Science Gateways for Open Science

Roberto Barbera – University of Catania and INFN
(roberto.barbera@ct.infn.it)

DPHEP Collaboration Workshop
CERN – 8 June 2015

Outline

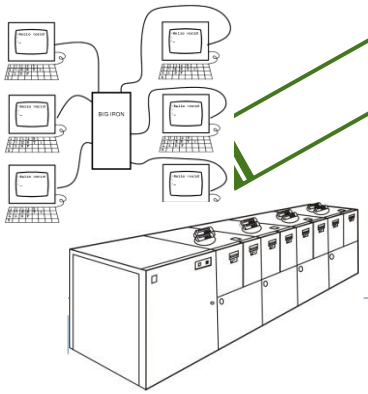
- Introductory concepts and driving considerations
- INFN approach to Open Access and Open Science
- Open Science with the Catania Science Gateway Framework
- Summary and conclusions

Evolution of Distributed Computing

Cost of hw
Cost of networks

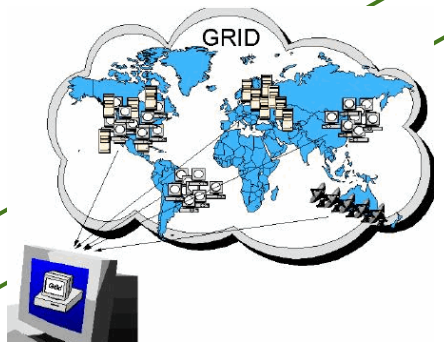
Time

Mainframe Computing

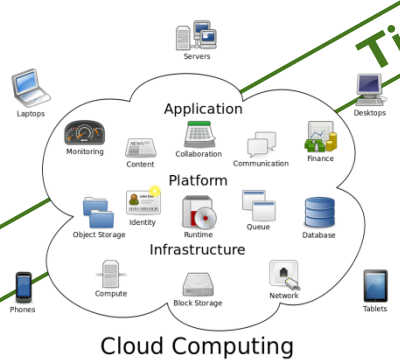


80's-90's Cluster Computing

90's-00's Grid Computing



00's-10's Cloud Computing



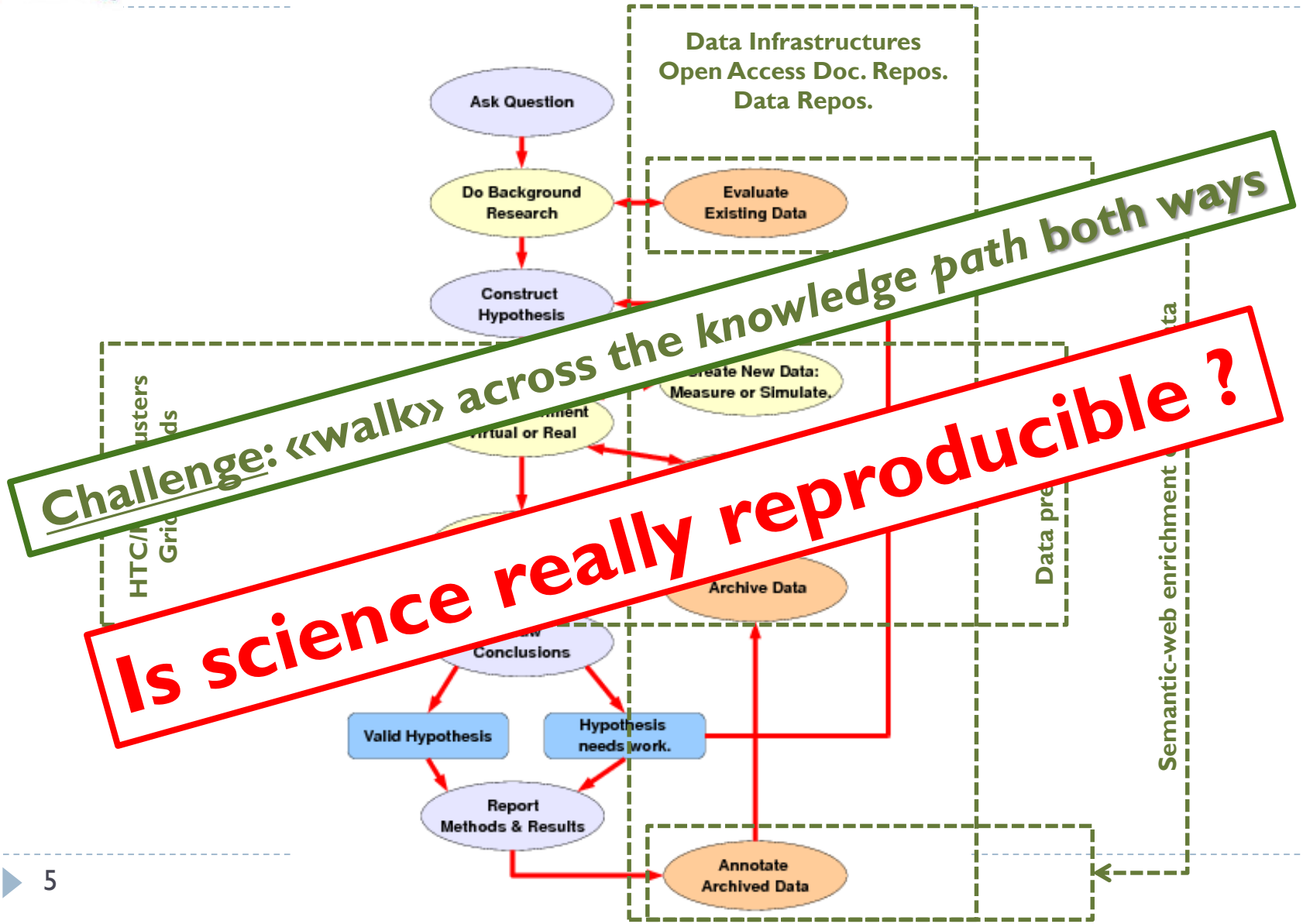
Cloud Computing

Power of COTS
WAN bandwidth

e-Science

- Virtual Research Communities
- e-Infrastructure
- Applications
- Data
- Instruments/sensors

How e-Infrastructures support the Scientific Method



Reproducibility crisis

COMMENT

AVIAN INFLUENZA Shift expertise to track mutations where they emerge p.534

EARTH SYSTEMS Past climates give valuable clues to future warming p.537

HISTORY OF SCIENCE Descartes' lost letter tracked using Google p.540

OBITUARY Wylie Vale and an elusive stress hormone p.542

Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal

Raise stan

Must try harder

Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigoro at the data — and at themselves.

Error prone

Biologists must realize the pitfalls of w massive amounts of data

If a i

...um on ensuring that ...an F. Russell.

...case for open computer programs

Six red flags for suspect work

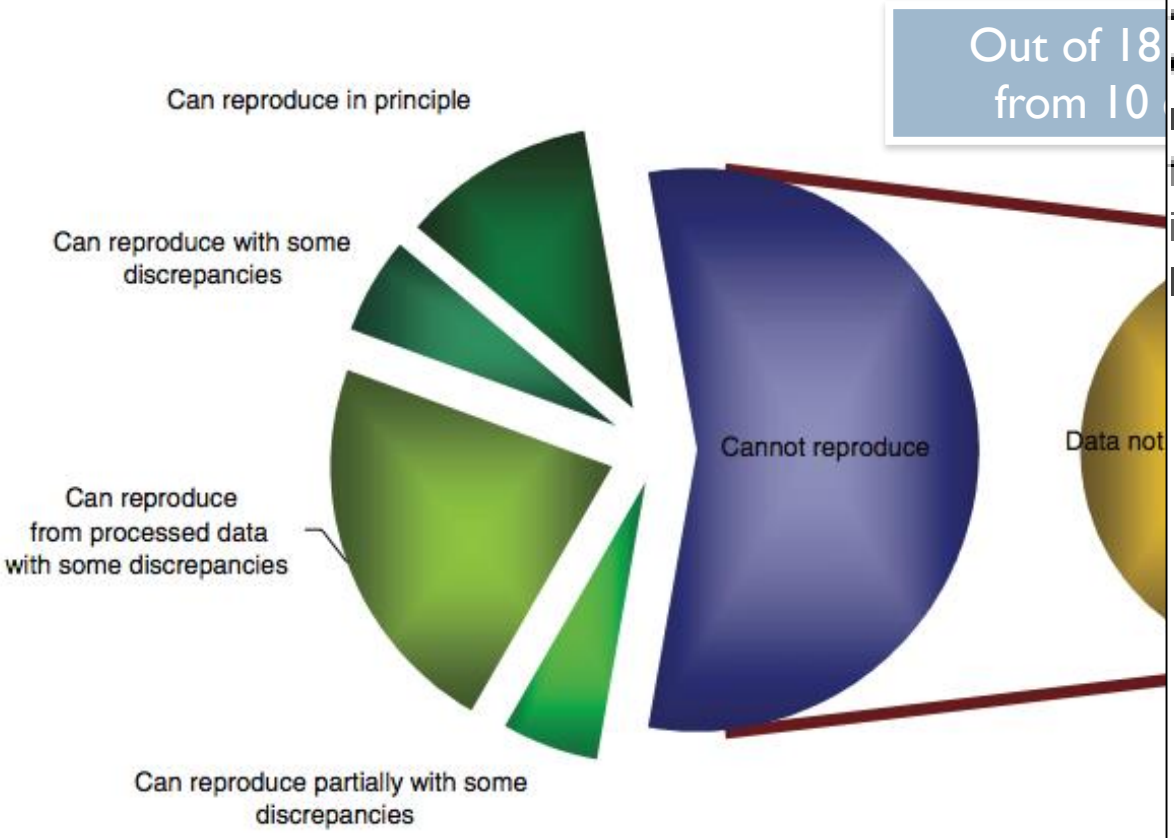
C. Glenn Begley explains how to recognize the preclinical papers in which the data won't stand up.

