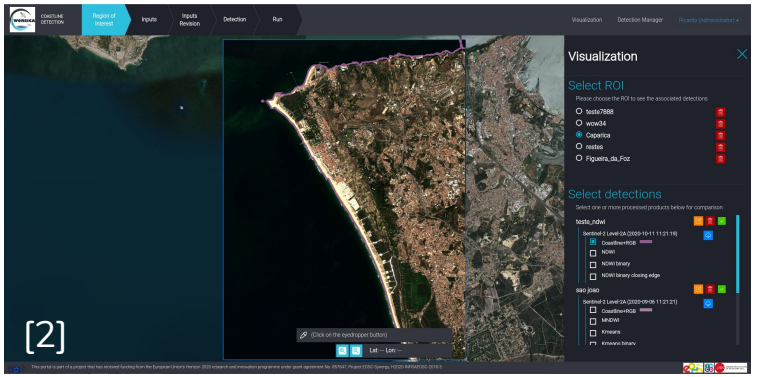
Water leak detection

Take advantage of the work developed in H2020-WADI project (with “low resolution” images from sentinel-2) and try to improve it using Pleiades and drone-based imagery.

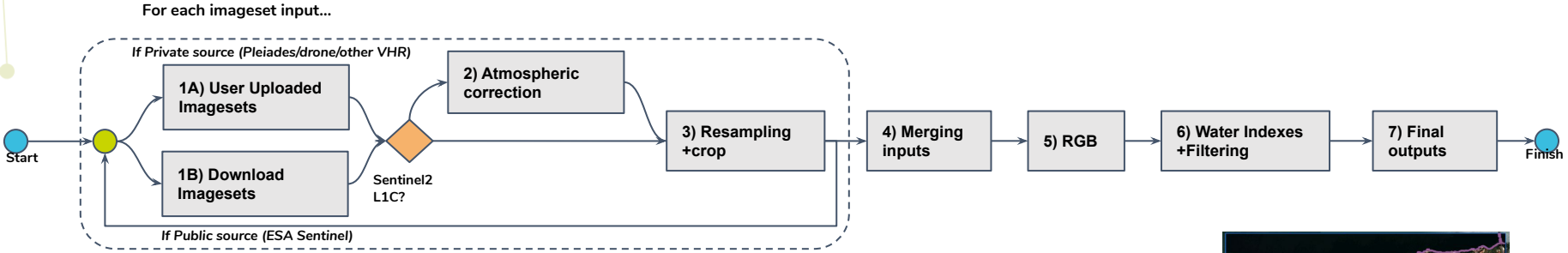
WORSICA
User Community and Usage model

- **Users**: researchers in coastal engineering as well as water irrigation networks management.
- **Access:** will be done via a single portal.
- **Usage**:
  - Configure and run the workflow, by choosing the Region of Interest (ROI) and the imagesets [1]. The service will download the imagesets, process them and generate the final products (water maps) [2].
  - The user will also be able to upload their own data from drone surveys or other private satellite images (e.g. Pleiades) to be processed.

