Science Reproducibility and Reusability with FutureGateway and a Zenodo-like repository: the PALMS experiment

R. Bruno[1*], R. Barbera[1,2], M. Fargetta[1], R. Rotondo[1], A. Anagnostou[3], S. J. Taylor[3]

1 Italian National Institute of Nuclear Physics, Division of Catania, Italy

2 Department of Physics and Astronomy “E. Majorana” of the University of Catania, Italy

3 Modelling & Simulation Group, Department of Computer Science, Brunel University London, UK

* riccardo.bruno@ct.infn.it

Driving considerations

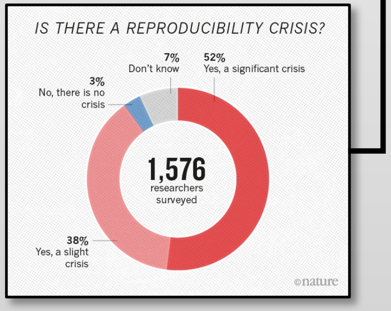
*"Open Science refers to a scientific culture that is characterized by its openness. Scientists **share results** almost immediately and with a very wide audience"* [1]

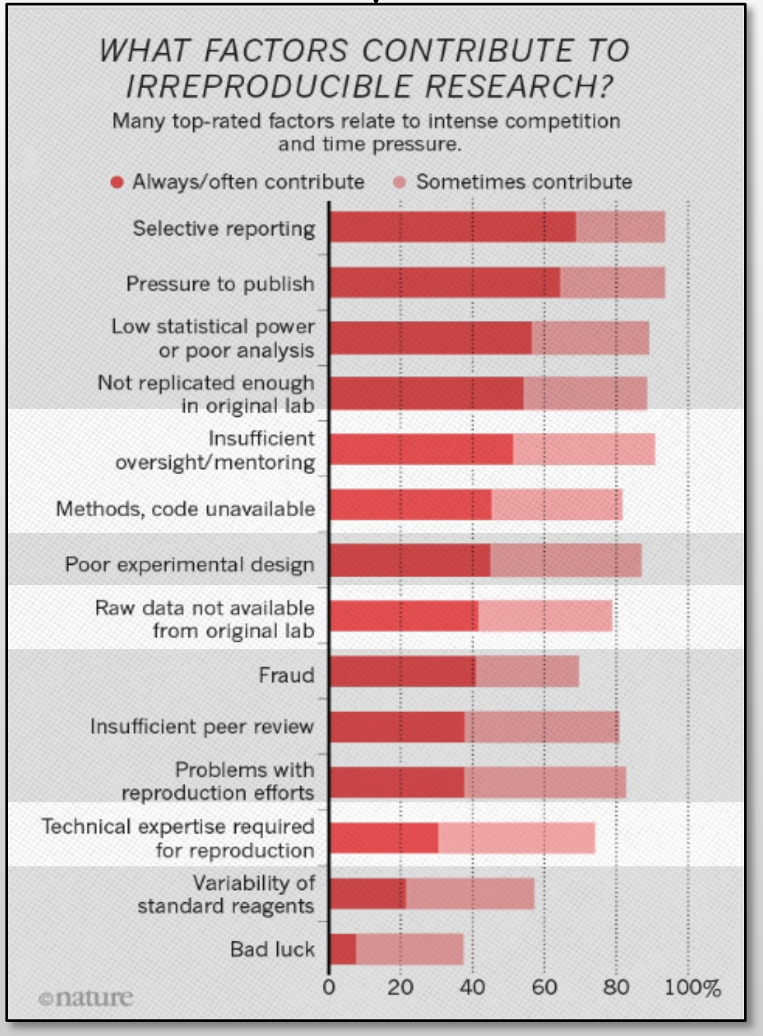
*"Open science is a means and not an end in itself and it is much more than just open access to publications or data; it includes many aspects and stages of research processes thus enabling full **reproducibility and re-usability** of scientific results."* [2]

nature
International journal of science

Reality check on reproducibility (Survey) [3]

- Insufficient oversight/mentoring
- Methods, code unavailable
- Raw data not available from original lab
- Technical expertise requires for reproducibility


IS THERE A REPRODUCIBILITY CRISIS?


WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.

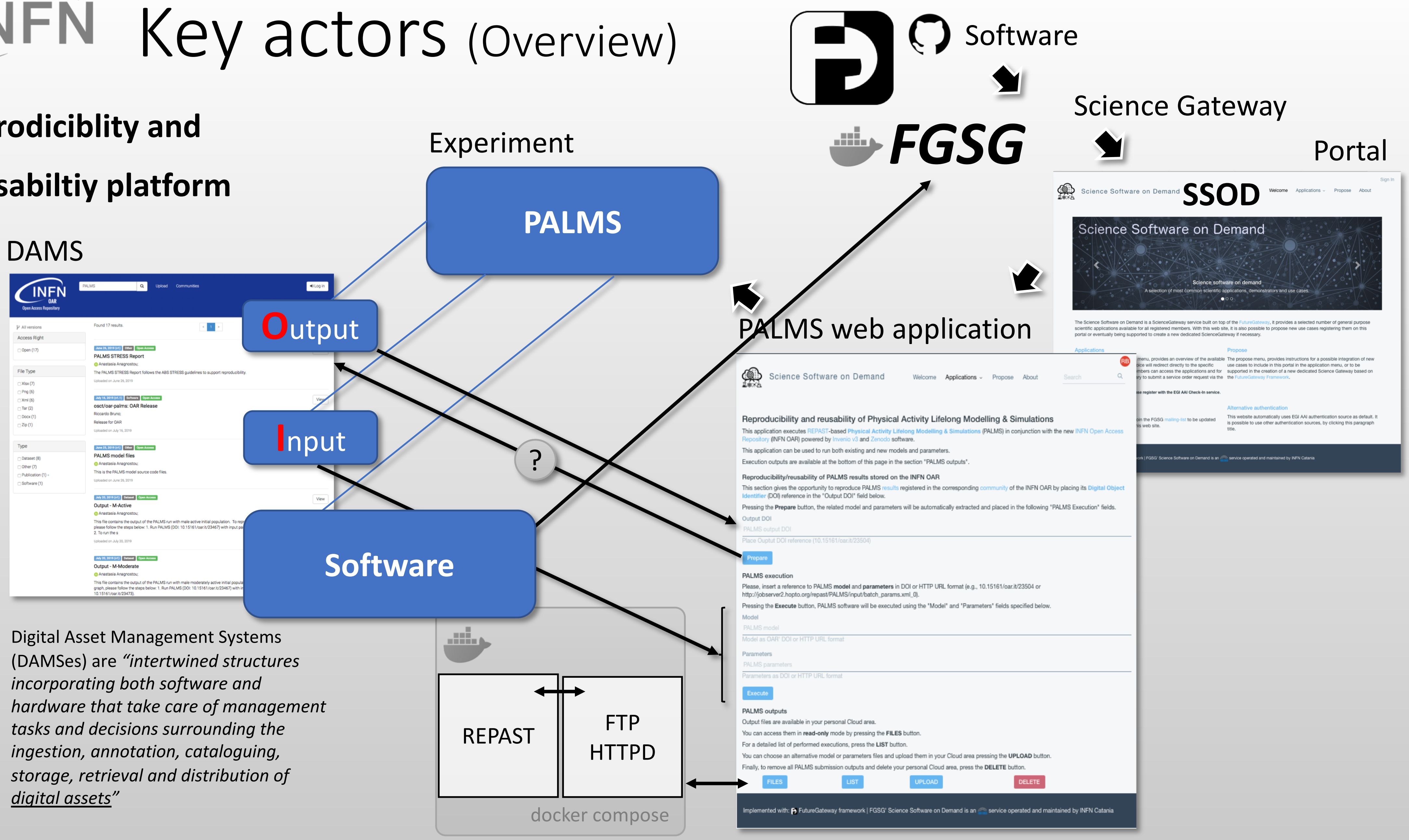
[1] Opening Science: The Evolving Guide on How the Internet is Changing Research (DOI: 10.1007/978-3-319-00026-8)

[2] Making Open Science a Reality: http://dx.doi.org/10.1787/5jrs2f963zs1-en

[3] Reality check on reproducibility: https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Key actors (Overview)

Reprodiciblity and
Reusabiltiy platform



PALMS | Upload | Communities | Log in

INFN OAR Open Access Repository

All versions

Access Right
- □ Open (17)

File Type
- □ Xlsx (7)
- □ Png (6)
- □ Xml (6)
- □ Tar (2)
- □ Docx (1)
- □ Zip (1)

Type
- □ Dataset (8)
- □ Other (7)
- □ Publication (1)
- □ Software (1)

Found 17 results.

June 26, 2019 (v1) | Other | Open Access
PALMS STRESS Report
Anastasia Anagnostou;
The PALMS STRESS Report follows the ABS STRESS guidelines to support reproducibility.
Uploaded on June 26, 2019

July 16, 2019 (v1.1) | Software | Open Access
osct/oar-palms: OAR Release
Riccardo Bruno;
Release for OAR
Uploaded on July 16, 2019

June 25, 2019 (v1) | Other | Open Access
PALMS model files
Anastasia Anagnostou;
This is the PALMS model source code files.
Uploaded on June 26, 2019

July 20, 2019 (v1) | Dataset | Open Access
Output - M-Active
Anastasia Anagnostou;
This file contains the output of the PALMS run with male active initial population. To repr please follow the steps below: 1. Run PALMS (DOI: 10.15161/oar.it/23467) with input pa 2. To run the s
Uploaded on July 20, 2019

July 20, 2019 (v1) | Dataset | Open Access
Output - M-Moderate
Anastasia Anagnostou;
This file contains the output of the PALMS run with male moderately active initial popula graph, please follow the steps below: 1. Run PALMS (DOI: 10.15161/oar.it/23467) with in 10.15161/oar.it/23473).

Science Software on Demand | Welcome | Applications | Propose | About | Sign In

Science Software on Demand

Science software on demand
A selection of most common scientific applications, demonstrators and use cases.