Know when your numbers are significant

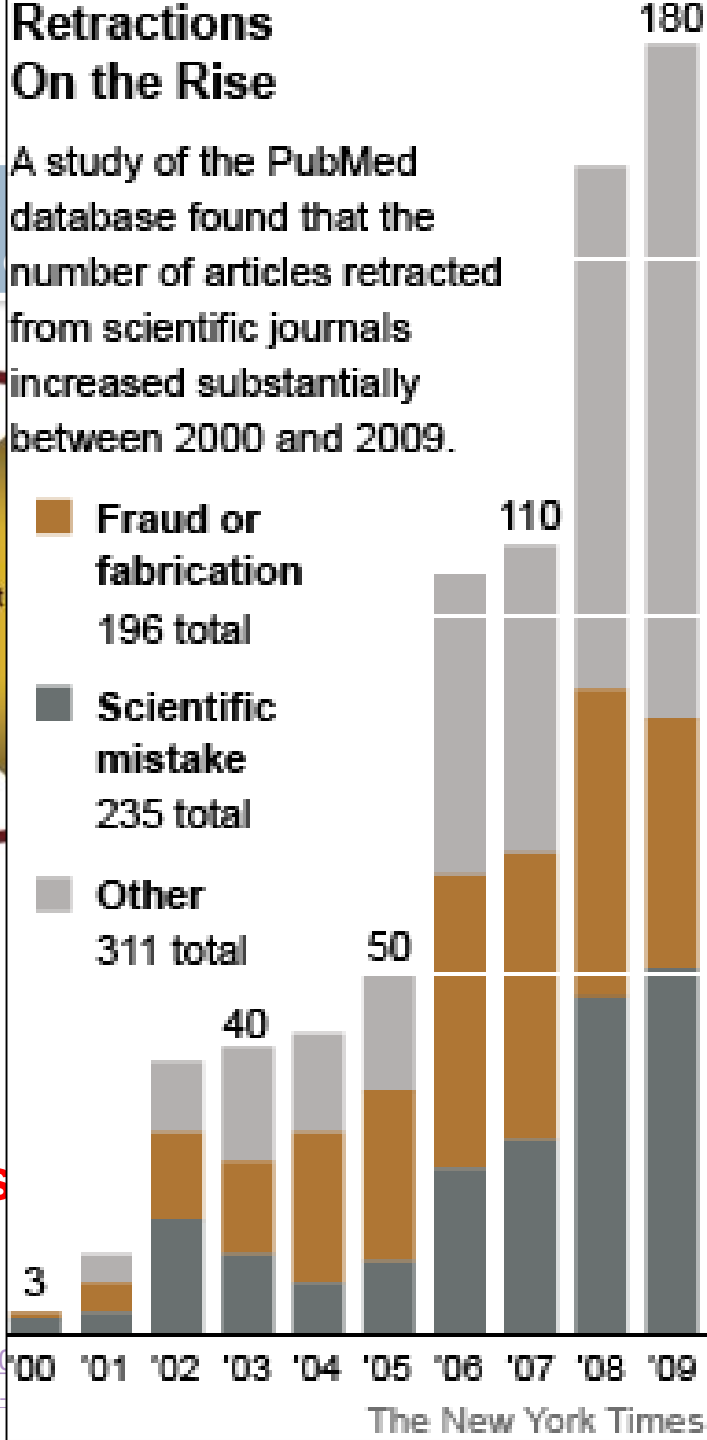
47/53 «landmark» publications could not be replicated

[Begley & Ellis, Nature 483, 2012]

Reproducibility crisis



Out of 18
from 10

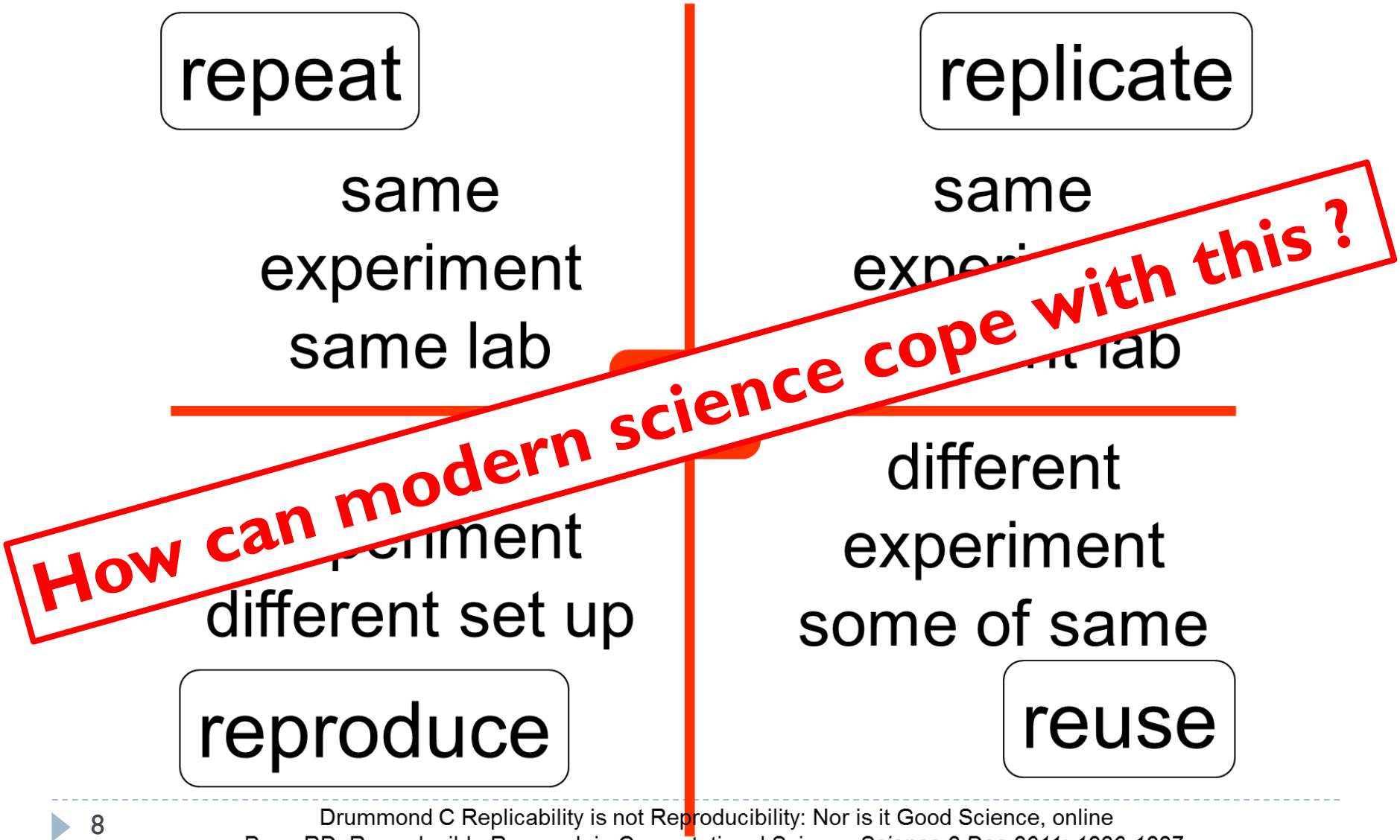


More retractions:

- **>15x increase between 2000-2009**
- **At current rate, by 2045 as many papers published as retracted**

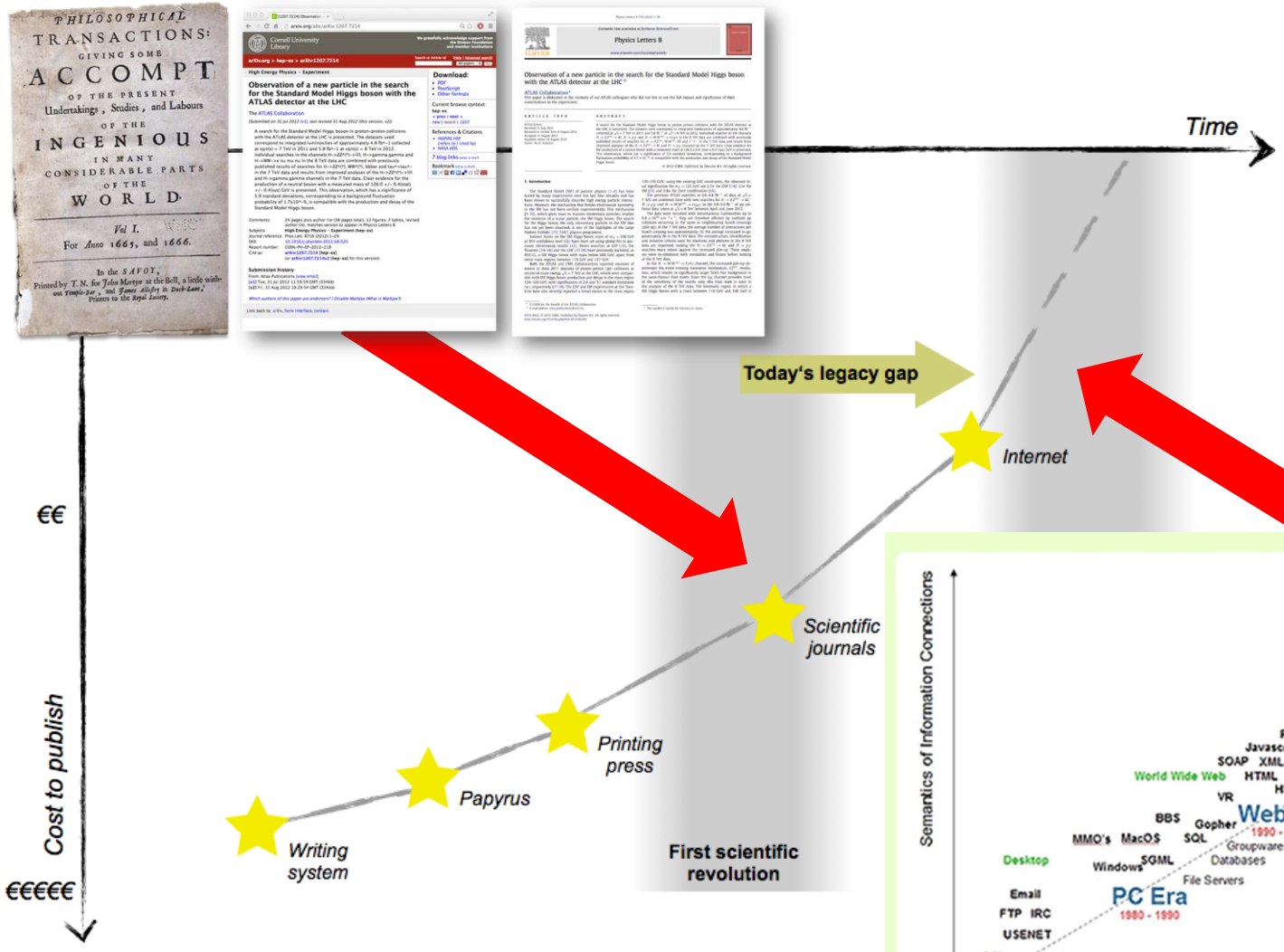
1. Ioannidis et al., 2009. Repeatability of published microarray gene expression analyses.
2. Science publishing: The trouble with retractions http://www.nature.com/news/2011/1110
3. Bjorn Brembs: Open Access and the looming crisis in science https://theconversation.com/open-