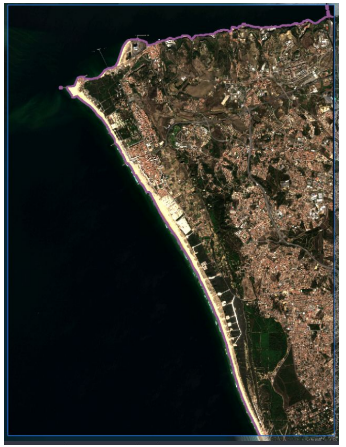




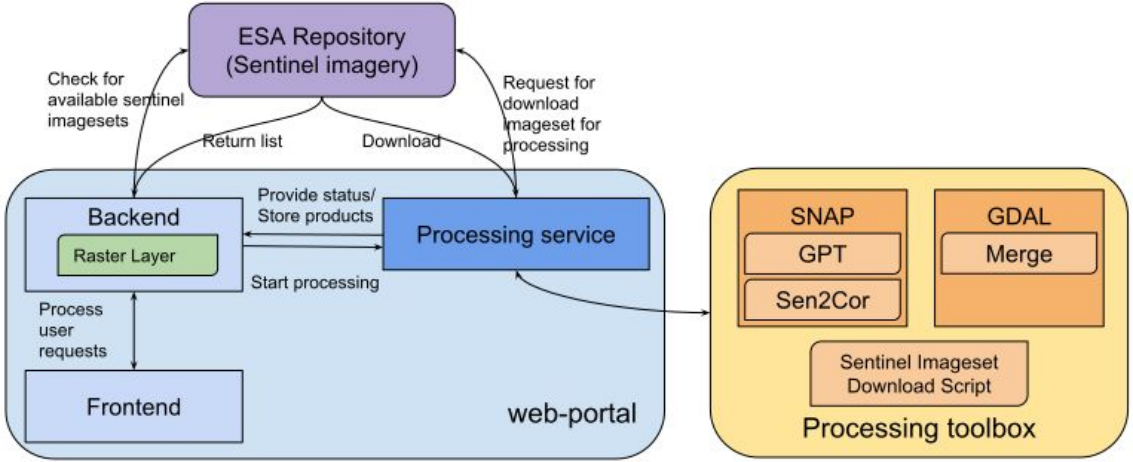
WORSICA
Processing Workflow



- **Workflow**
  - This processing workflow is done sequentially.
  - For bigger ROIs, more inputs, more time and more resources are needed to process.
  - Applying this workflow to a big processing job (e.g water leak detection), that require processing **hundreds** of imagesets, takes a lot of time to complete (days).
  - For a ROI, for the same day, it can have more than one imageset available for download (different orbits), thus the need for merging first inputs.



WORSICA
Architecture (State **before** EOSC-Synergy)



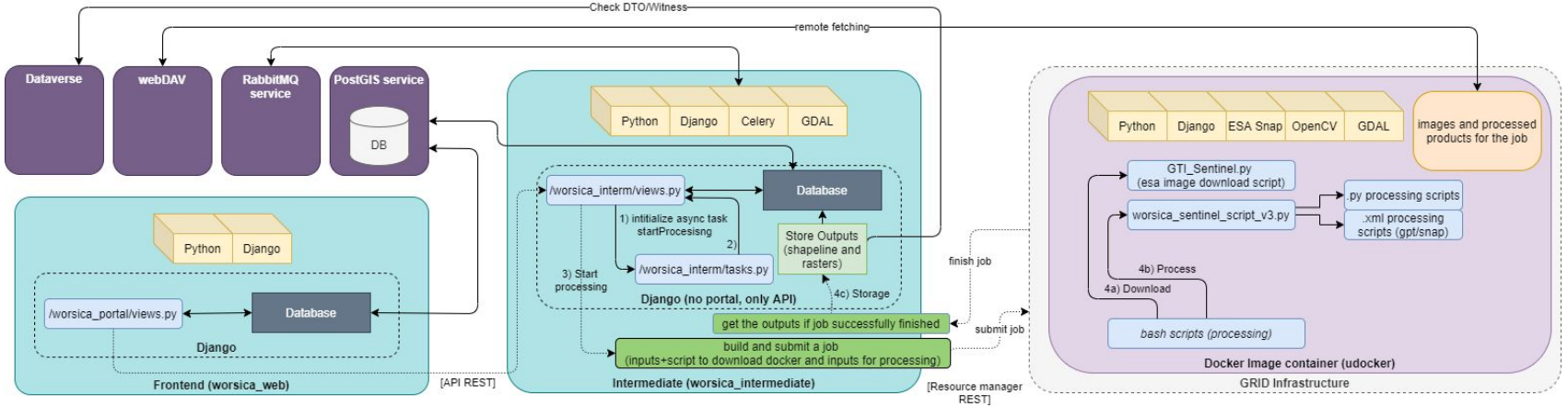
Architecture

- Everything on the Cloud.
- No dockers.
- Frontend provides the portal to the user and communicates with it's backend
- Backend receives and responds to user requests, checks available imagesets and stores final products to the DB.
- A 'service' will download the imageset, then start processing it using the available tools from the toolbox.