The Science Software on Demand is a ScienceGateway service built on top of the FutureGateway, it provides a selected number of general purpose scientific applications available for all registered members. With this web site, it is also possible to propose new use cases registering them on this portal or eventually being supported to create a new dedicated ScienceGateway if necessary.

Applications
menu, provides an overview of the available oice will redirect directly to the specific mbers can access the applications and for ry to submit a service order request via the
se register with the EGI AAI Check-In service.
oin the FGSG mailing-list to be updated is web site.

Propose
The propose menu, provides instructions for a possible integration of new use cases to include in this portal in the application menu, or to be supported in the creation of a new dedicated Science Gateway based on the FutureGateway Framework.

Alternative authentication
This website automatically uses EGI AAI authentication source as default. It is possible to use other authentication sources, by clicking this paragraph title.

ork | FGSG' Science Software on Demand is an service operated and maintained by INFN Catania

Science Software on Demand | Welcome | Applications | Propose | About | Search

Reproducibility and reusability of Physical Activity Lifelong Modelling & Simulations

This application executes REPAST-based **Physical Activity Lifelong Modelling & Simulations** (PALMS) in conjunction with the new INFN Open Access Repository (INFN OAR) powered by Invenio v3 and Zenodo software.

This application can be used to run both existing and new models and parameters.

Execution outputs are available at the bottom of this page in the section "PALMS outputs".

Reproducibility/reusability of PALMS results stored on the INFN OAR

This section gives the opportunity to reproduce PALMS results registered in the corresponding community of the INFN OAR by placing its **Digital Object Identifier** (DOI) reference in the "Output DOI" field below.

Pressing the **Prepare** button, the related model and parameters will be automatically extracted and placed in the following "PALMS Execution" fields.

Output DOI
PALMS output DOI
Place Ouptut DOI reference (10.15161/oar.it/23504)

Prepare

PALMS execution

Please, insert a reference to PALMS **model** and **parameters** in DOI or HTTP URL format (e.g., 10.15161/oar.it/23504 or http://jobserver2.hopto.org/repast/PALMS/input/batch_params.xml_0).

Pressing the **Execute** button, PALMS software will be executed using the "Model" and "Parameters" fields specified below.

Model
PALMS model
Model as OAR' DOI or HTTP URL format

Parameters
PALMS parameters
Parameters as DOI or HTTP URL format

Execute

PALMS outputs

Output files are available in your personal Cloud area.

You can access them in **read-only** mode by pressing the **FILES** button.

For a detailed list of performed executions, press the **LIST** button.

You can choose an alternative model or parameters files and upload them in your Cloud area pressing the **UPLOAD** button.

Finally, to remove all PALMS submission outputs and delete your personal Cloud area, press the **DELETE** button.

FILES | LIST | UPLOAD | DELETE

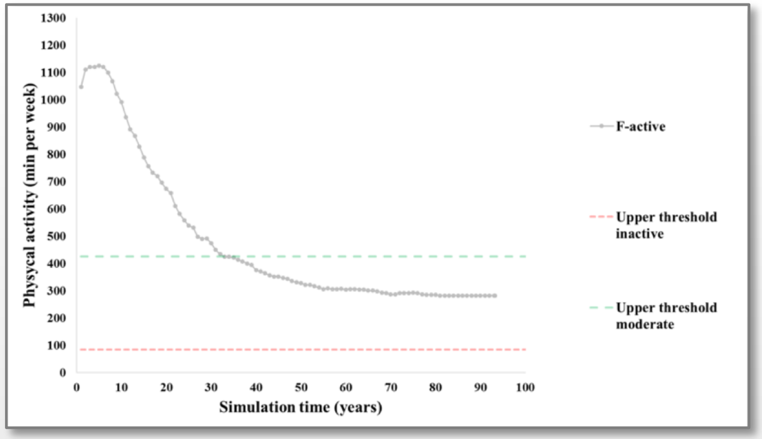
Implemented with: FutureGateway framework | FGSG' Science Software on Demand is an service operated and maintained by INFN Catania

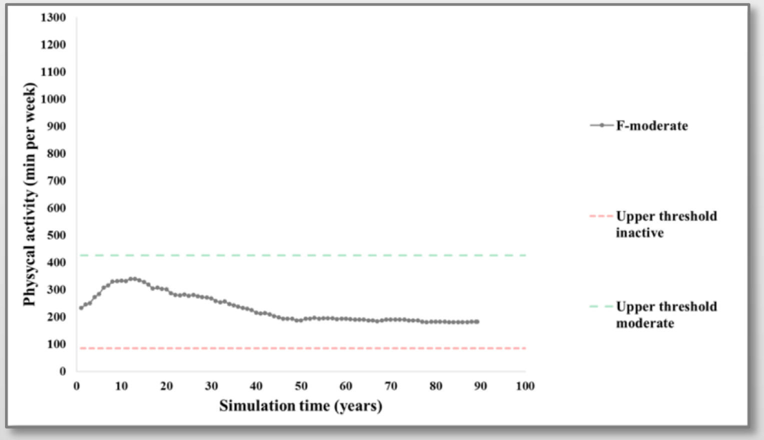
Digital Asset Management Systems (DAMSes) are *"intertwined structures incorporating both software and hardware that take care of management tasks and decisions surrounding the ingestion, annotation, cataloguing, storage, retrieval and distribution of digital assets"*

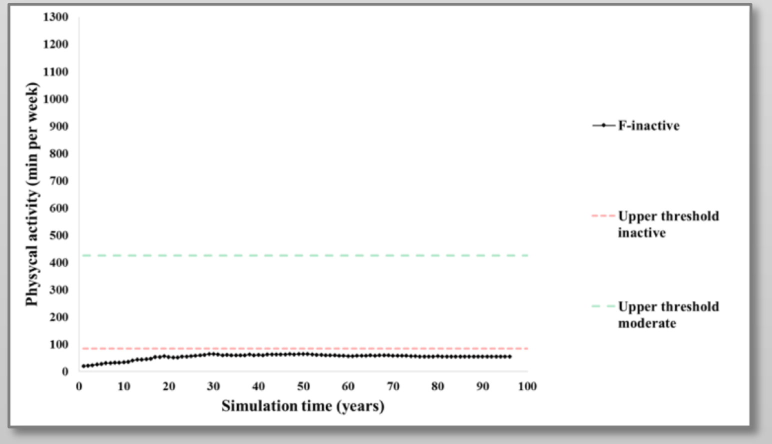
PALMS

- **P**hysical **A**ctivity **L**ifelong **M**odelling & **S**imulations
  - Is an agent based micro-simulation that predicts the lifelong physical activity behaviour of a population taking into account individual characteristics and their effect on physical activity over time
  - Produces individual and aggregated quantitative outputs for quality of life and health conditions related costs

- The software
  - Uses REPAST [4] an open source agent-based modeling and simulation platform
  - A specific dockerhub image exists for PALMS executions (osabuoun/repast) [5]
  - Two inputs necessary: model file (REPAST) and a parameters' file







[4] https://repast.github.io
[5] https://hub.docker.com/r/osabuoun/repast

FutureGateway Framework

futuregatewayframework.github.io

INFN software project aiming to build secure and reliable Science Gateways [6]

Three core components:

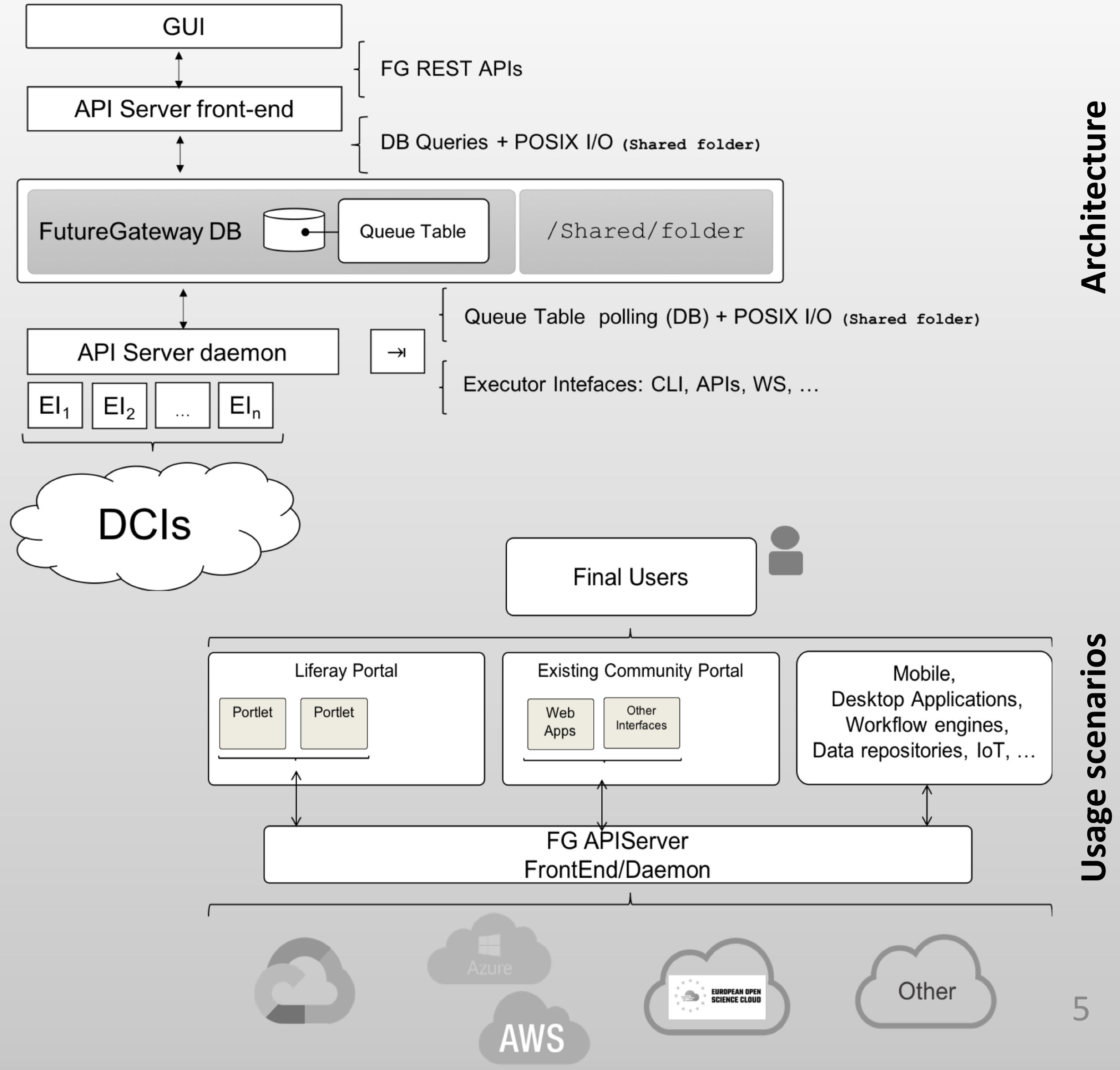
- Database, APIServer front-end, APIServer daemon + Executor Interfaces