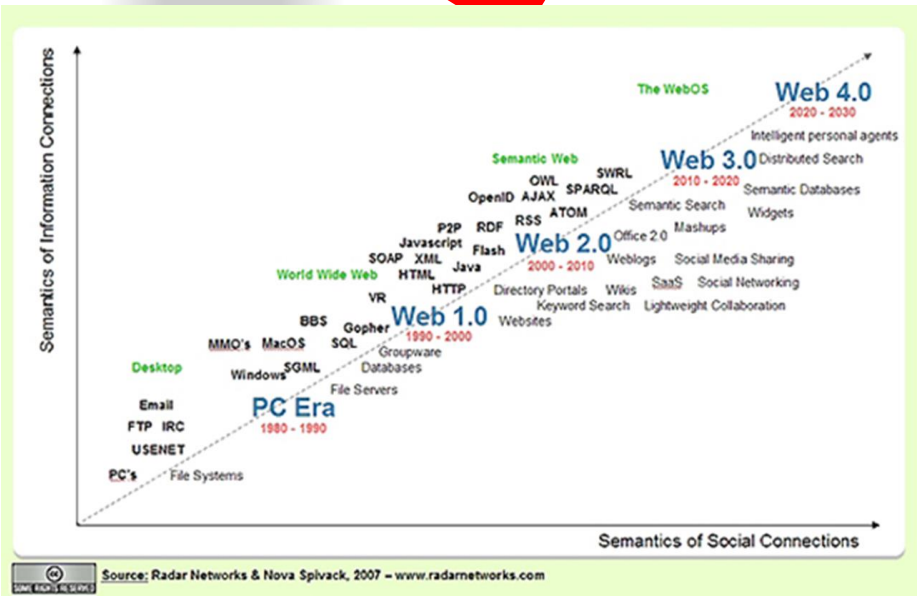
Repeatability and Reproducibility are not all



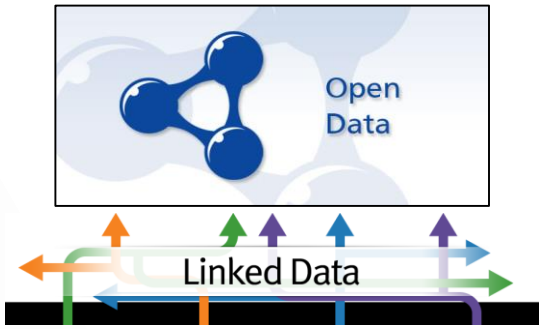
Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Peng RD, Reproducible Research in Computational Science *Science 2 Dec 2011: 1226-1227.*

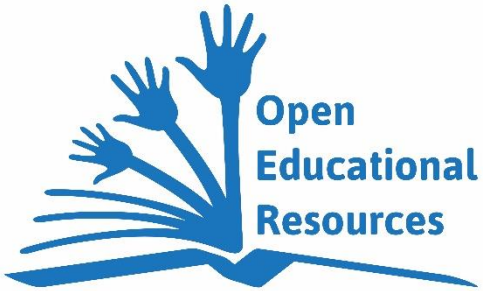
Evolution of Scientific Research
(http://book.openingscience.org)

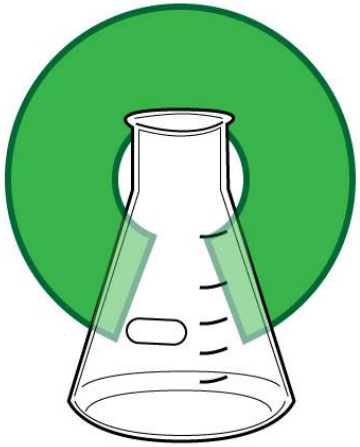




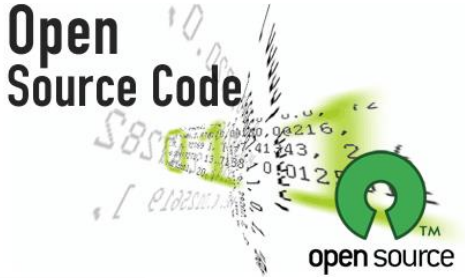
o-Science
(Open Science)





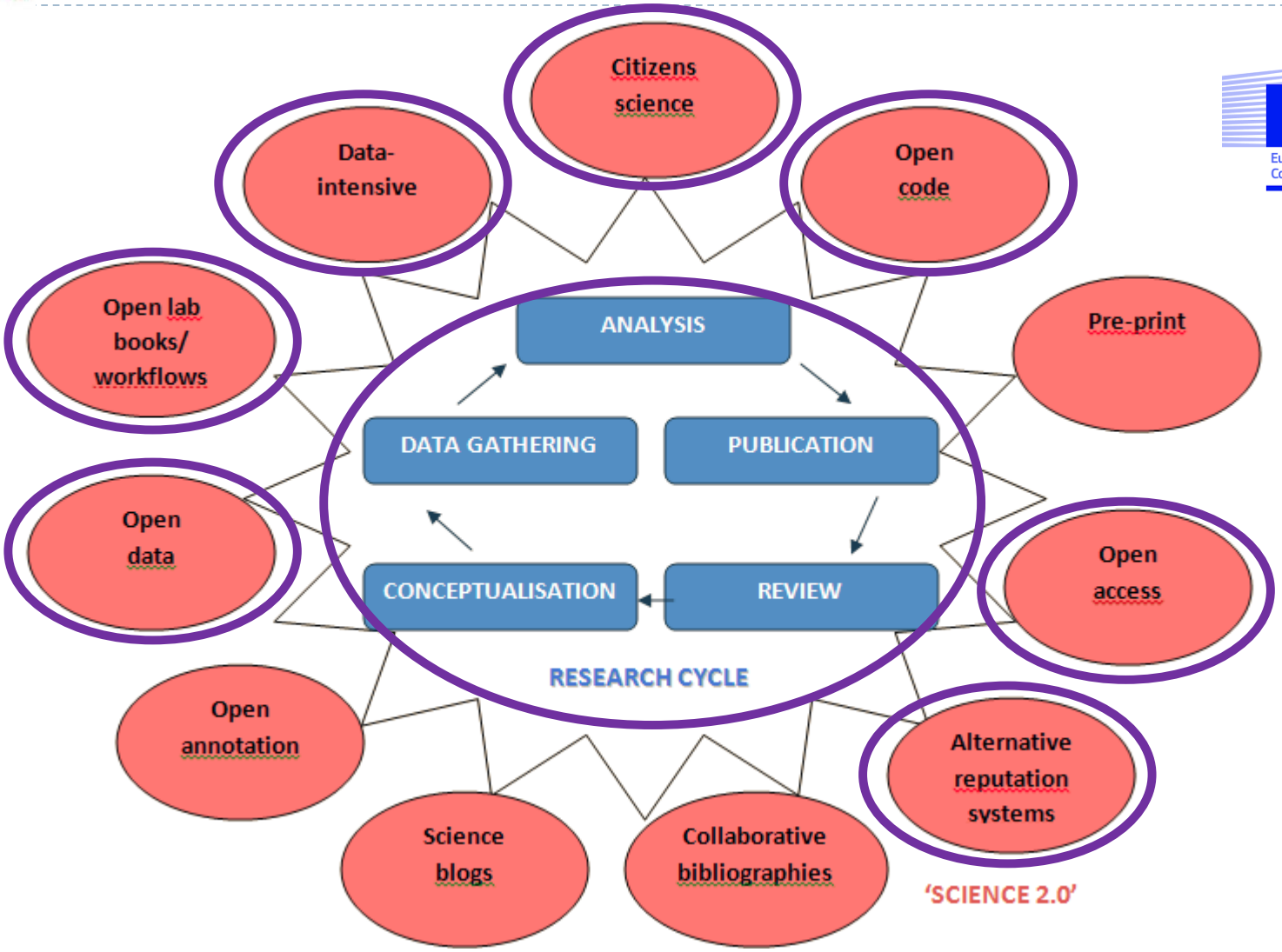


open science



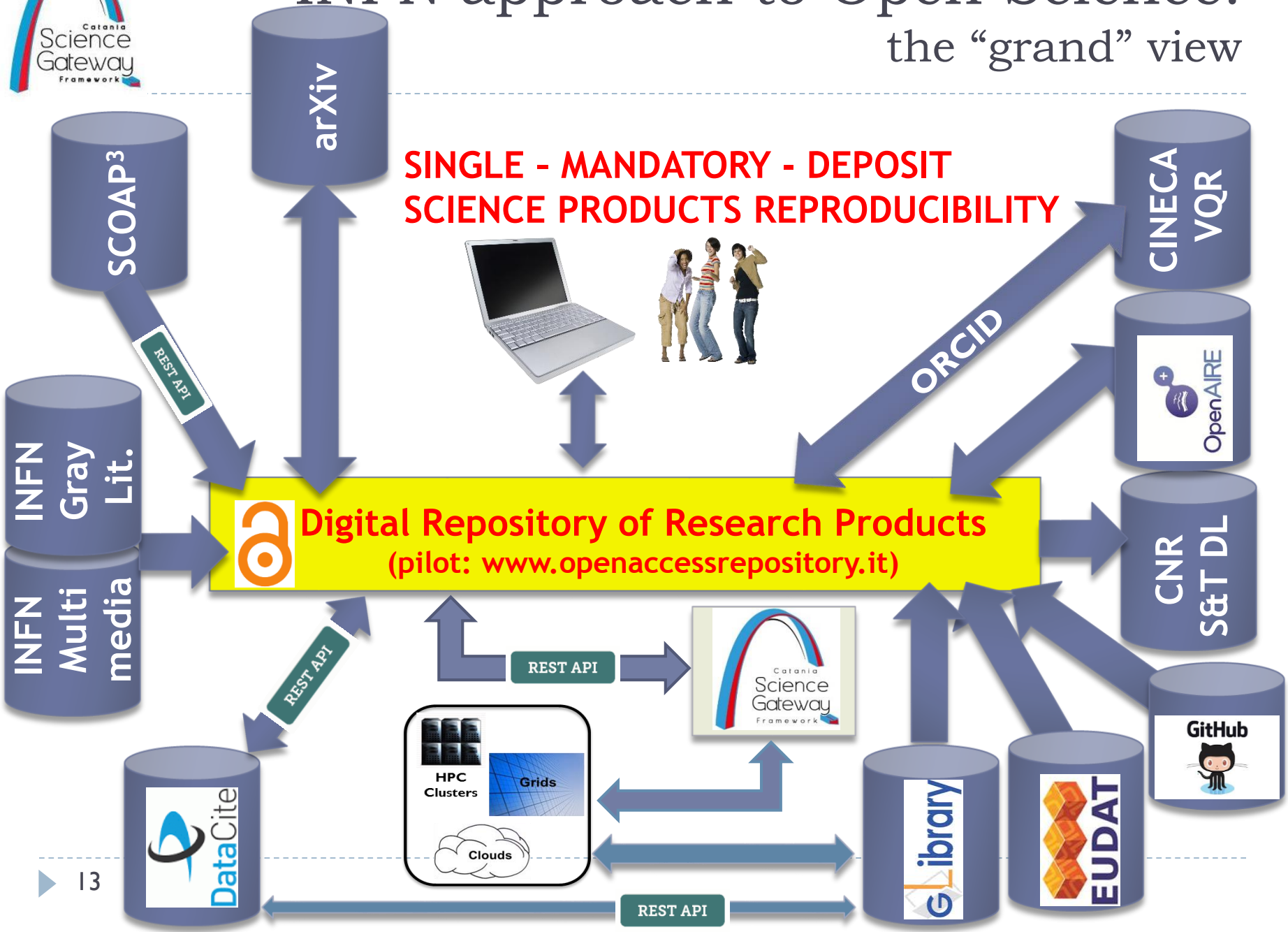
Science in transition → towards Science 2.0