WORSICA
Challenges on EOSC-Synergy

- **Processing**:
  - Network: Download speeds and number of simultaneous downloads of satellite data from operational providers (e.g. ESA or Pleiades);
  - Storage: of the intermediate and final products;
    - Each input is a imageset with ~1,1GB of size,
    - Intermediate and final outputs will be produced during the processing
  - Computation resources: where the GPU and RAM are highly recommended to speedup the image processing and to prevent bottlenecks on using the service during processing.
    - Processing a imageset requires at least 8GB of RAM
    - More will be required if imageset requires atmospheric correction.
    - Processing workflow needs to be parallelized.
  - Assure Resilience and Speedup.
- **Thematic service**:
  - Assure Scalability, Redundancy, Portability of the service
  - Assure (inter)operability with other thematic services
- **Other**:
  - Assure Data FAIRness of the user generated products on this thematic service

WORSICA
Architecture (**Actual** state during EOSC-Synergy)



Architecture:

- Frontend component (web portal on Cloud)
- Intermediate component (simple task orchestrator w/ Celery on Cloud)
- Processing component (an instance that will be sent to GRID infrastructure).
- Dataverse/Storage/Broker/Database are isolated components

Advantages

- Provides redundancy, scalability and flexibility
- Good weight distribution on each component
- Components are using dockers, easier for portability and installation
- Processing jobs submissions are sent by the WORSICA Intermediate service to a GRID infrastructure

Technical Description: Services

	Service Used		Provider	
	Before	**Planned**	**Before**	**Planned**
AAI	Local	EGI Check-In	INCD	EGI Federation
Workload Mng.	Local batch system	ArcCE, SLURM	INCD	EOSC-Synergy
Resource Mng.	Manual	IM (TOSCA)	INCD	EOSC-Synergy
Data Storage	Local	Nextcloud	INCD	EOSC-Synergy
Monitoring	-	ARGO	-	EGI Service Monitoring
Other: Hydrodynamic water forecasts	-	OPENCoastS	-	EOSC-hub marketplace
Other: Dataverse	-	Dataverse	-	EOSC-Synergy
Computing Resources	Local	FedCloud and EGI HTC	INCD	EOSC-Synergy
Storage Resources	Local	FedCloud and EGI Online storage	INCD	EOSC-Synergy

Technical Description: Planned Services

- **Authentication**:
  - **EGI Check-In:** federated authentication is required on WORSICA to have access to the available EOSC services and resources.
- **Workload Managers**:
  - **ArcCE** with **SLURM**: This allows efficient management of the available GRID resources for HPC in order to speed up the processing jobs.
- **Data Manager**:
  - **Nextcloud**: to store processed job submission data input/outputs.
  - **Dataverse**: to register processed job submission metadata information for data FAIRsFAIR compliance. (more on the Service QA and Dataverse presentation)
- **Ansible and IM**:
  - **IM**: for deployment of the infrastructure required for job processing, repositories and microservices.
  - **SLURM** and **Kubernetes** clusters are deployed using **TOSCA** template over IaaS service and the remaining services will be installed from Docker images. Configurations for SLURM and Kubernetes are set up by ansible playbooks.
- **CI/CD for the automatization of the service integration in EOSC infrastructure:**
  - **Jenkins** pipelines and **unitary/functional** tests were also developed to be compliant with the SQAaaS methodologies developed in EOSC-Synergy (more on the Service QA and Dataverse presentation)

Conclusions

WORSICA Achievements on EOSC-Synergy

- **Resilience+Speedup**: a better GRID infrastructure to process Sentinel and other VHR imagery, using robust and tested software.
- **Scalability+Redundancy**: possibility to adjust the resources according to the usage. More users, more resources for the computation
- **Portability+Deployment**: possibility to port this thematic service to any other infrastructure depending of the computational needs
- **Federated access+Support:** the need for a federated access is a requirement on WORSICA to have access and support to the resources and other EOSC services.
- **Interoperability:** this thematic service can be a connection for other thematic services, and can connect with other thematic services to provide additional products.
- **Data FAIRness**: provide data FAIRness to the WORSICA user generated products.

Future work

- Improve existing water processing algorithms and/or implement new ones
- Improve user interaction with the portal.
- Improve interoperability with other thematic services
- Continuous improvement/update of the IT services implemented in the WORSICA thematic service

Thank you.