The framework:

- Core components are enriched with a suite of tools, APIs and installation + maintenance scripts
- Open Source code available on GitHub

Targets:

- Desktop and Mobile applications, Workflow Engines, IoT and **Open Science**


GUI
FG REST APIs
API Server front-end
DB Queries + POSIX I/O (Shared folder)
FutureGateway DB
Queue Table
/Shared/folder
Queue Table polling (DB) + POSIX I/O (Shared folder)
API Server daemon
Executor Intefaces: CLI, APIs, WS, …
EI_1
EI_2
...
EI_n
DCIs
Architecture
Final Users
Liferay Portal
Portlet
Portlet
Existing Community Portal
Web Apps
Other Interfaces
Mobile, Desktop Applications, Workflow engines, Data repositories, IoT, …
FG APIServer FrontEnd/Daemon
Azure
AWS
EUROPEAN OPEN SCIENCE CLOUD
Other
Usage scenarios

[6] https://www.xsede.org/ecosystem/science-gateways

FG APIs

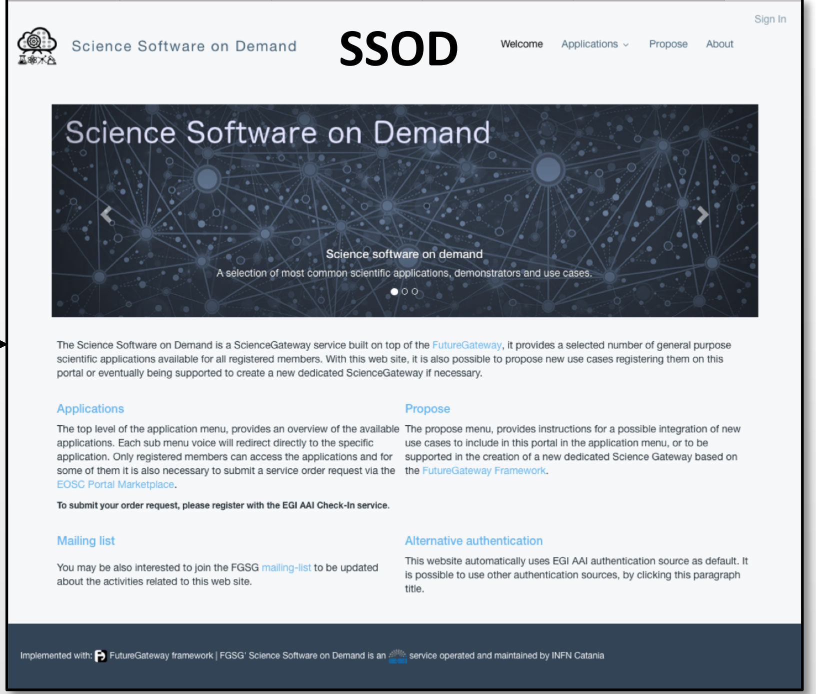
- FG uses a set of RESTful APIs to perform operations on Distributed Computing Infrastructures (DCI)
- APIs are splitted in three families:
  - **IAT**, Specifications [7]
    - Infrastructures – Specify how to target the DCI
    - Applications – Specify which kind of activity perform on the DCI
    - Tasks – Application instances, relative statuses and output retrieval
  - **UGR**, Documented
    - Users – FG manages its own set of API users
    - Groups – Membership can be grouped to assign different set of Roles
    - Roles – Are assigned to groups and specify the operations allowed to perform
  - **AAA**, Design
    - Authorisation Auditing and Accounting (not yet available, db level functionality)

[7] https://fgapis.docs.apiary.io/#

FGSG – FutureGateway based Science Gateway

Fully docker containerised environment built in the context of the EOSC-hub project [8], to provide a General purposes Science Gateway: the EGI [9] **Science Software on Demand** (SSOD) [10]

- The system allows to dynamically instantiate and destroy docker containers (it supports docker compose as well as docker swarm)
- FG core services + SSOD service
- SSOD service powered by an enterprise portal framework (Liferay)
  - One section dedicated to the **Reproducibility & Reusability platform**
  - The platform exploits the FutureGateway and the INFN Open Access Repository



SSOD

Science Software on Demand
Sign In
Welcome
Applications
Propose
About
Science Software on Demand
Science software on demand
A selection of most common scientific applications, demonstrators and use cases.

The Science Software on Demand is a ScienceGateway service built on top of the FutureGateway, it provides a selected number of general purpose scientific applications available for all registered members. With this web site, it is also possible to propose new use cases registering them on this portal or eventually being supported to create a new dedicated ScienceGateway if necessary.

Applications

The top level of the application menu, provides an overview of the available applications. Each sub menu voice will redirect directly to the specific application. Only registered members can access the applications and for some of them it is also necessary to submit a service order request via the EOSC Portal Marketplace.

To submit your order request, please register with the EGI AAI Check-In service.

Propose

The propose menu, provides instructions for a possible integration of new use cases to include in this portal in the application menu, or to be supported in the creation of a new dedicated Science Gateway based on the FutureGateway Framework.

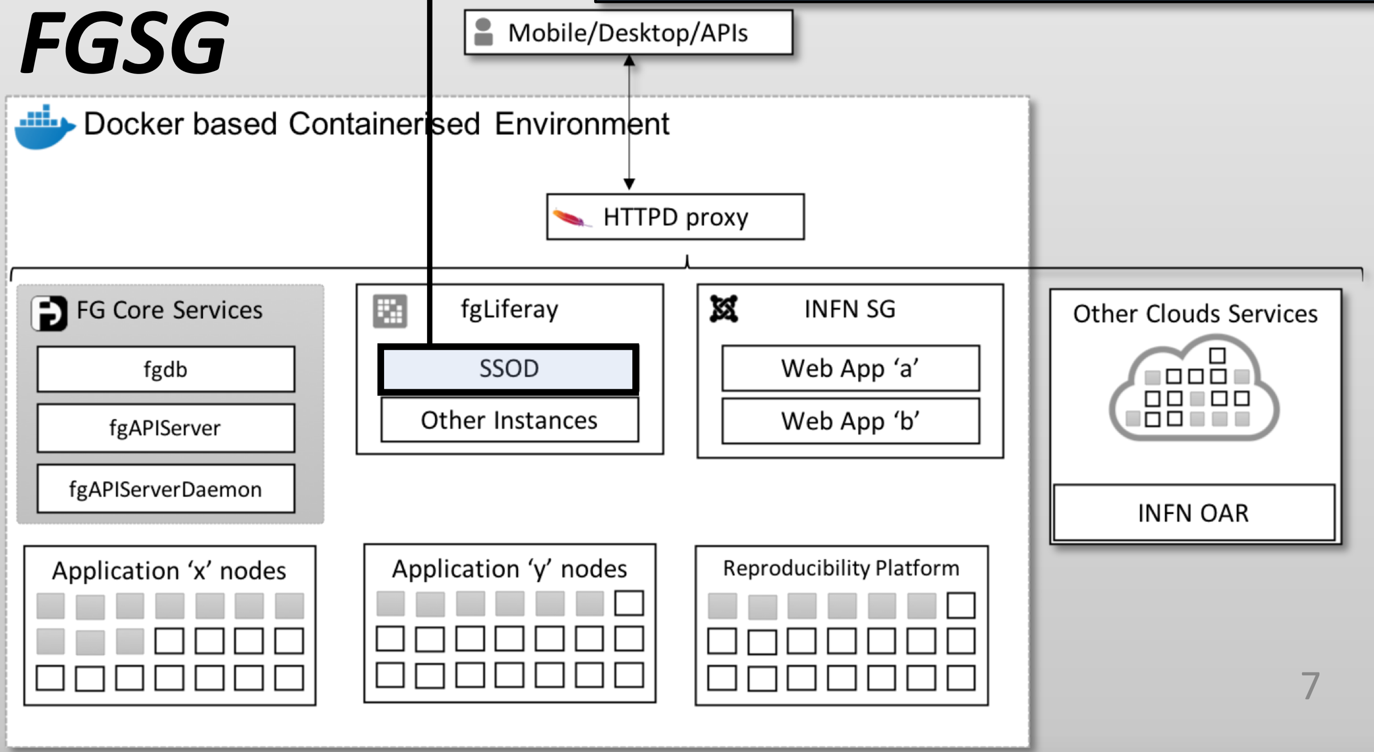
Mailing list

You may be also interested to join the FGSG mailing-list to be updated about the activities related to this web site.

Alternative authentication

This website automatically uses EGI AAI authentication source as default. It is possible to use other authentication sources, by clicking this paragraph title.