(http://ec.europa.eu/research/consultations/science-2.0/background.pdf)



- Final report: http://ec.europa.eu/digital-agenda/en/news/final-report-science-20-public-consultation

INFN approach to Open Science:
the "grand" view



INFN approach to Open Science:
requirements, choices and motivations

- Requirements:
  - Open source
  - Standard compliant
  - Well supported
  - Scalable up to O(10^6)-O(10^7) resources (to begin with)
- Choice:
  - Invenio (www.invenio-software.org) + our «add-ons»
- Motivations:
  - Fully compliant with OAI-PMH and Marc21 standards
  - Co-developed by an international collaboration comprising institutes such as CERN, DESY, EPFL, FNAL, SLAC and used by about 30 scientific institutions worldwide
  - ZENODO (OpenAIRE flagship repository) and SCOAP^3 repositories are based on Invenio
  - The CERN Document Server (http://cds.cern.ch/) contains more than 1.3 million documents

The INFN Open Access Repository
(www.openaccessrepository.it)

Open Access Repository BETA

Search | Submit | Personalize | Help | Administration

INVENIO

barbera :: logout

Search 4,347 records for:

any field | Search | Browse

Search Tips :: Advanced Search

- Audio-Video Recordings (0)
  - INFN (0) Others (0)
- Datasets (185)
  - INFN (185) Others (0)
- Images (0)
  - INFN (0) Others (0)
- Presentations (4)
  - INFN (3) PSTS (0) Others (1)
- Posters (4)
  - INFN (4) Others (0)
- Publications (3,968)
  - INFN (1,147) PSTS (1) Others (2,820)
- Software (186)
  - INFN (186) Others (0)

Automatic ingestion in place from:

Cornell University Library

arXiv.org

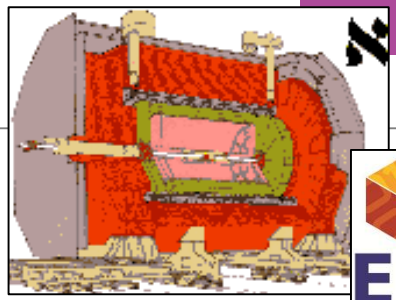
papers

SCOAP³

REST API

Crontab

data



EUDAT

ABOUT THIS SITE

Welcome to the beta version of the Open Access Repository managed and operated by INFN. Feel free to browse all the features and contents of this site as well as to upload your own open access documents and data.
To sign-up or sign-in, click on login

CERTIFICATION AND COMPLIANCE

This site is both an OAI conforming repository and an official OpenDOAR data provider. It is also one of the official OpenAIRE archives, compliant with version 3.0 of its guidelines

SEE ALSO

INFN, PSTS

federated authentication

GIDP eduGAIN idem garr aai

Open Access Repository :: Search :: Submit :: Personalize :: Help
Info :: Terms of use :: Privacy Policy :: Support/Feedback
Powered by Invenio v1.1.3.15-fe13-dirty
Maintained by INFN Catania librarian@openaccessrepository.it
Last updated: 20 Oct 2014, 14:03

This site is also available in the following languages:
English Italiano
This is a Service Provider of:

Resources' automatic harvesting & ingestion: the SCOAP^3 use case

Step 1

Call SCOAP^3 API

Connect to the SCOAP^3 HTTP endpoint and pull required information

```
public static HttpMethod callAPISCOAP3(String date,int jrec, int num_rec) {
    HttpMethod method =null;
    method=new GetMethod("http://repo.scoap3.org/search?of=xm&p=datecreated:"+date+"&jrec=" + jrec + "&rg=" + num_rec);
    return method;
```

Step 2

Get SCOAP^3 records

Save retrieved records in local XML file

```
public static void getRecordsScoap3(String date,int jrec, int num_rec) {
    String responseXML = null;
    HttpClient client = new HttpClient();
    HttpMethod method = callAPISCOAP3(date,jrec, num_rec);

    try {
        client.executeMethod(method);

        if (method.getStatusCode() == HttpStatus.SC_OK) {

            method.getResponseBody();
            responseXML = convertStreamToString(method.getResponseBodyAsStream());
            FileWriter fw = new FileWriter("MARCXML_SCOAP3_"+date+"/marcXML_scoap3_" + jrec + "_" + num_rec +

            fw.append(responseXML);

            fw.close();
```

Step 3

Format SCOAP^3 records

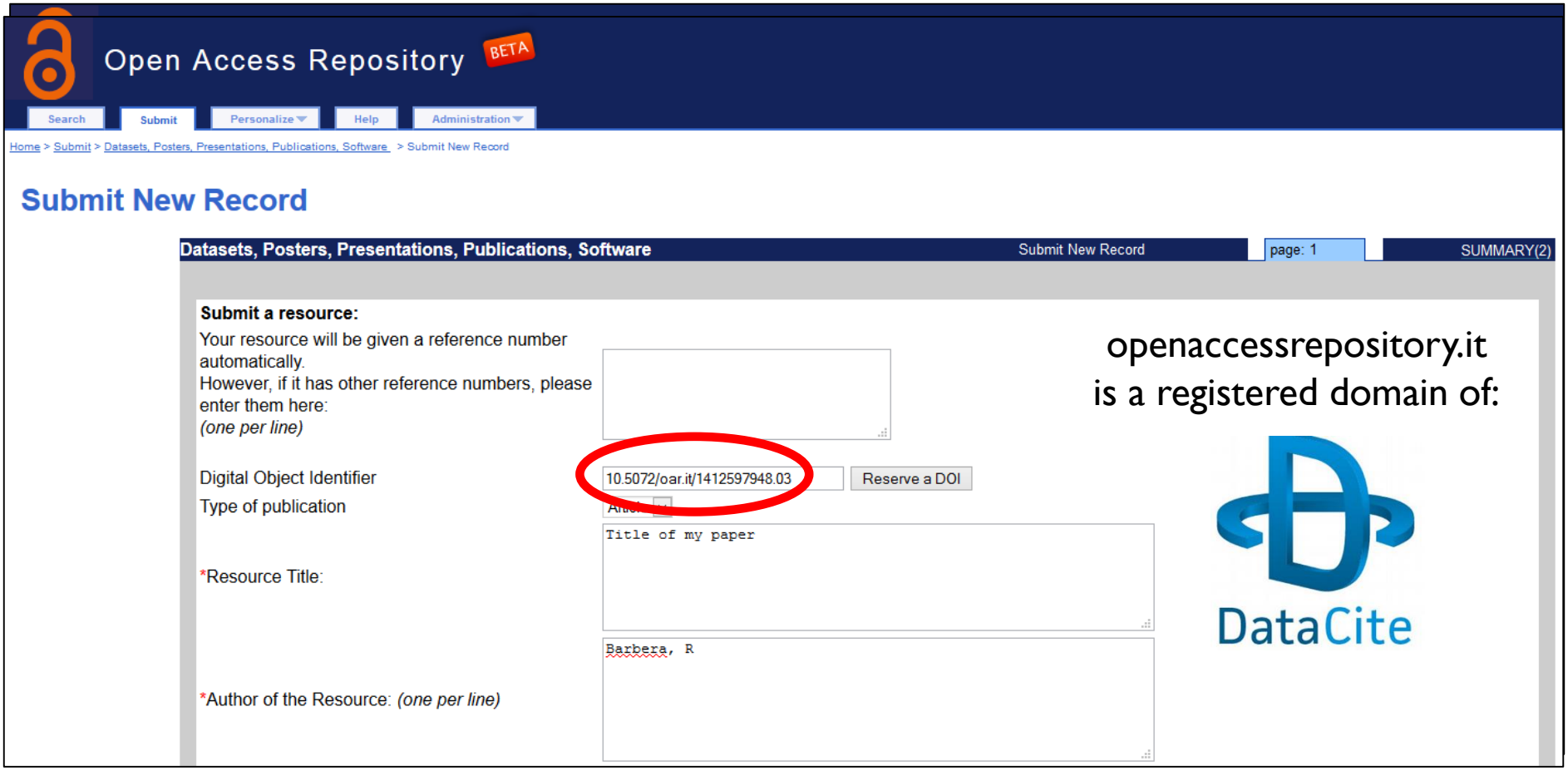
Create compliance with Open Access Repository scheme (modify MarcXML tags)