Ricardo Martins

rjmartins@lnec.pt

http://worsica.lnec.pt/

Service QA and Dataverse

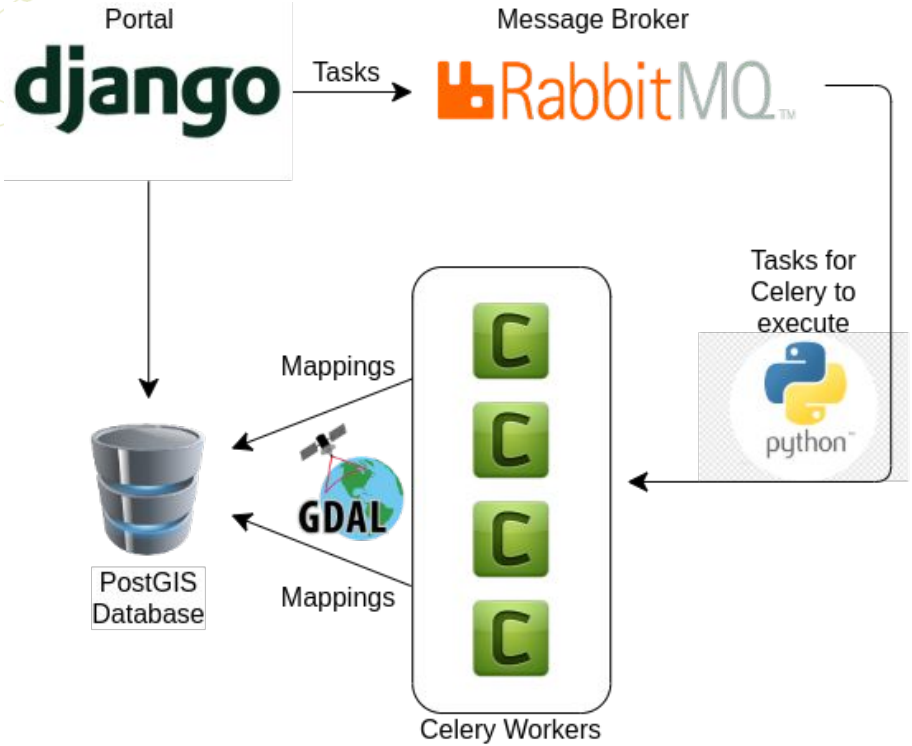
Speaker:
Vyacheslav Tykhonov (DANS-KNAW) on behalf of EOSC-Synergy

Collaborators:
Jorge Gomes (LIP), João Pina (LIP), Mário David (LIP), Ricardo Martins (LNEC), Alberto Azevedo (LNEC), Samuel Bernardo (LIP), Pablo Orviz (CSIC), Isabel Campos (CSIC), Germán Moltó (UPV) and Miguel Caballer (UPV)

Introducing Dataverse data repository

- Open source project developed by IQSS of Harvard University
- Great product with very long history (from 2006) and dynamic and experienced development team
- Clear vision and understanding of research communities requirements, public roadmap
- Well developed architecture with rich APIs allows to build application layers around Dataverse
- Strong community behind of Dataverse is helping to improve the basic functionality and develop it further. DANS-KNAW is leading SSHOC task to deliver production ready Dataverse for all partners

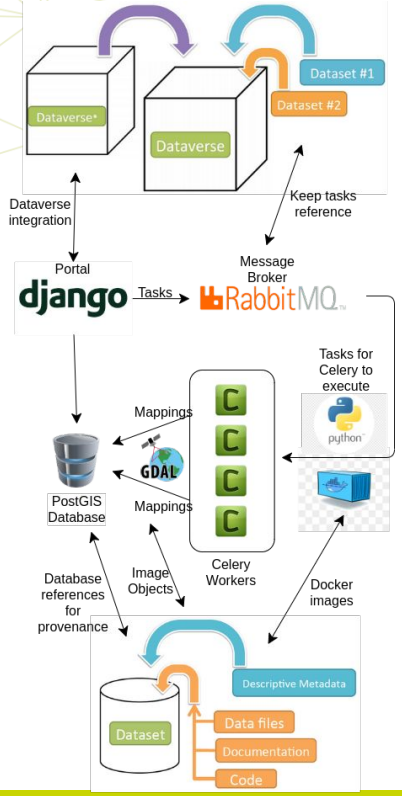
WORSICA and Dataverse Repository (1)



Initial service architecture:

- Data missing global unique identifier
- Data stored in multiple places internal to the services and not accessible
- Inexistent metadata detailed provenance association
- Data access not following controlled vocabularies that apply FAIR principles