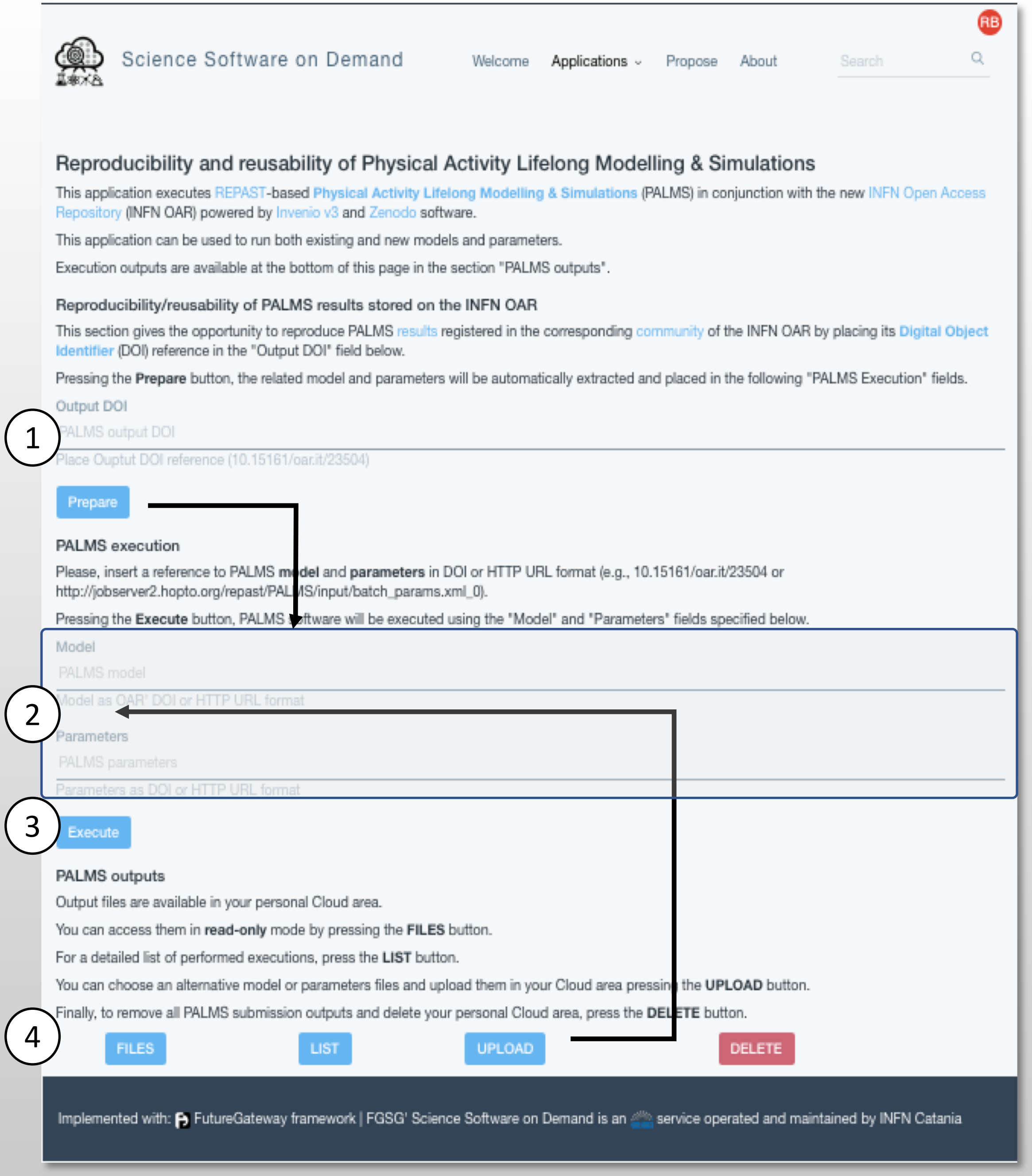
Implemented with: FutureGateway framework | FGSG' Science Software on Demand is an service operated and maintained by INFN Catania



[8] https://www.eosc-hub.eu/
[9] https://www.egi.eu/
[10] https://fgsg.egi.eu/egissod/web/ssod/

R&R for PALMS in SSOD

- Web application making use of FG REST APIs to reproduce and reuse PALMS results.
- Usage:
  1. Specify a PALMS output DOI registered in the INFN Open Access Repository -> ‘Prepare’
  2. Specify PALMS input DOIs (model+prameters). They can also be HTTP URLs
  3. Execute PALMS with input specified at 2.
  4. User can access PALMS reproduced outputs; User can upload modified input files (reusability)



Reproducibility and reusability of Physical Activity Lifelong Modelling & Simulations

This application executes REPAST-based **Physical Activity Lifelong Modelling & Simulations** (PALMS) in conjunction with the new INFN Open Access Repository (INFN OAR) powered by Invenio v3 and Zenodo software.

This application can be used to run both existing and new models and parameters.

Execution outputs are available at the bottom of this page in the section "PALMS outputs".

Reproducibility/reusability of PALMS results stored on the INFN OAR

This section gives the opportunity to reproduce PALMS results registered in the corresponding community of the INFN OAR by placing its **Digital Object Identifier** (DOI) reference in the "Output DOI" field below.

Pressing the **Prepare** button, the related model and parameters will be automatically extracted and placed in the following "PALMS Execution" fields.

Output DOI

PALMS output DOI

Place Ouptut DOI reference (10.15161/oar.it/23504)

PALMS execution

Please, insert a reference to PALMS **model** and **parameters** in DOI or HTTP URL format (e.g., 10.15161/oar.it/23504 or http://jobserver2.hopto.org/repast/PALMS/input/batch_params.xml_0).

Pressing the **Execute** button, PALMS software will be executed using the "Model" and "Parameters" fields specified below.

Model

PALMS model

Model as OAR' DOI or HTTP URL format

Parameters

PALMS parameters

Parameters as DOI or HTTP URL format

PALMS outputs

Output files are available in your personal Cloud area.

You can access them in **read-only** mode by pressing the **FILES** button.

For a detailed list of performed executions, press the **LIST** button.

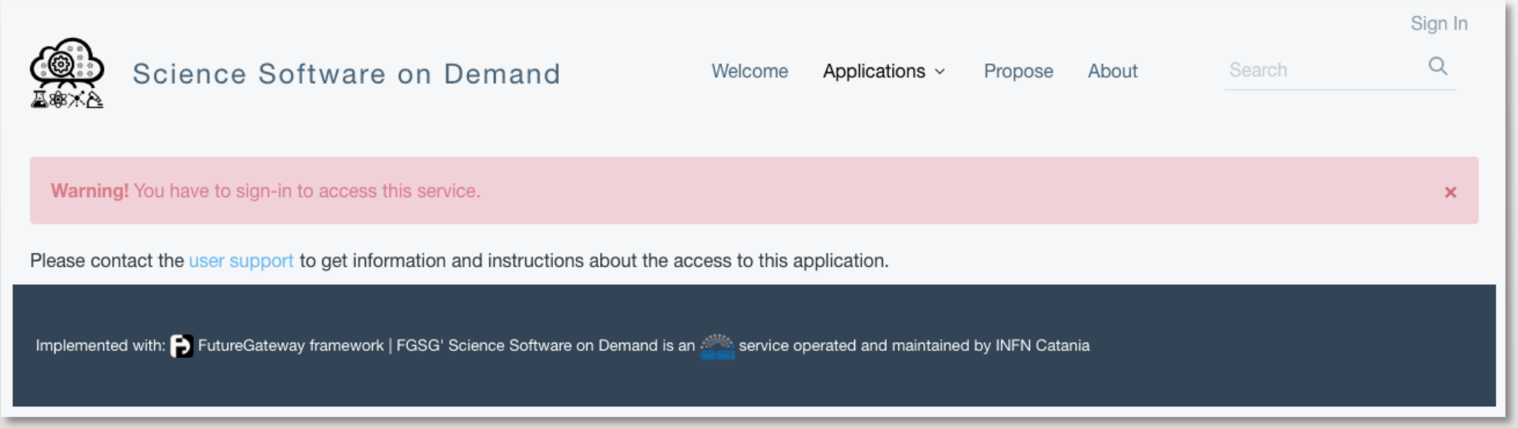
You can choose an alternative model or parameters files and upload them in your Cloud area pressing the **UPLOAD** button.

Finally, to remove all PALMS submission outputs and delete your personal Cloud area, press the **DELETE** button.

Implemented with: FutureGateway framework | FGSG' Science Software on Demand is an service operated and maintained by INFN Catania

User Auditing and Accounting in SSOD

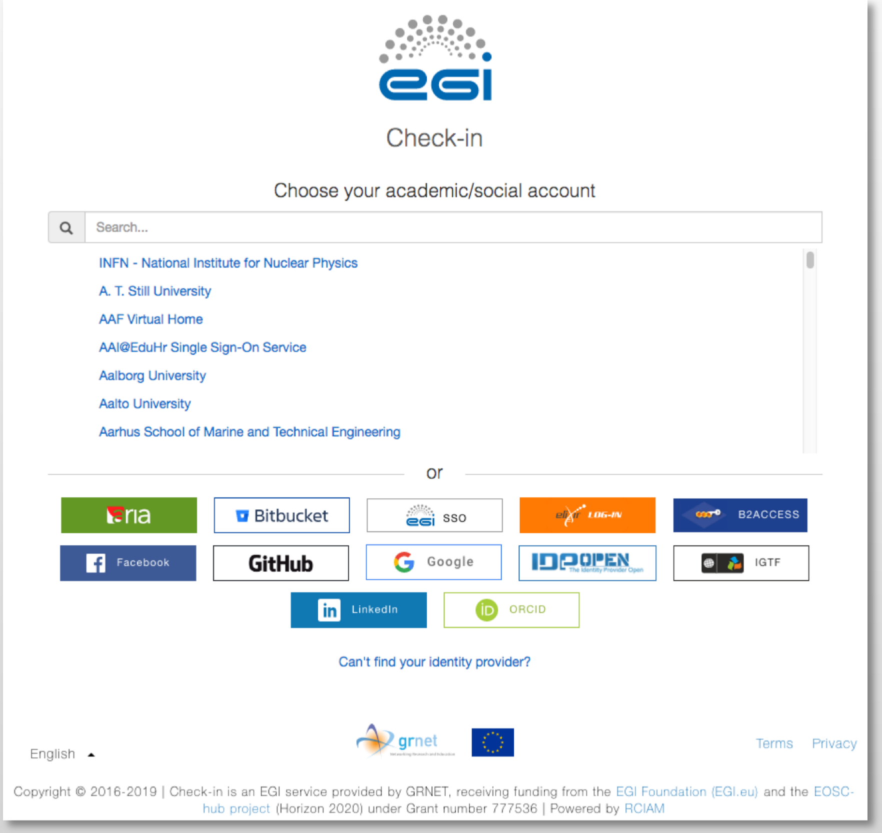
- Only registered users can access the application



Warning! You have to sign-in to access this service. ×

Please contact the user support to get information and instructions about the access to this application.