Step 4

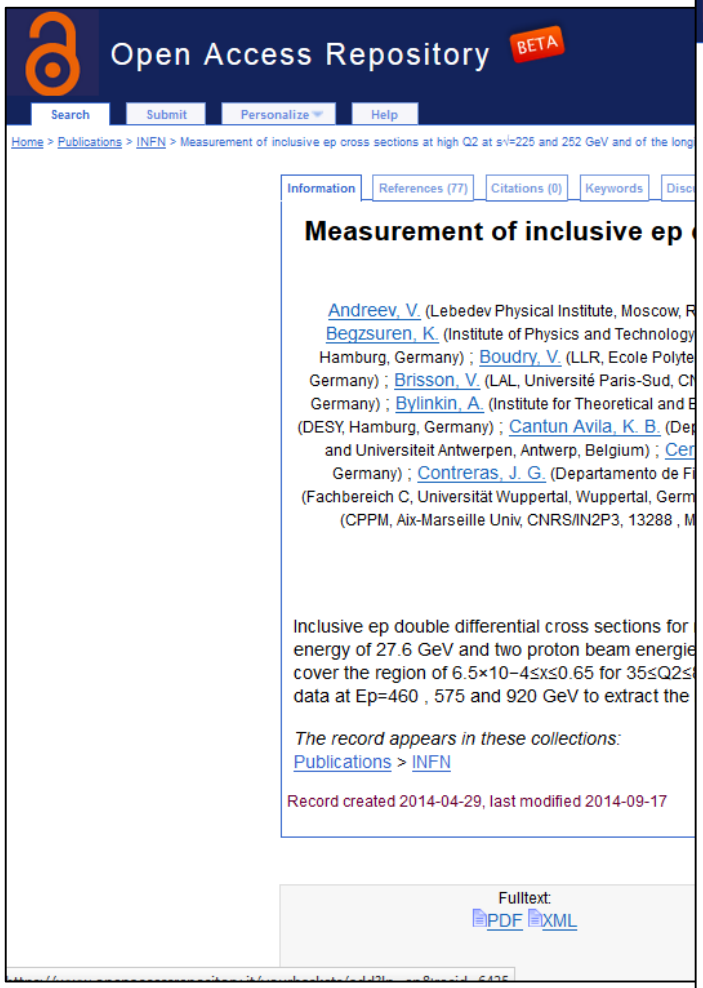
Ingest final data

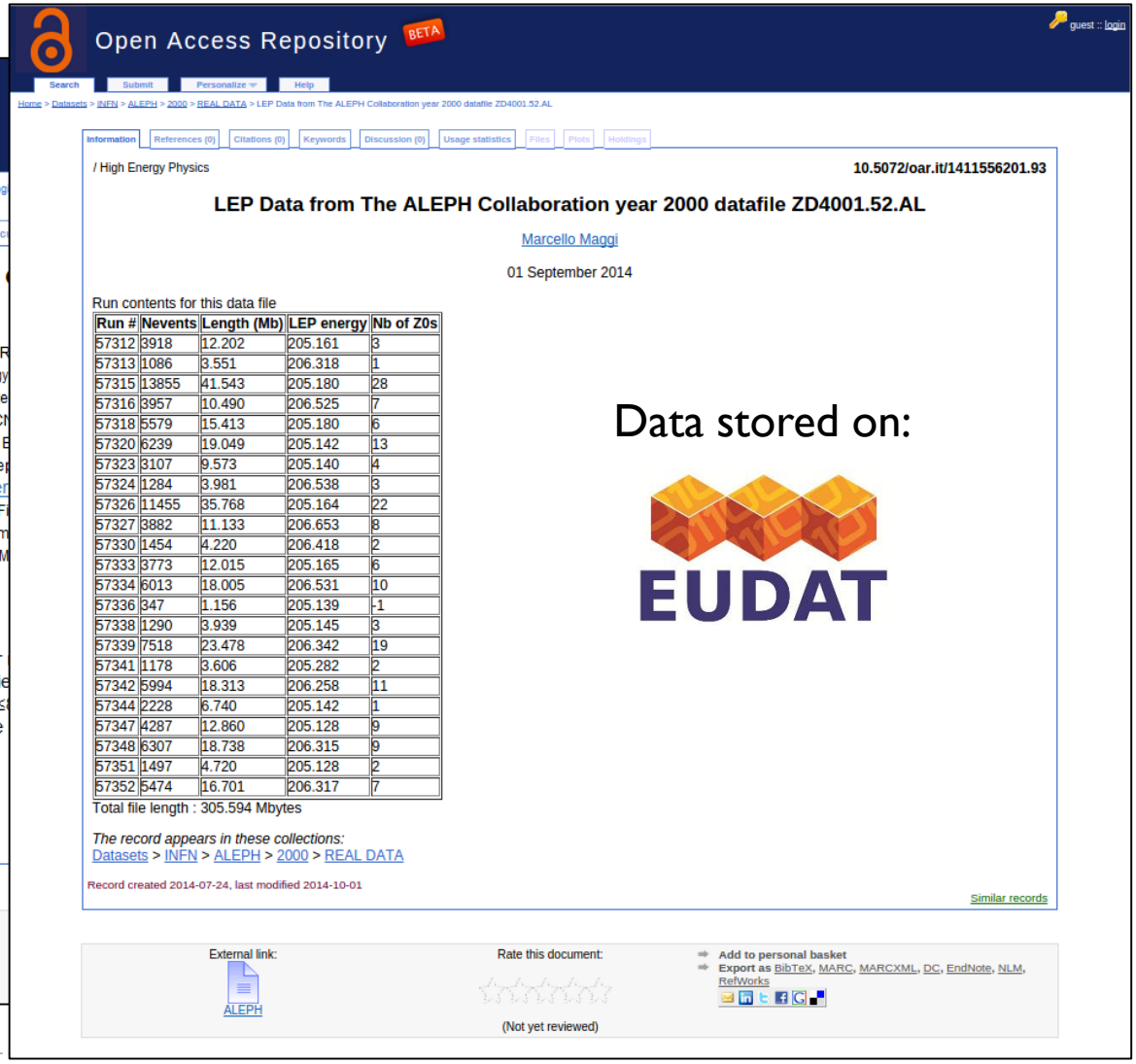
Ingest final, reformatted data in Open Access Repository

(Single) resources upload and DOI registration



Examples of document and data resources

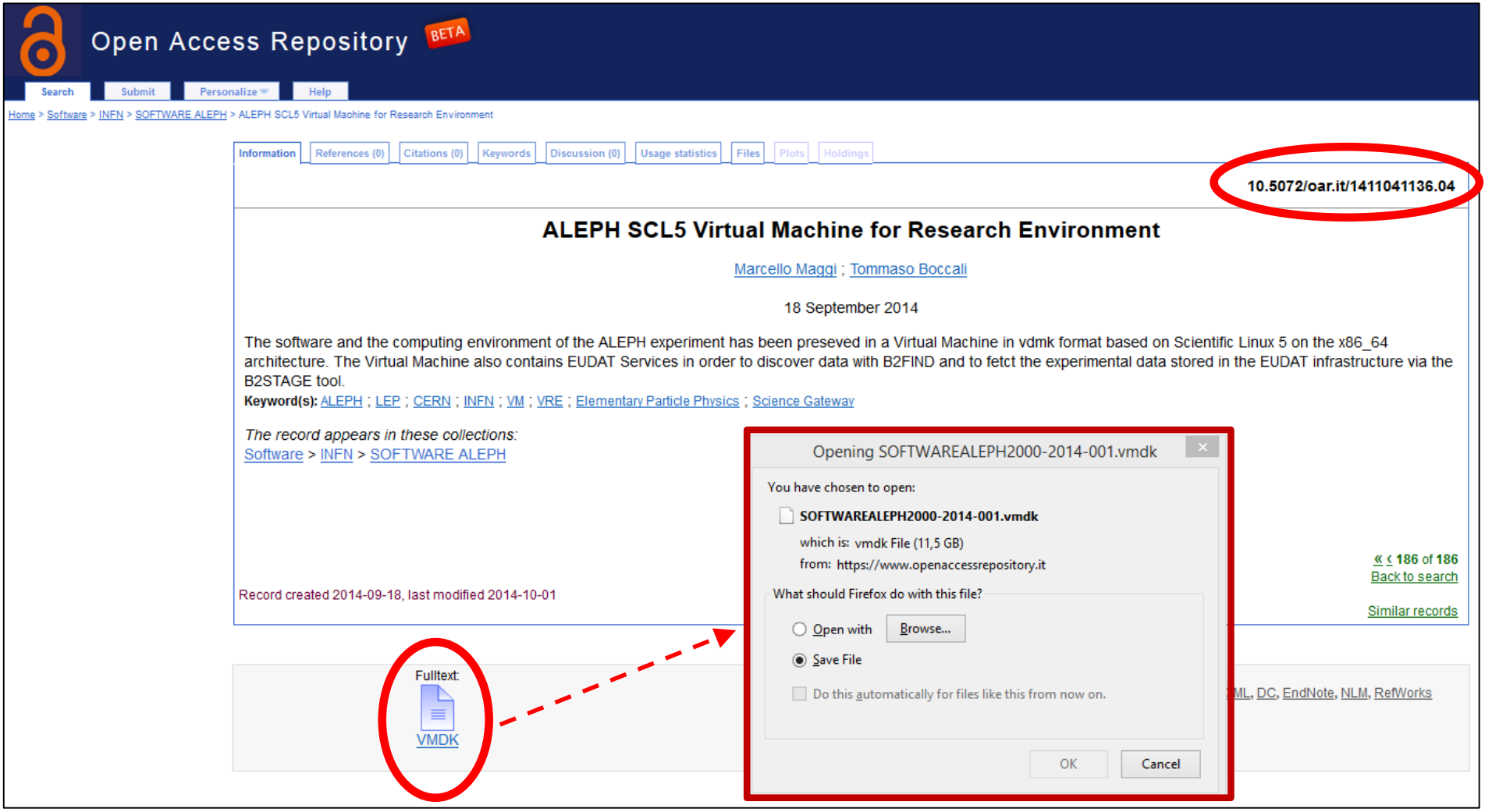




Data stored on:

EUDAT

Example of software resources:
the ALEPH Virtual Research Environment



Visibility and compliance
(full conforming with OAI)

Registration Record

element name	element value
Base URL	http://www.openaccessrepository.it/oai2d
Repository Name	Open Access Repository
Protocol Version	2.0
Email	librarian@openaccessrepository.it
Registration Date	2014-05-06T10:47:29Z
Date Last Validated	Tue May 6 10:47:29 2014
OAI Repository ID	www.openaccessrepository.it

If you are the maintainer of this repository, you may to update the information recorded to match new information exposed via the Identify response by running the validation/registration process again. Go to the validation page and select "Register this site".

Fri May 23 09:57:55 2014

Visibility and compliance (registered as an OpenDOAR data provider)

OpenDOAR

OpenDOAR Development Blog >

Directory of Open Access Repositories

Home | Find | Suggest | Tools | FAQ | About | Contact Us

Search or Browse for Repositories

Recent Additions RSS1 Feed

"Open Access Repository"

Any Subject Area | Any Content Type | Any Repository Type

Any Country | Any Language | Any Software | Search

Full records | 1 | per page. Sort by: Repository Name | New Query

To search the *contents* of the repositories listed in *Open*DOAR, please see our Content Search page.