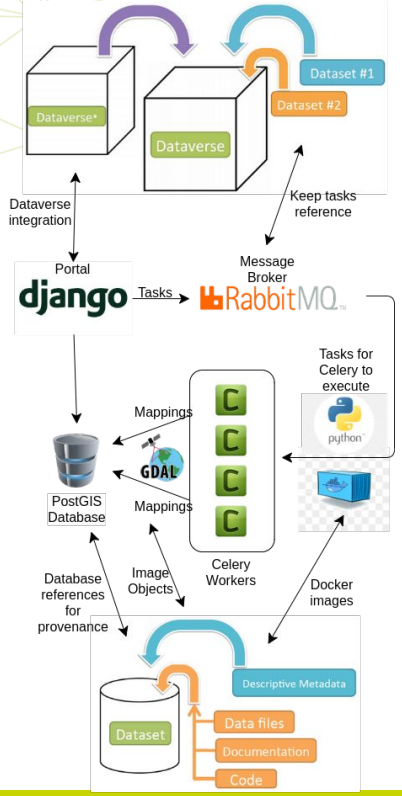
Dataverse Repository (2)



FAIR service architecture:

- Dataverse provides the repository solution that complies with the FAIR principles
- Define a dataverse and associate a persistent identifier namespace
- Associate metadata with the provided and produced data
- Use Data Commons to allow data sharing between all teams and projects
- Metadata is by default associated with CC0 Creative Commons license and publicly accessible

Dataverse Repository (3)



Integrate code with Dataverse REST API:

- Very useful to implement in any language only being dependent with the provided interface without any library requirements
- Easy to maintain WORSICA code in parallel with Dataverse service updates
- Current Dataverse REST API is very complete and allows to run all necessary operations
- Share sensitive data with confidence using DataTags System, that will allow to use a set of security features and access requirements for file handling

Interoperability in WORSICA/Dataverse

- Most of variables in WORSICA datasets can be linked in Dataverse to the appropriate ontologies to increase interoperability and data FAIRness
- Variable names can be included in datasets metadata in the native language (Portuguese) and get URI identifiers for those entities in controlled vocabularies
- Standardized metadata fields available in Linked Open Data Cloud through standard machine-to-machine interfaces available in Dataverse

WORSICA
Dataverse Metadata (1)

Note:
- **Remember, these variables may change during WORSICA development.**
- **If you see (Multiple), it means the field allows to add more than one object on it.**
- **This is for Dataverse 4.19.**
- **For further information, we recommend to check Dataverse documentation.**