Implemented with: FutureGateway framework | FGSG' Science Software on Demand is an service operated and maintained by INFN Catania



Check-in

Choose your academic/social account

Search...

- INFN - National Institute for Nuclear Physics
- A. T. Still University
- AAF Virtual Home
- AAI@EduHr Single Sign-On Service
- Aalborg University
- Aalto University
- Aarhus School of Marine and Technical Engineering

or

Can't find your identity provider?

Copyright © 2016-2019 | Check-in is an EGI service provided by GRNET, receiving funding from the EGI Foundation (EGI.eu) and the EOSC-hub project (Horizon 2020) under Grant number 777536 | Powered by RCIAM

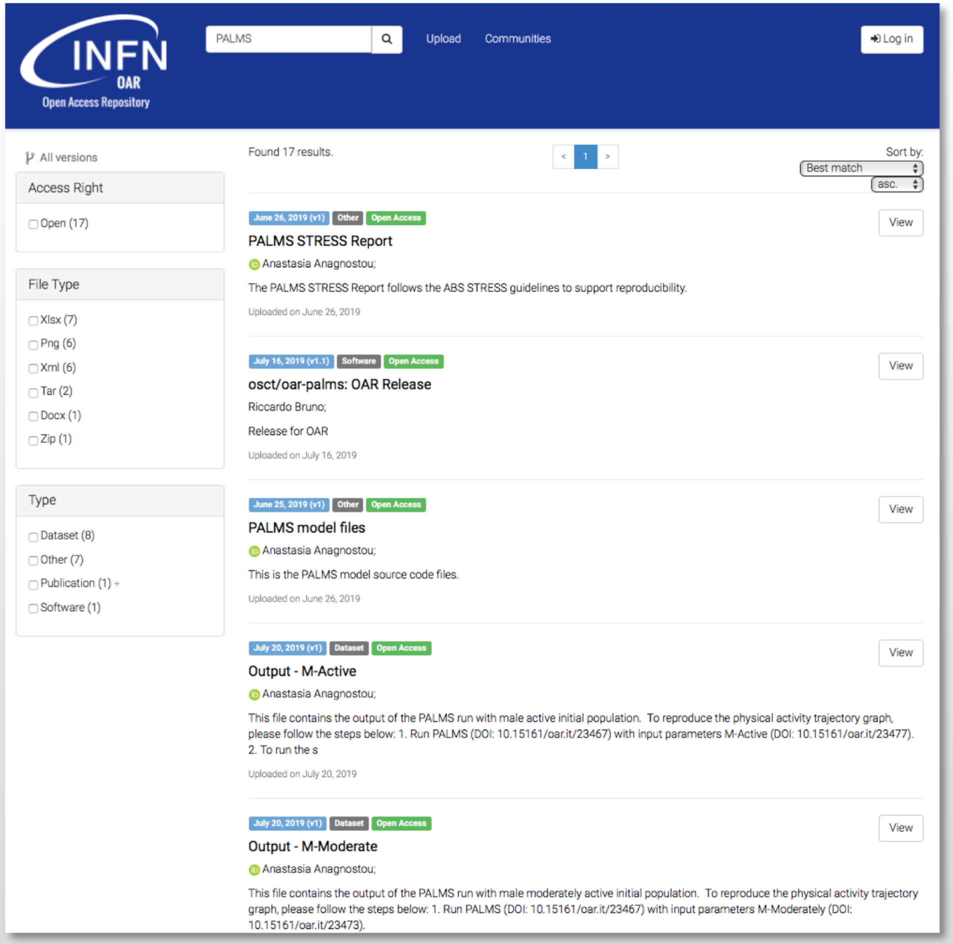
- SSOD uses EGI Check-in Proxy [11] (Federated Authentication)
  - It supports many authentication sources available in EU and supports Social Networks accounts as well
  - User activity performed by users is tracked by the FG with AAA API set (currently querying the DB)
  - Once logged the access to the application has to be authorized (not at the moment)

[11] https://www.egi.eu/services/check-in/

INFN OAR

Open Access Repository

- INFN joined the Plan S [12] initiative to promote open access
- INFN OAR is a DAMS hosted at INFN Catania [13] running on a dedicated Kubernets cluster
- It uses Zenodo [14] open source DAMS software
- PALMS input, output, software and papers files are registered with a referencing DOI
- INFN OAR allows to define references among registered DOIs and supports DOI versioning
- Software used by *FGSG* to run PALMS simulations is published as well on the INFN OAR (linked from GitHub)



Found 17 results.

Access Right

□ Open (17)

File Type

□ Xlsx (7)
□ Png (6)
□ Xml (6)
□ Tar (2)
□ Docx (1)
□ Zip (1)

Type

□ Dataset (8)
□ Other (7)
□ Publication (1) +
□ Software (1)

June 26, 2019 (v1) Other Open Access

PALMS STRESS Report

Anastasia Anagnostou;

The PALMS STRESS Report follows the ABS STRESS guidelines to support reproducibility.

Uploaded on June 26, 2019

July 16, 2019 (v1.1) Software Open Access

osct/oar-palms: OAR Release

Riccardo Bruno;

Release for OAR

Uploaded on July 16, 2019

June 25, 2019 (v1) Other Open Access

PALMS model files

Anastasia Anagnostou;

This is the PALMS model source code files.

Uploaded on June 26, 2019

July 20, 2019 (v1) Dataset Open Access

Output - M-Active

Anastasia Anagnostou;

This file contains the output of the PALMS run with male active initial population. To reproduce the physical activity trajectory graph, please follow the steps below: 1. Run PALMS (DOI: 10.15161/oar.it/23467) with input parameters M-Active (DOI: 10.15161/oar.it/23477). 2. To run the s

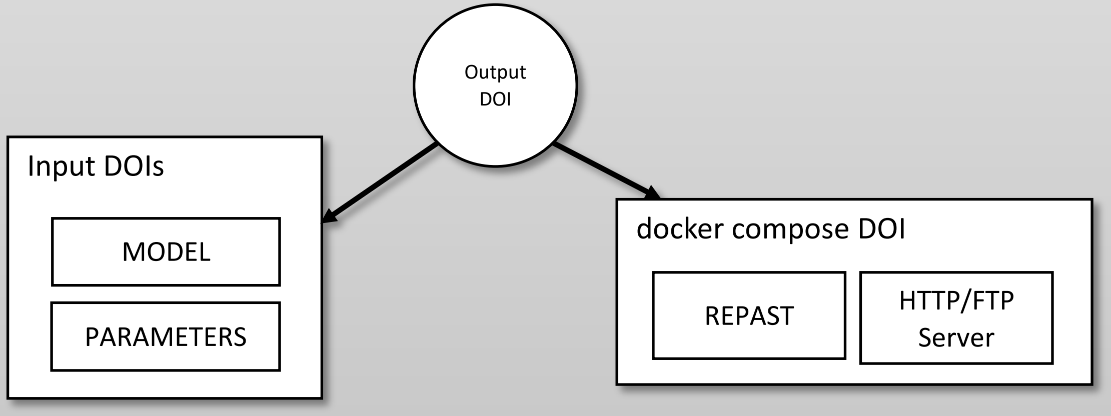
Uploaded on July 20, 2019

July 20, 2019 (v1) Dataset Open Access

Output - M-Moderate

Anastasia Anagnostou;

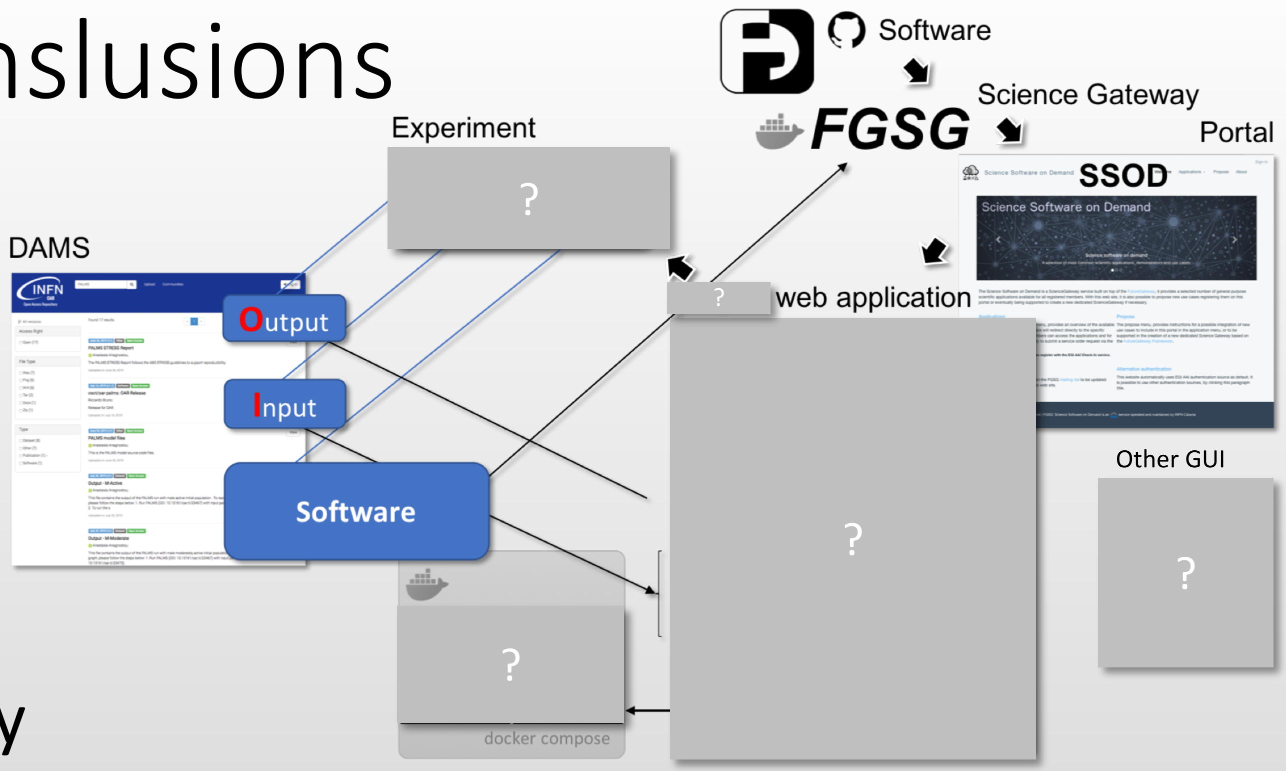
This file contains the output of the PALMS run with male moderately active initial population. To reproduce the physical activity trajectory graph, please follow the steps below: 1. Run PALMS (DOI: 10.15161/oar.it/23467) with input parameters M-Moderately (DOI: 10.15161/oar.it/23473).