Result **1** of **1**.

Page: << Previous 1 Next >>

Open Access Repository

URL: http://www.openaccessrepository.it/
Organisation: INFN
Address: Via Santa Sofia, 62, Catania
Country: Italy
Location: *Latitude*: 37.526700 & *Longitude*: 15.073100, Google Map
Description: Open Access repository of INFN publications and data, to be eventually extended to other organisations. The interface is in English.
Type: Institutional - Operational
Size: 2026 items (2014-05-19)
OAI-PMH: http://www.openaccessrepository.it/oai2d
Software: invenio
Subjects: Physics and Astronomy
Content: Articles; Conferences; Datasets; Multimedia; Software; Special
Languages: English
Contacts: 1. Roberto Barbera (roberto.barbera@ct.infn.it), Administrator
2. Roberto Barbera (librarian@openaccessrepository.it), Administrator
***Open*DOAR ID:** 3061, *Last reviewed*: 2014-05-16, Suggest an update for this record
Link to this record: http://opendoar.org/id/3061/

Visibility and compliance (compliant with OpenAIRE guidelines)

Browse Compatibility Test Results

Refresh (F5) to see the progress of the latest activities.

Guidelines 3.0

Type	Status	Score	User	Started	Duration	Repository	Rule Set	Reports
OAI Content Validation	finished	100	roberto.barbera@ct.infn.it	2014-05-22 15:53:18	5 mins 8 secs	http://www.openaccessrepository.it/oai2d	Custom	Show Results
OAI Usage Validation	finished	100	roberto.barbera@ct.infn.it	2014-05-22 15:53:18	1 secs	http://www.openaccessrepository.it/oai2d	Custom	Show Results

Re: Registration of a data provider — Repository Technical Support

23 May 2014, 08:51

Dear Roberto,

your repository seems to have successfully passed the registration test.

" A new repository/journal is ready to be added in the list of OpenAIRE compliant repositories/journals: Open Access Repository

Visibility and compliance (registered as an OpenAIRE data provider)

OpenAIRE

PARTICIPATE DEPOSIT, JOIN | SEARCH PUBLICATIONS, DATA, PROJECTS | STATISTICS OA, PROJECTS, TOPICS | SUPPORT FAQ, HELPDESK, GUIDES | OPEN ACCESS IN EUROPE

Ocean Science (OS)	-	-	-	OpenAIRE 2.0 (EC funding)
OceanRep	Germany	**20382** (updated 04/04/2014)	Helmholtz-Zentrum für Ozeanforschung Kiel	OpenAIRE 2.0+ (DRIVER OA, EC funding)
OMA	Belgium	**15057** (updated 20/02/2014)	Vlaams Instituut voor de Zee (Flanders Marine Institute)	OpenAIRE 2.0 (EC funding)
Online-Publikations-Server der Hochschule Osnabrück	Germany	**12** (updated 19/05/2014)	Hochschule Osnabrück	OpenAIRE basic (DRIVER OA)
Open Access LMU	Germany	**16316** (updated 13/02/2014)	Ludwig-Maximilians Universität München	OpenAIRE basic (DRIVER OA)
Open Access Repository	Italy	-	Istituto Nazionale di Fisica Nucleare (INFN)	OpenAIRE 3.0 (OA, funding)
Open Access Repository of Indian Theses	-	-	-	OpenAIRE basic (DRIVER OA)
Open Natuur Archief (Open Nature Archive)	Belgium	-	Research Institute for Nature and Forest	OpenAIRE basic (DRIVER OA)

Includes funding information: EU FP7, NOP-IT, INFN DB, <other CRISes>

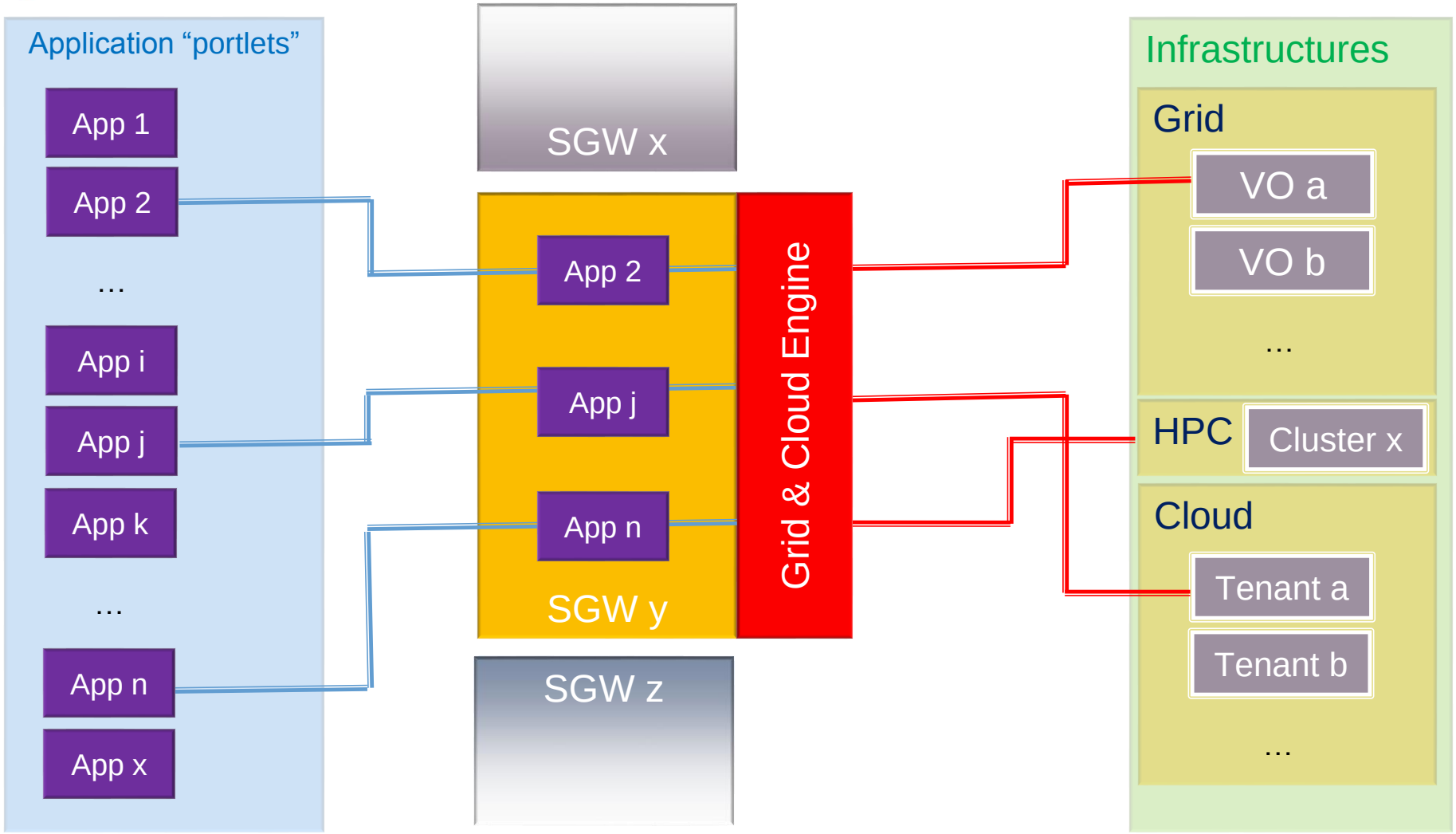
The Catania Science Gateway Framework (CSGF) in a nutshell

http://www.catania-science-gateways.it

To better access e-Infrastructures Worldwide

- Born in 2010 to hide Grid – and now Cloud - complexity (especially security-wise)
- Designed to be:
  - Sustainable (Fully based on standards)
  - Scalable (e.g., through Glassfish)
  - Secure (integrated AAAAI)
  - Interoperable (one system → many infrastructures)
  - Accessible anytime from anywhere (including mobile devices)

CSGF Architecture



CSGF Components

Users having different roles and privileges

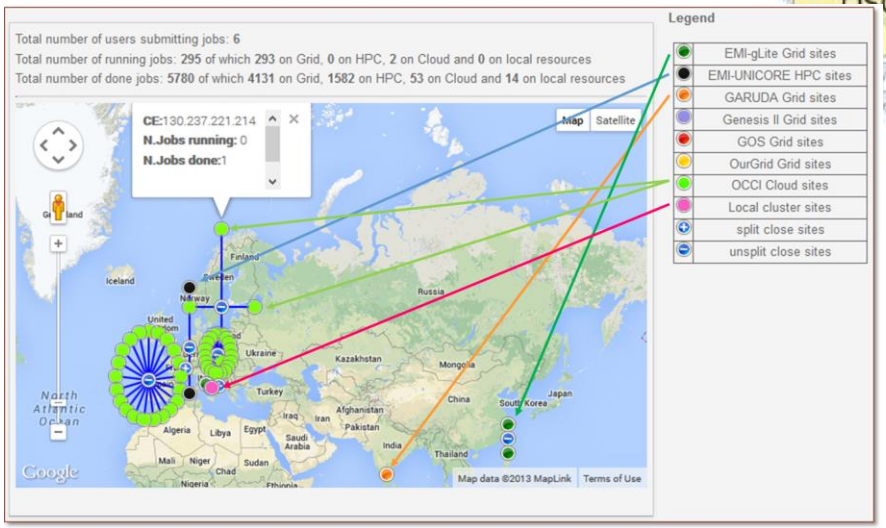
- Administrators
- Power users
- Basic users
- VRC members
- etc,

LIFERAY
Enterprise. Open Source. For Life.

AAI

Portlets

Grid&Cloud Engine



TeraGrid
XSEDE
OSG
EUMedGrid
EUChinaGrid
NAREGI
EUIndiaGrid
EUAsiaGrid
HP/UNESCO «Brain Gain» eI4Africa
NKN & Garuda
EELA
Grid
Cloud
HPC
GISELA
SAGrid & SANREN

European Commission co-funded projects

Projects with other funding

AAI in the CSGF

access.egi.eu

EGI UMP

OpenID

OpenID Connect

AuthN/AuthZ
handled separately

LIFERAY
Enterprise. Open Source. For Life.

Authentication

Shibboleth.

SAML 2.0

Authorization

OpenLDAP
http://www.OpenLDAP.org

Identity Federations

eduGAIN

idem garr aai

GRIDP

...