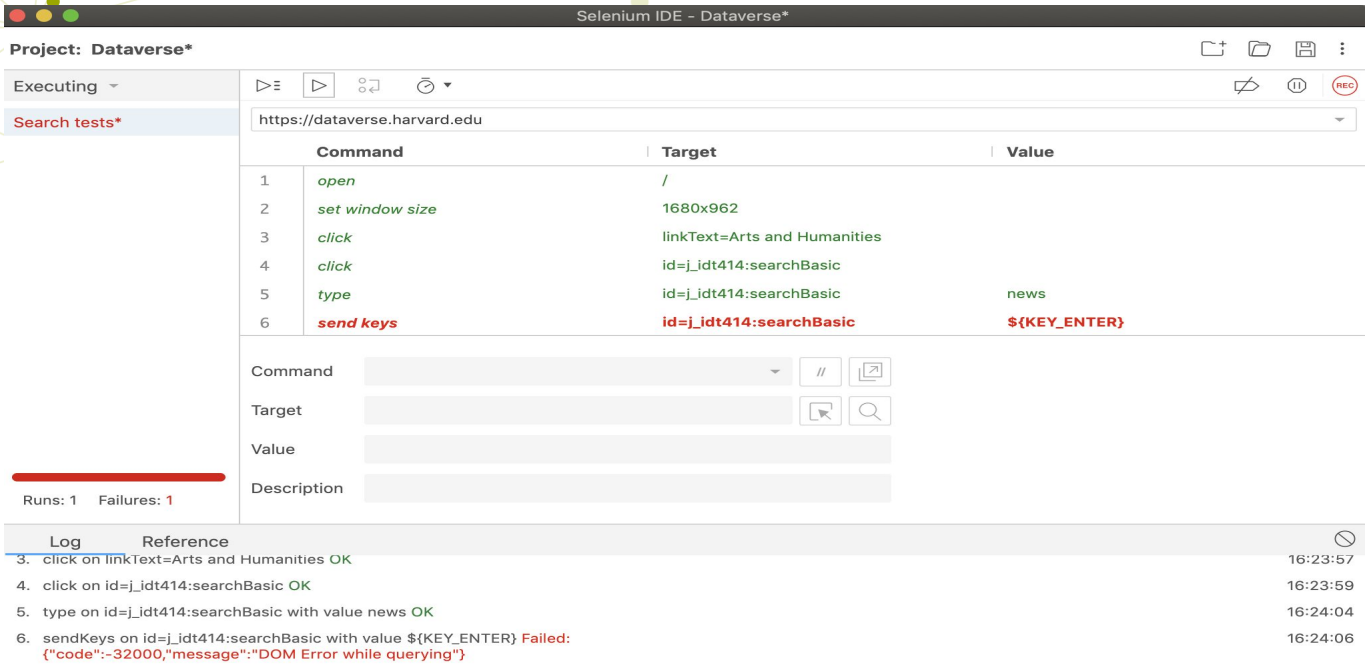
Scope	Type Name		Description	Example	Vocabulary URI
Citation	title	-	A title for this dataverse	“Figueira da Foz (ROI110 Simulation230)”	-
	author **(Multiple)**	authorName authorAffiliation	Author or a list of authors who ran the WORSICA processing to generate this dataverse	“Ricardo Martins” “LNEC”	-
	datasetContact **(Multiple)**	datasetContactName datasetContactEmail	Author or a list of authors who own this dataset	“Ricardo Martins” “rjmartins(at)lnec.pt”	-
	dsDescription **(Multiple)**	dsDescriptionValue	A detailed description or various descriptions of the dataset	"Simulation230 in ROI110 (38,-8,37,-7) with minimum bath depth threshold of -10m and maximum topo depth threshold of 10m"	-
	subject **(Multiple)**	-	An array of thematic subject(s) that identify the dataverse. Take note these subjects must match with the ones provided from Dataverse.	["Computer and Information Science", "Earth and Environmental Sciences"]	https://www.wikidata.org/wiki/Q21198
	dataSource **(Multiple)**	-	List the data sources	(list the name of the processed image sets...)	-
	otherReferences **(Multiple)**	-	Other references to mention on this dataset	(list of URLs of the processed WORSICA products from the storage…)	-
	software **(Multiple)**	softwareName softwareVersion	List of software used for the processing	“GDAL” “3.0.4”	https://www.wikidata.org/wiki/Q676202
	displayName	-	A name to this dataverse scope	“Figueira da Foz (ROI110 Simulation230) Citation Metadata”	-

WORSICA
Dataverse Metadata (2)

Note:
- **Remember, these variables may change during WORSICA development.**
- **If you see (Multiple), it means the field allows to add more than one object on it.**
- **This is for Dataverse 4.19.**
- **For further information, we recommend to check Dataverse documentation.**

Scope	Type Name			Description	Example	Vocabulary URI
Geospatial	fields	geograhicCoverage **(Multiple)**	country state city otherGeographicCoverage	Details representing the processed ROI	Portugal Lisbon Lisbon Av. do Brasil	https://www.wikidata.org/wiki/Q6256 https://www.geonames.org/countries/PT/portugal.html
		geographicUnit **(Multiple)**	-	A list of Default units	[“m”,”m”]	-
		geographicBoundingBox **(Multiple)**	westLongitude eastLongitude northLongitude southLongitude	A bounding box or a list of bounding boxes that represent the processed ROI	-9 -8 38 37	https://www.wikidata.org/wiki/Property:P625
	displayName	-	-	A name to this dataverse scope	“Figueira da Foz (ROI110 Simulation230) Geospatial Metadata”	-