[12] https://www.coalition-s.org
[13] https://www.openaccessrepository.it
[14] https://github.com/zenodo/zenodo

Summary and conslusions

- PALMS use case has a quite simple structure
- The R&R platform developed for PALMS can be easily used to host other use cases
- OAR + R&R Platform are open to any interested community



“Reusing” Reprodiciblity and Reusabiltiy platform

- FutureGateway is a mature product successfully used by:
  - Desktop and Mobile Applications, Workflow engines, IoT and **OpenScience**
- Investigations are in progress to extend this work to a more general and widely usable solution

?

Further info

FAIR Principles in Open Science

As expressed by GOFAIR , to be compliant with the Open Science paradigm, research data should be FAIR:

- **F**indable

  F1. (Meta)data are assigned a globally unique and persistent identifier
  F2. Data are described with rich metadata (defined by R1 below)
  F3. Metadata clearly and explicitly include the identifier of the data they describe
  F4. (Meta)data are registered or indexed in a searchable resource

- **A**ccessible

  A1.1 The protocol is open, free, and universally implementable
  A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

- **I**nteroperable

  I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  I2. (Meta)data use vocabularies that follow FAIR principles
  I3. (Meta)data include qualified references to other (meta)data

- **R**eusable

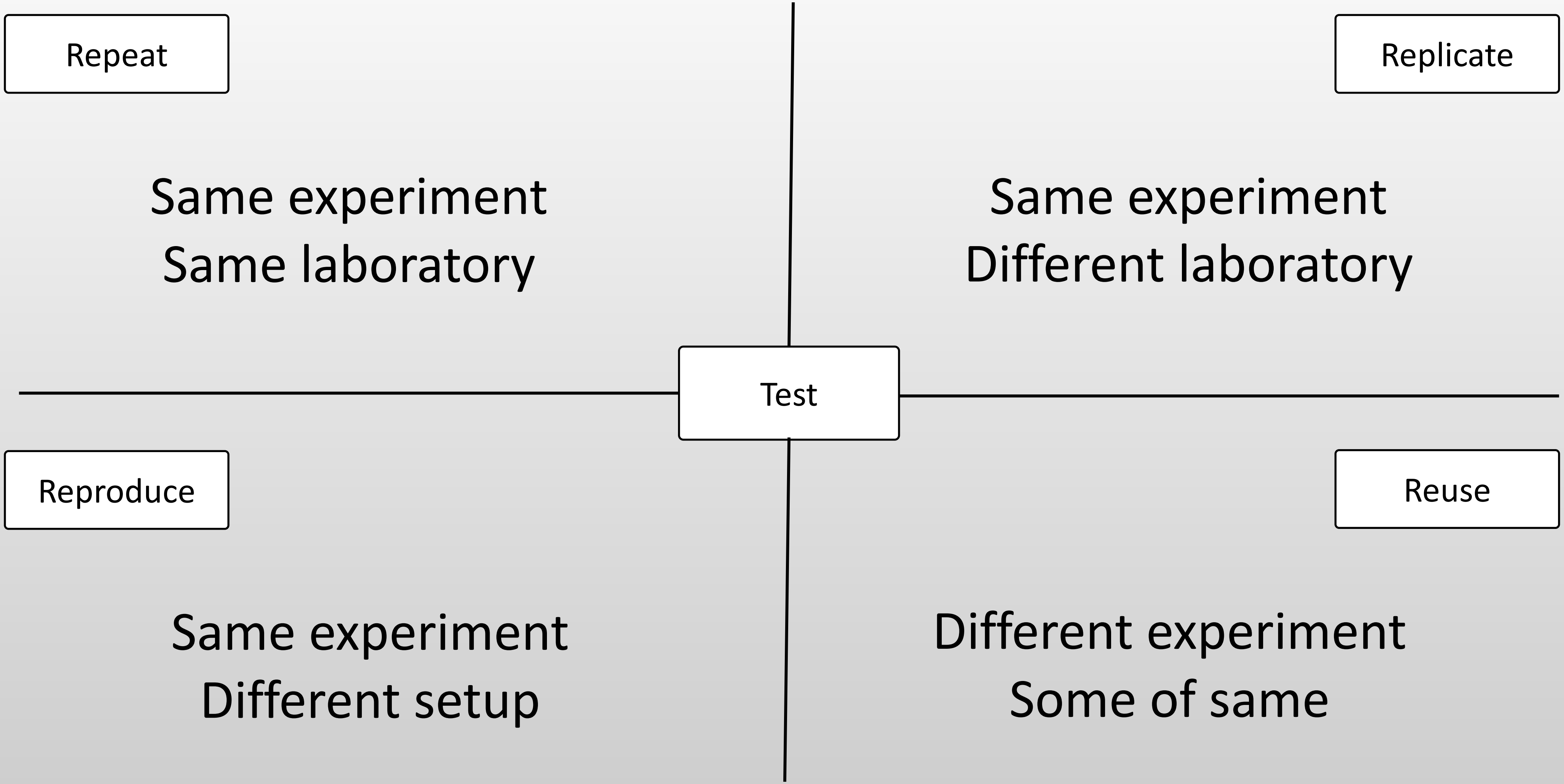
  R1.1. (Meta)data are released with a clear and accessible data usage license
  R1.2. (Meta)data are associated with detailed provenance
  R1.3. (Meta)data meet domain-relevant community standards

The «pillars» of the Scientific Method

- **Repeatability**
  - The closeness of agreement between independent results obtained with the same method on identical test material, under the same conditions (same operator, same apparatus, same laboratory and after short intervals of time)
  - Affected by *random errors*

- **Reproducibility**
  - The closeness of agreement between independent results obtained with the same method on identical test material but under different conditions (different operators, different apparatus, different laboratories and/or after different intervals of time)
  - Affected by *systematic errors*

The (Re-)'s of Science



Repeat

Same experiment
Same laboratory

Replicate

Same experiment
Different laboratory

Test

Reproduce

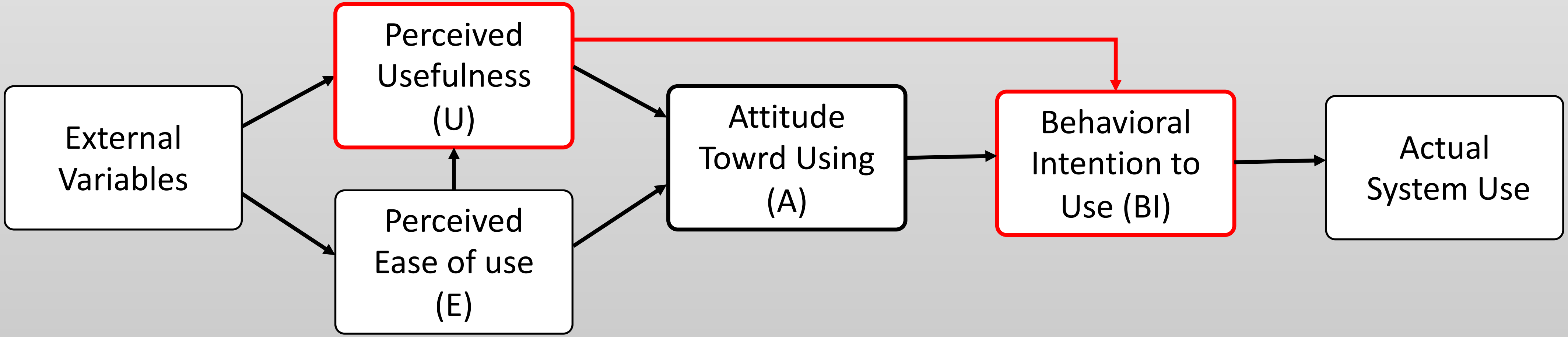
Same experiment
Different setup

Reuse

Different experiment
Some of same

Open Science (a receipe rather than a definition)

Open science and research refers to effort to promote open procedures in scientific research activities. The key objective is, in the context set by research ethics and legal frameworks, to publish research outputs (research publications, research data, research methods) so the can be examined and user by any interested party.
Open science and research involves practices, such as promoting open access to research publications, open availability fo research data, harnessing open source software and open standards, and open documentation of research process.