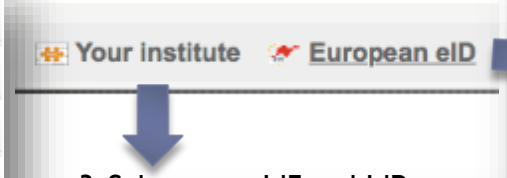
IdP 1

IdP 2

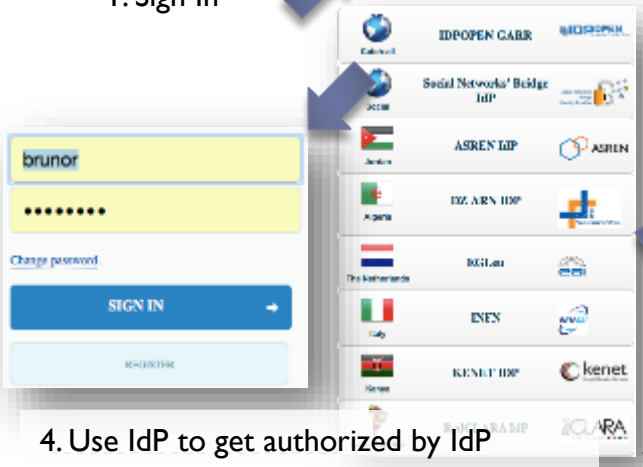
IdP n

1. Sign In

2. Select eID or Your institute IdFs



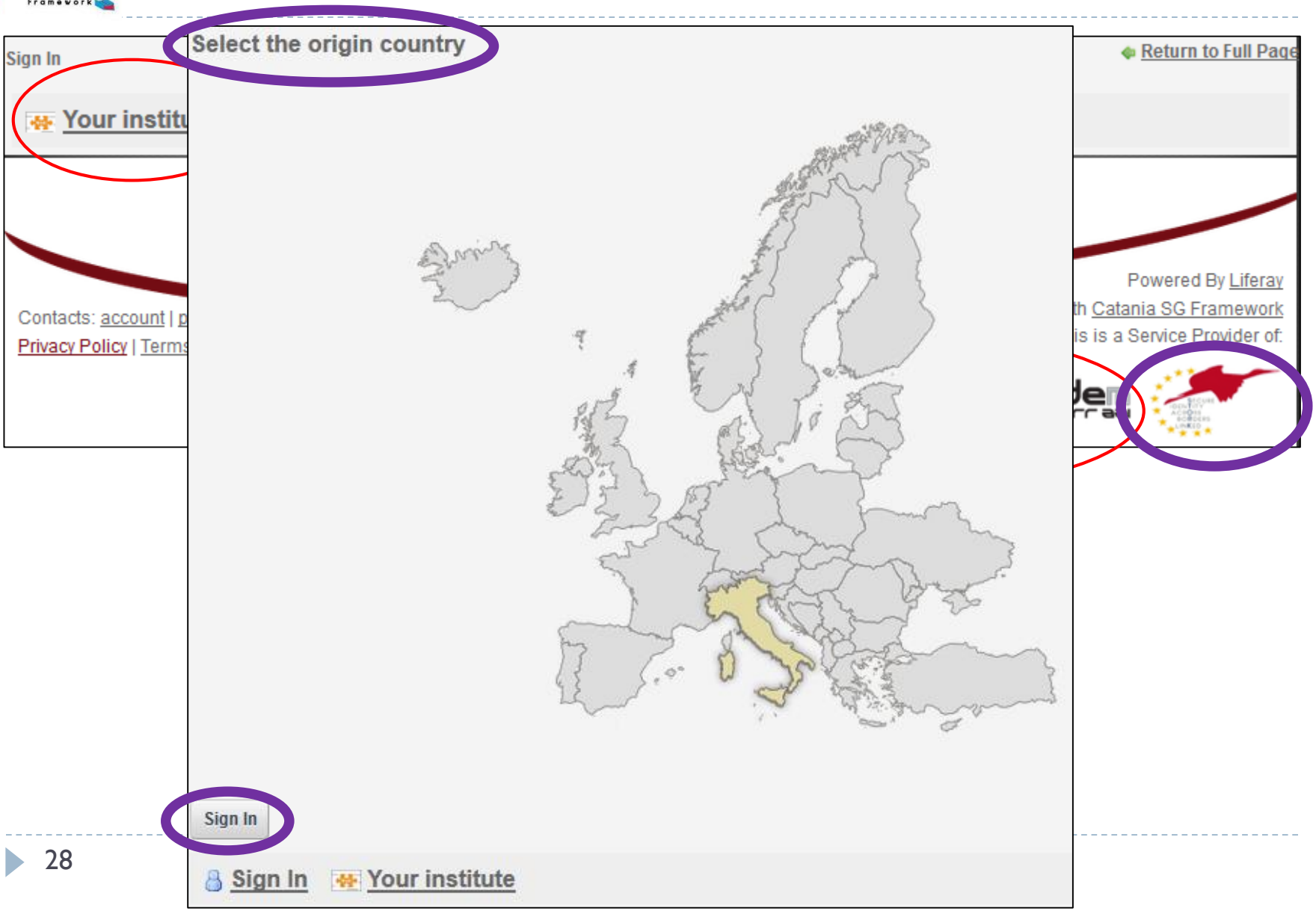
3. Select your IdF and IdP



4. Use IdP to get authorized by IdP

eid-stork.eu

Support for the most diverse and widest possible communities of users (1/3)



Support for the most diverse and widest possible communities of users (2/3)