SQA process with Selenium tests for Dataverse



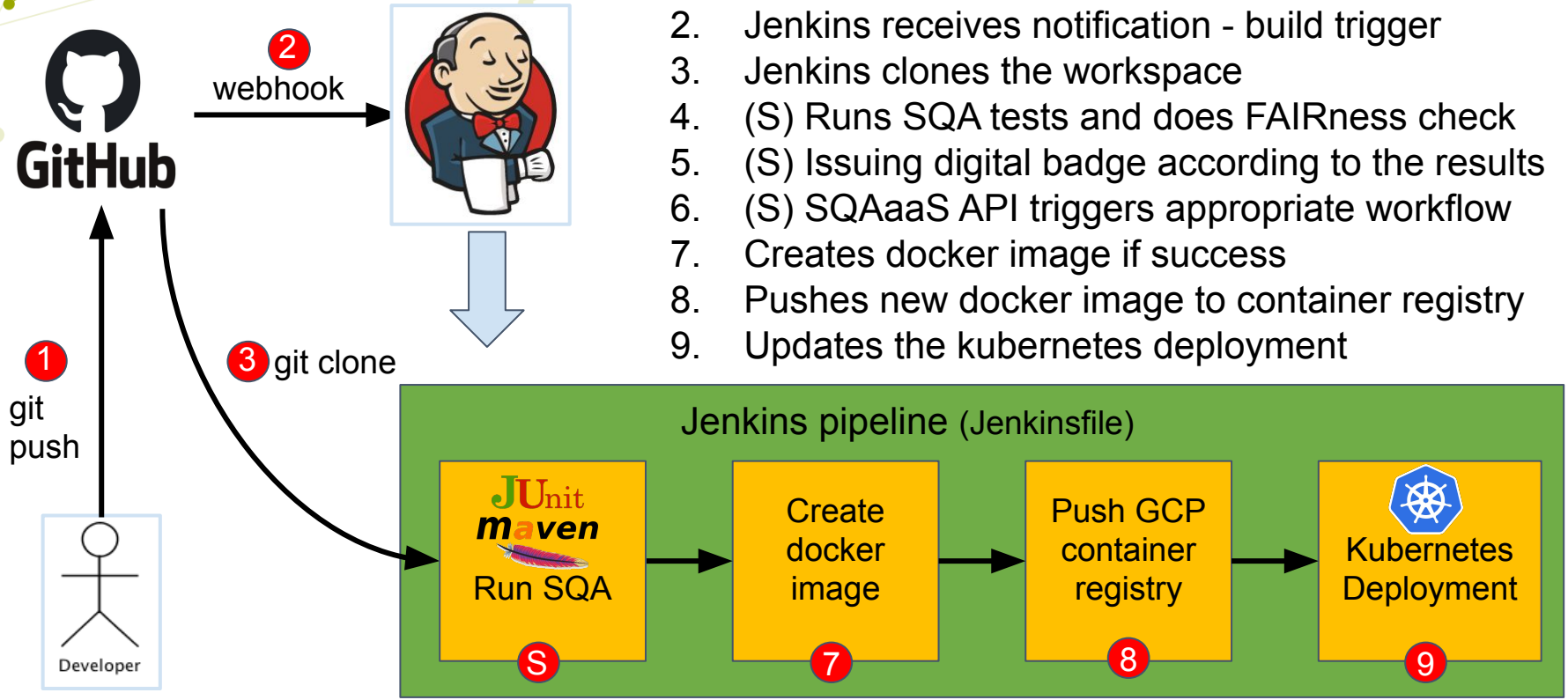
Selenium IDE allows to create and replay all UI tests in your browser

Shared tests can be reused by community to increase reproducibility

SQA for the service maturity = unit tests + integration tests

CI/CD pipeline with SQAaaS (S)

1. Developer pushes code to GitHub
2. Jenkins receives notification - build trigger
3. Jenkins clones the workspace
4. (S) Runs SQA tests and does FAIRness check
5. (S) Issuing digital badge according to the results
6. (S) SQAaaS API triggers appropriate workflow
7. Creates docker image if success
8. Pushes new docker image to container registry
9. Updates the kubernetes deployment



Final remarks

Dataverse pros:

- provides a FAIR repository for thematic services and has a rich REST interface
- open source software with Apache License v2.0
- allows to manage public and private data
- data commons sharing along teams / projects

Dataverse cons:

- software integration for data management using Dataverse couldn’t be as quick as expected because of required learning curve
- an account and associated namespace must be acquired for a fee from a DOI or HDL provider for persistent identifiers be citable

Thank you

For further information:

communications@eosc-synergy.eu

www.eosc-synergy.eu