Source: Davis, F. D. (1989), "Perceived usefulness, perceived ease of use, and user acceptance of information technology», MIS Quarterly, **13**(**3**): 319-340

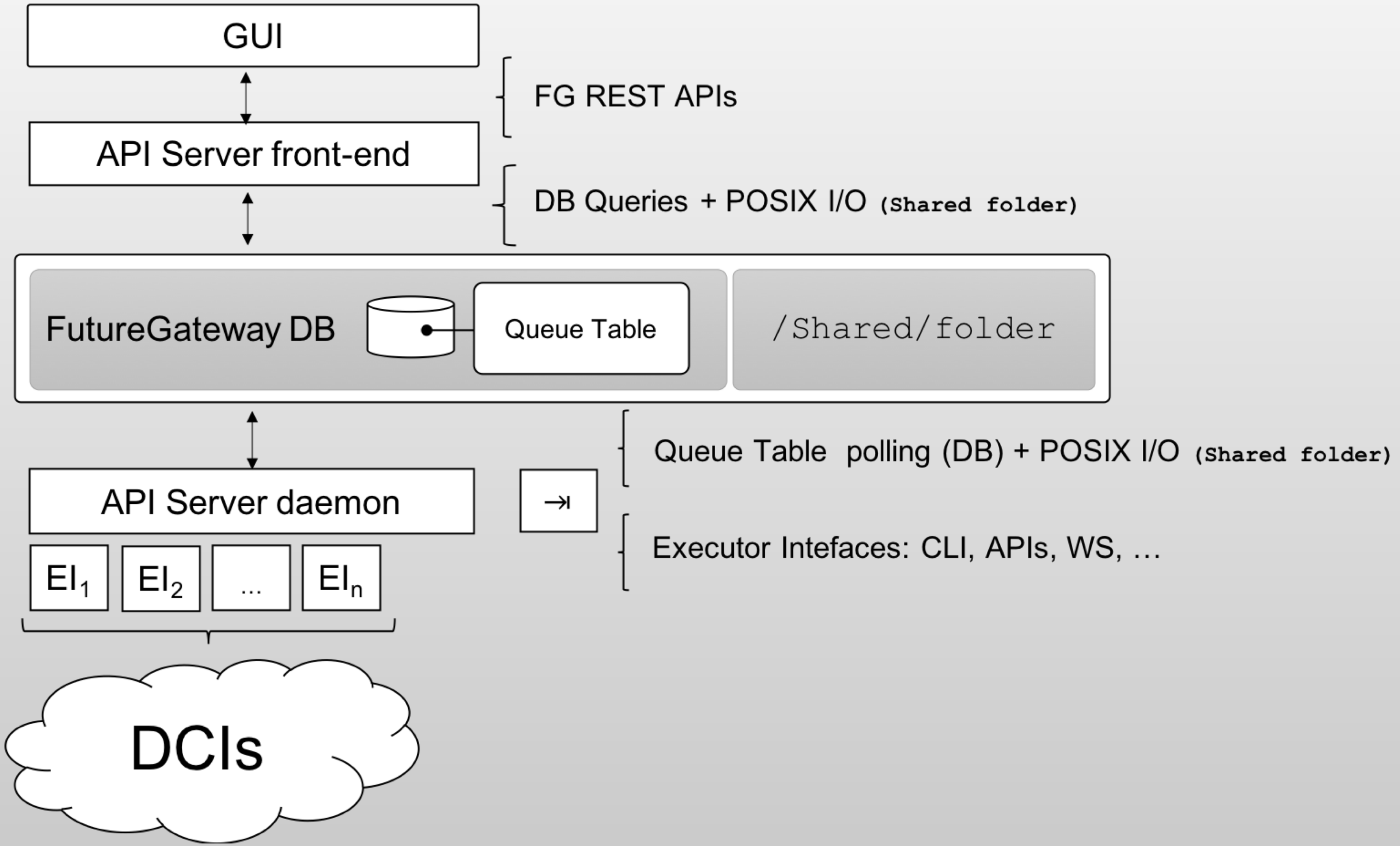
Concepts and definitions

- **Open Access Repositories** (OARs) are powered by **Digital Asset Management Systems** (DAMSes), which are "intertwined structures incorporating both software and hardware that take care of management tasks and decisions surrounding the ingestion, annotation, cataloguing, storage, retrieval and distribution of **digital assets**"

- A **digital asset** in essence is "anything that exists in a binary format and comes with the right to use"

- "Types of digital assets include, but are not exclusive to, photography, logos, illustrations, animations, audio-visual media, presentations, spreadsheets, Word and/or PDF documents, data and a multitude of other digital formats and their respective metadata"

FutureGateway Architecture



GUI

FG REST APIs

API Server front-end

DB Queries + POSIX I/O (Shared folder)

FutureGateway DB

Queue Table

/Shared/folder

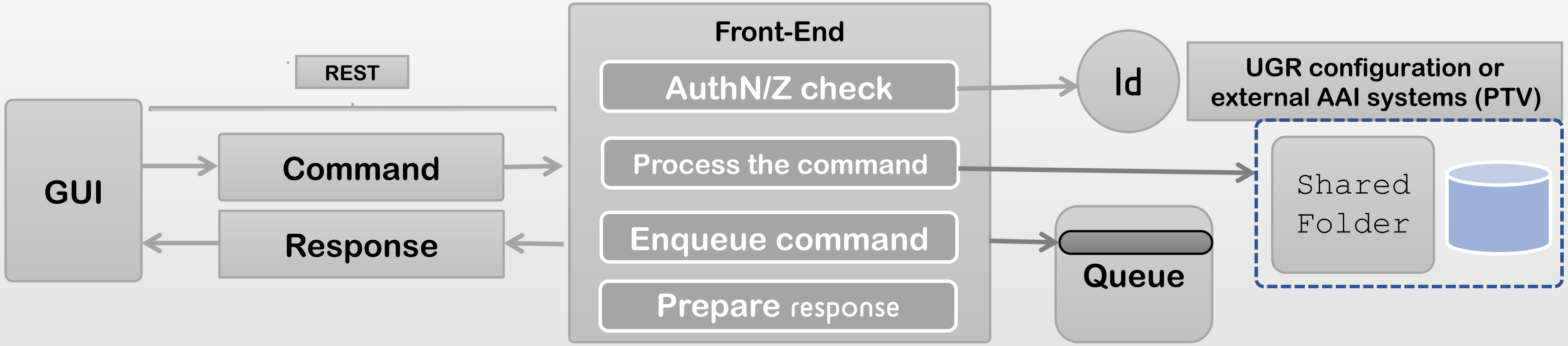
Queue Table polling (DB) + POSIX I/O (Shared folder)

API Server daemon

→|

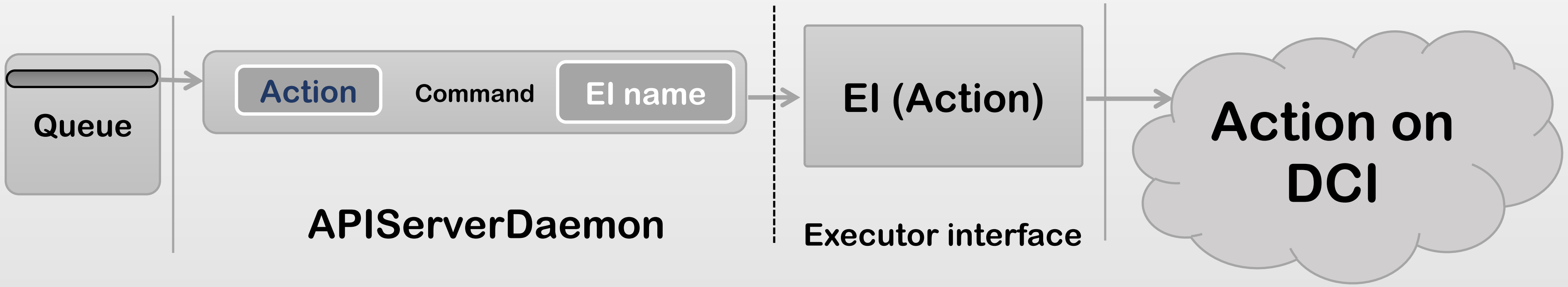
Executor Intefaces: CLI, APIs, WS, …

API Server front-end



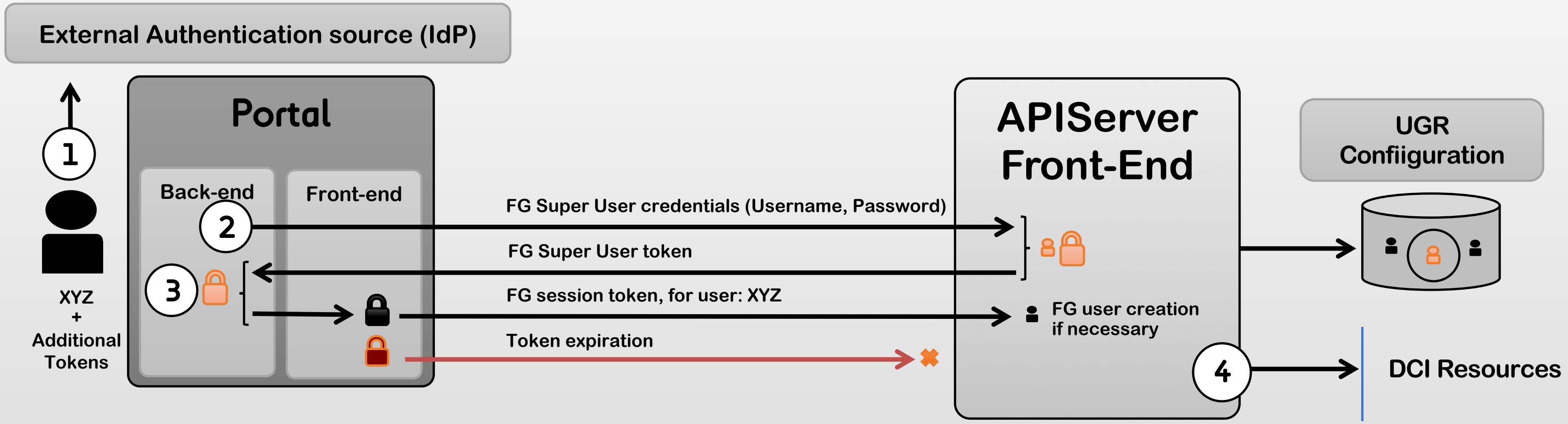
- Operations
  - GUIs send a command via REST APIs
  - The ‘command’ may contain a JSON stream specifying command parameters
  - The Front-End first check for requestor Authorization and Authentication eventually using UGR configuration or external AAI mechanisms using PTV information
  - The command is processed querying and/or updating the DB accordingly and/or updating the shared folder
  - Commands to be finalized by the APIServer daemon, are stored in the queue table
  - Command output is returned back into the response as a JSON stream

API Server daemon



- Operations
  - Commands (Tasks=Command(Action,EI)) are extracted from the front-end queue
  - Each 'command' contains the 'Target Executor' field which specifies the Executor Interface name
  - Executor interfaces are dynamically instantiated by the APIServer by its name, applying the specified action on DCI
  - Other queue daemons may extract commands from the queue having their own EIs implemented.
  - New EIs can be easily developed just inheriting a standard Executor Interface class
  - Current available EIs:
    - GridEngine (A core component of the CSGF using JSAGA and targeting: ssh, rOCCI and wms)
    - ToscaIDC (INDIGO-dc orchestrator with IAM)

FutureGateway AuthN/Z (UGR: user-behalf)



1 **Log-In**

The portal authenticate users using an external Authorization source (IdP). Once authenticated it receives a unique user name (XYZ) + eventually additional tokens.

3 **User session token**

The UGR APIs can obtain a user session token on behalf of another user, only super-user can do this operation (see role: user_impersonate).

During this opertation, the back-end Portal should create a new FG user if not yet present.

2 **FG Super user access token**

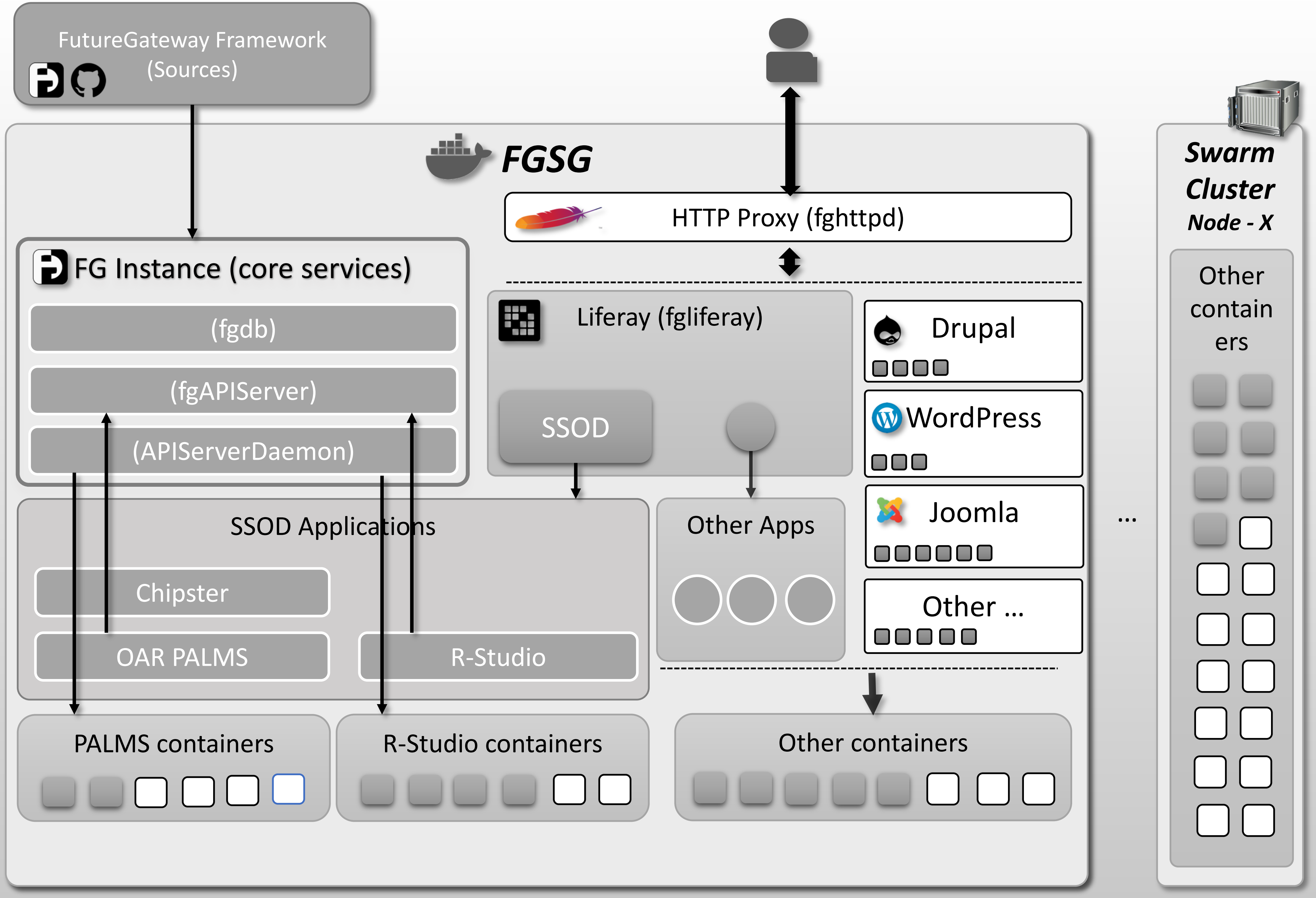
With UGR API calls, it is possible to obtain a session token having FG super user capabilities. This operation should be done in a protected environment such as a portal back-end environment.

4 **Additional tokens**

During the external authentication, additional tokens may be retrieved. This information can be used as FG username or saved using URG APIs. This kind of tokens may be needed by FG applications to access DCIs resources (See INDIGO-dc IAM).

FGSG Structure

IBM Blade center-X
One slot with 32 cores
128 GB RAM
Ubuntu 18.04
Docker version 17.12.1-ce
150 GB on board RAID-2
5TB ext. Storage



FutureGateway Framework
(Sources)

FGSG

HTTP Proxy (fghttpd)

FG Instance (core services)
(fgdb)
(fgAPIServer)
(APIServerDaemon)

Liferay (fgliferay)
SSOD

Drupal
WordPress
Joomla
Other ...

SSOD Applications
Chipster
OAR PALMS
R-Studio

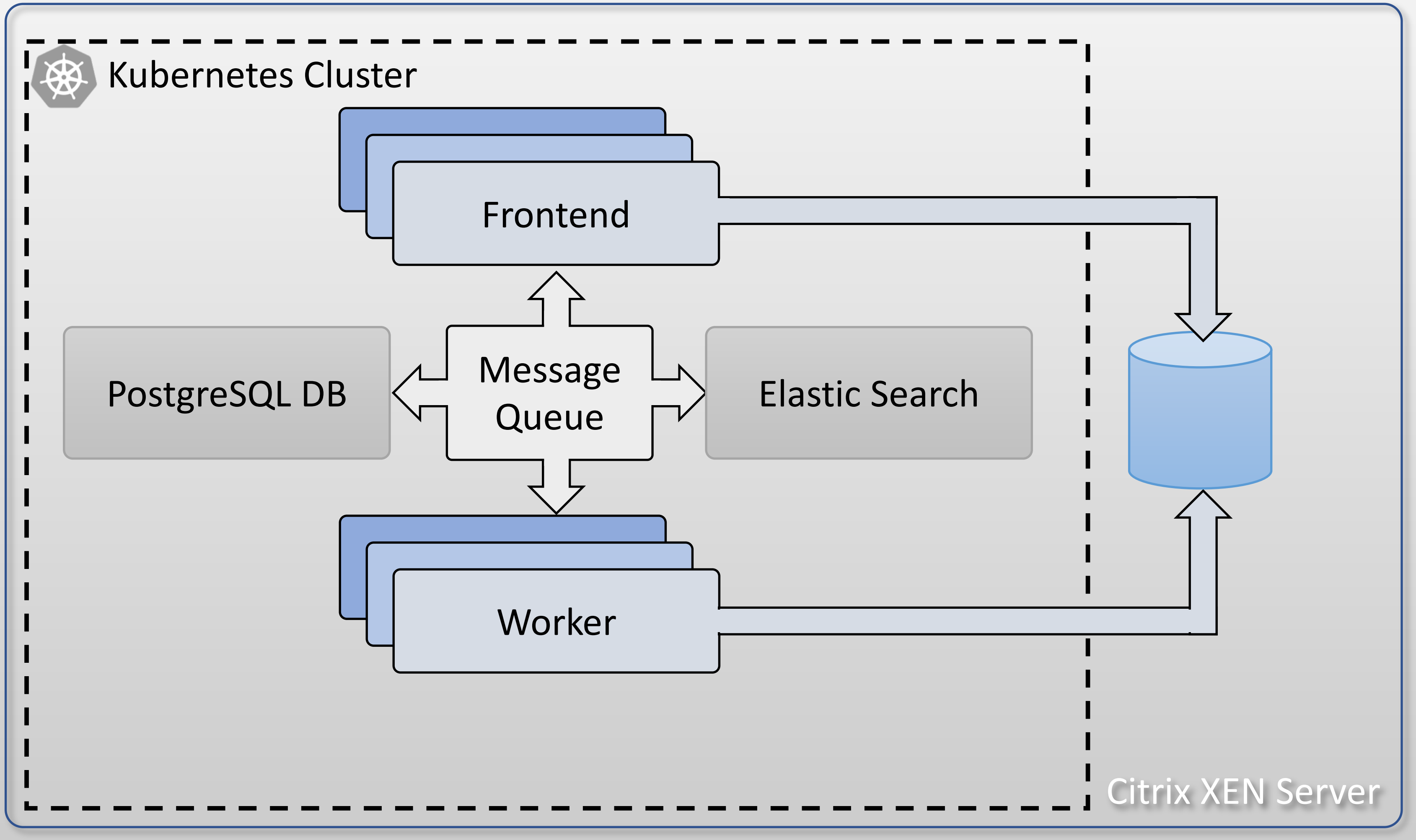
Other Apps

PALMS containers
R-Studio containers
Other containers

...

Swarm Cluster
Node - X
Other containers

INFN OAR Deployment Structure


Kubernetes Cluster
Frontend
PostgreSQL DB
Message Queue
Elastic Search
Worker
Citrix XEN Server