Millions of people can become potential users

A new way of thinking about citizen science

Nordea

BankID

aNcert

SIGEN-CA

LUXTRUST

KAART

.beID

CATCert

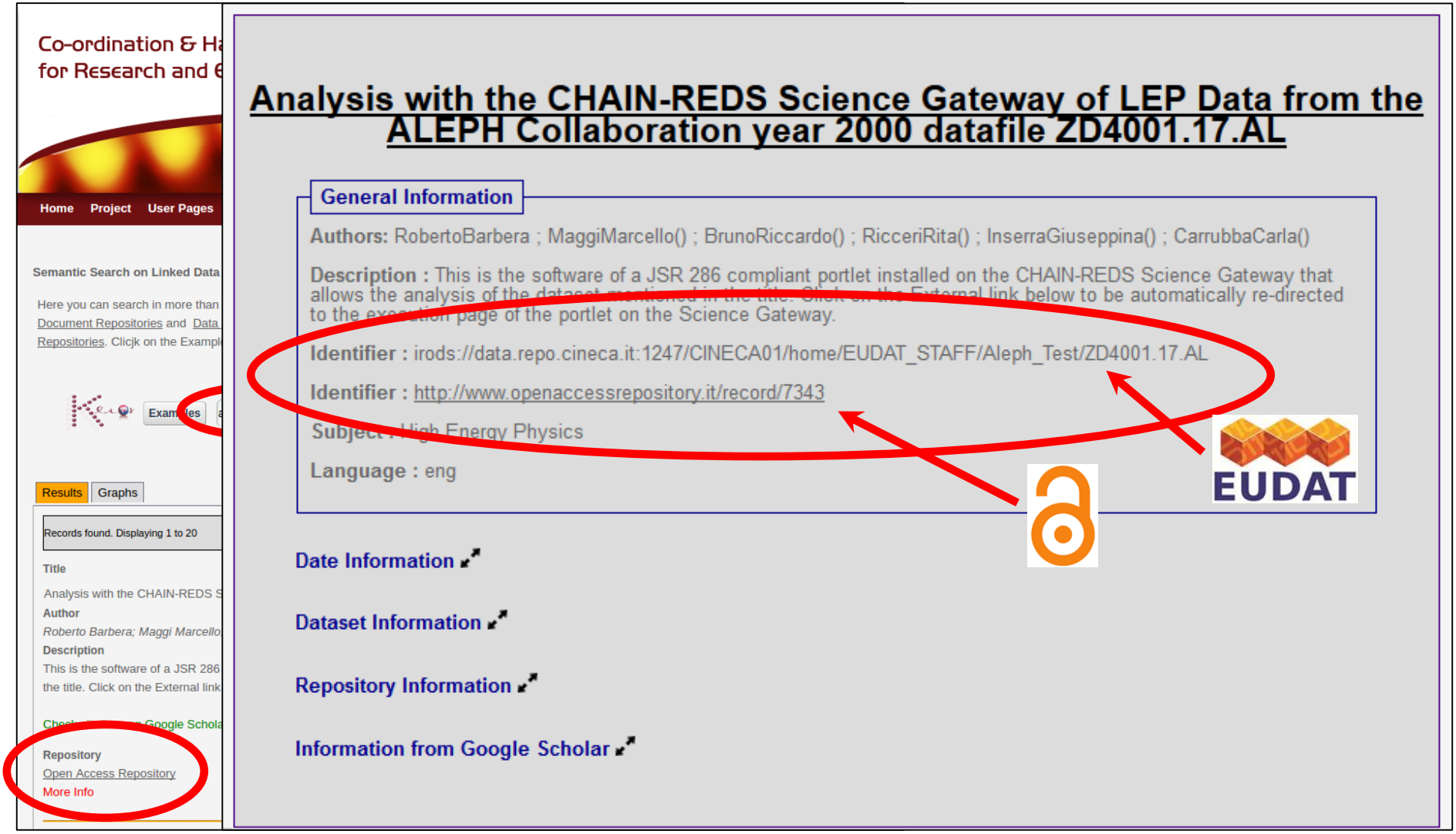
Agència Catalana de Certificació

Väestörekisterikeskus
Population Register Centre

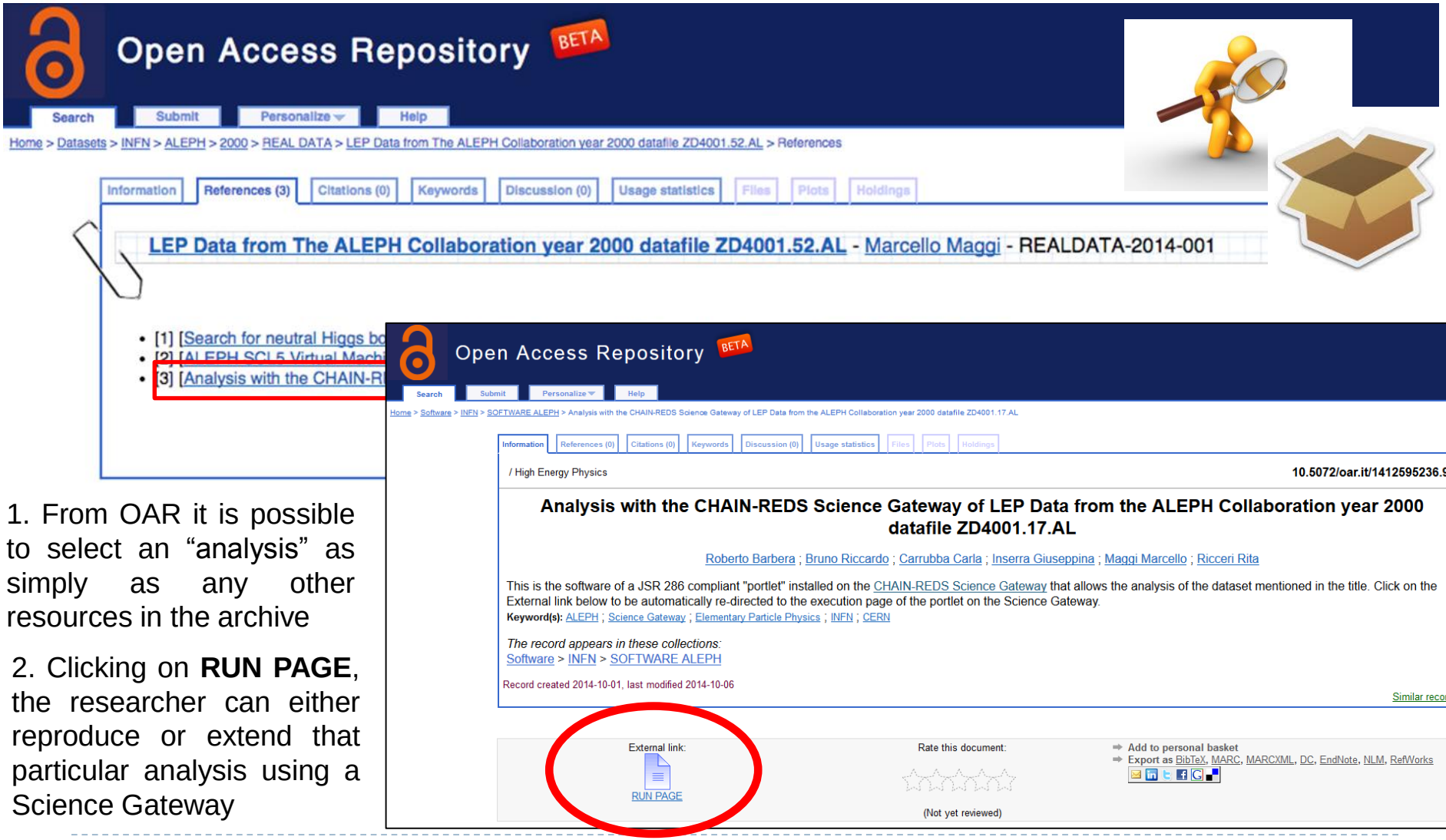
POŠTA CA
Certifikatska agencija Pošte Slover

The OAR Knowledge Workflow

The OAR Knowledge Workflow: data search & discovery



The OAR Knowledge Workflow: data inspection



1. From OAR it is possible to select an “analysis” as simply as any other resources in the archive

2. Clicking on **RUN PAGE**, the researcher can either reproduce or extend that particular analysis using a Science Gateway

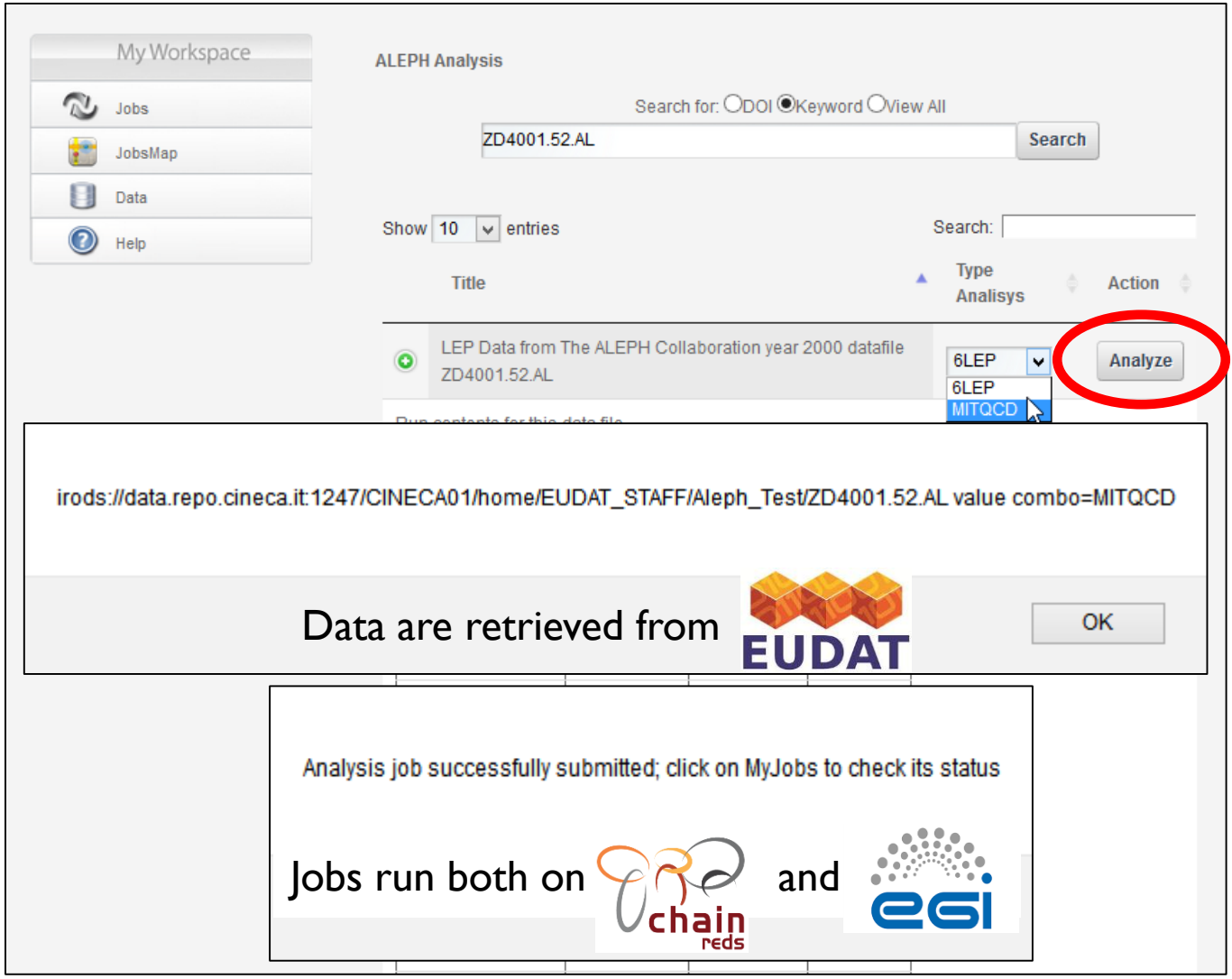
The OAR Knowledge Workflow: data analysis (1/2)

The Science Gateway collects from OAR, and allows user browse, metadata associated to the dataset(s) associated to that particular analysis

The OAR Knowledge Workflow: data analysis (2/2)

Using the JSAGA adaptor for all OCCI-compliant cloud-middleware, the Science Gateway starts a dedicated VM already configured with the all the experiment software

Both the CHAIN-REDS Cloud Testbed and the EGI Federated Cloud can be used as e-Infrastructures



Reproducibility of ALEPH data with the CHAIN-REDS Science Gateway (1/3)

Co-ordination & Harmonisation of Advanced e-Infrastructures for Research and Education Data Sharing

chain reds

Science Gateway

Welcome | Help | Applications | Job Status | My Workspace | My Cloud | Project Home

- "Hello World"
- Computational chemistry
- Computer Science and Mathematics
- Earth Sciences
- High Energy Physics
  - ALEPH Analysis
  - CORSIKA-LAGO
- Life Sciences
- Cloud Applications
- Data Applications
- Other
- Full List

By definition, a Scienc ...elopment set of tools, applications, and ...tal or a suite of applications, usually ... that is further customized to meet the ...

The present Science G... CHAIN and CHAIN-RE... Gateway paradigm and ...e-Infrastructures worldwide, based on di... ...res (Grid, HPC, Cloud or simply local ... each other, at user application level.

Thanks to the collabora... ...NEA-CRESCO, EUMEDGRID-Support, FutureGrid, GARUDA, GISELA, IGI, JSAGA, SAGrid and WeNMR projects, a set of Demo Applications has been deployed on various Grid (based on EMI - gLite and UNICORE - Genesis II, Globus, GOS, OurGrid middleware), Cloud (based on OCCI compliant - Okeanos, OpenNebula and OpenStack - stacks) and local (based on the Platform Computing resource manager) resources and you can execute them through this portal in a simple and easy way. Some of the Cloud sites belong to the EGI Federated Cloud.

Contributors

Follow us on Social Networks

This Includes the possibility to access the Science Gateway from within the Social Network page.

Reproducibility of ALEPH data with the CHAIN-REDS Science Gateway (2/3)

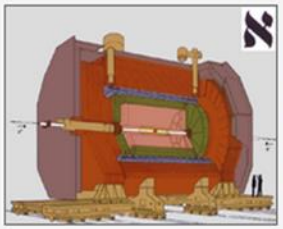
Co-ordination & Harmonisation of Advanced e-Infrastructures for Research and Education Data Sharing

chain reds

Science Gateway

Welcome Help Applications Job Status Project Home Sign In

ALEPH Analysis



ALEPH was a particle physics experiment installed on the Large Electron-Positron collider (LEP) at the CERN laboratory in Geneva/Switzerland. It was designed to explore the physics predicted by the Standard Model and to search for physics beyond it. ALEPH first measured events in LEP in July 1989. LEP operated at around 91 GeV – the predicted optimum energy for the formation of the Z particle. From 1995 to 2000 the accelerator operated at energies up to 200 GeV, above the threshold for producing pairs of W particles. The data taken, consisted of millions of events recorded by the ALEPH detector,allowed precision tests of the electro-weak Standard Model (SM) to be undertaken. The group here concentrated our analysis efforts mainly in Heavy Flavour (beauty and charm) physics, in searches for the the Higgs boson, the particles postulated to generate particle mass, and for physics beyond the SM, e.g. Supersymmetry, and in W physics.

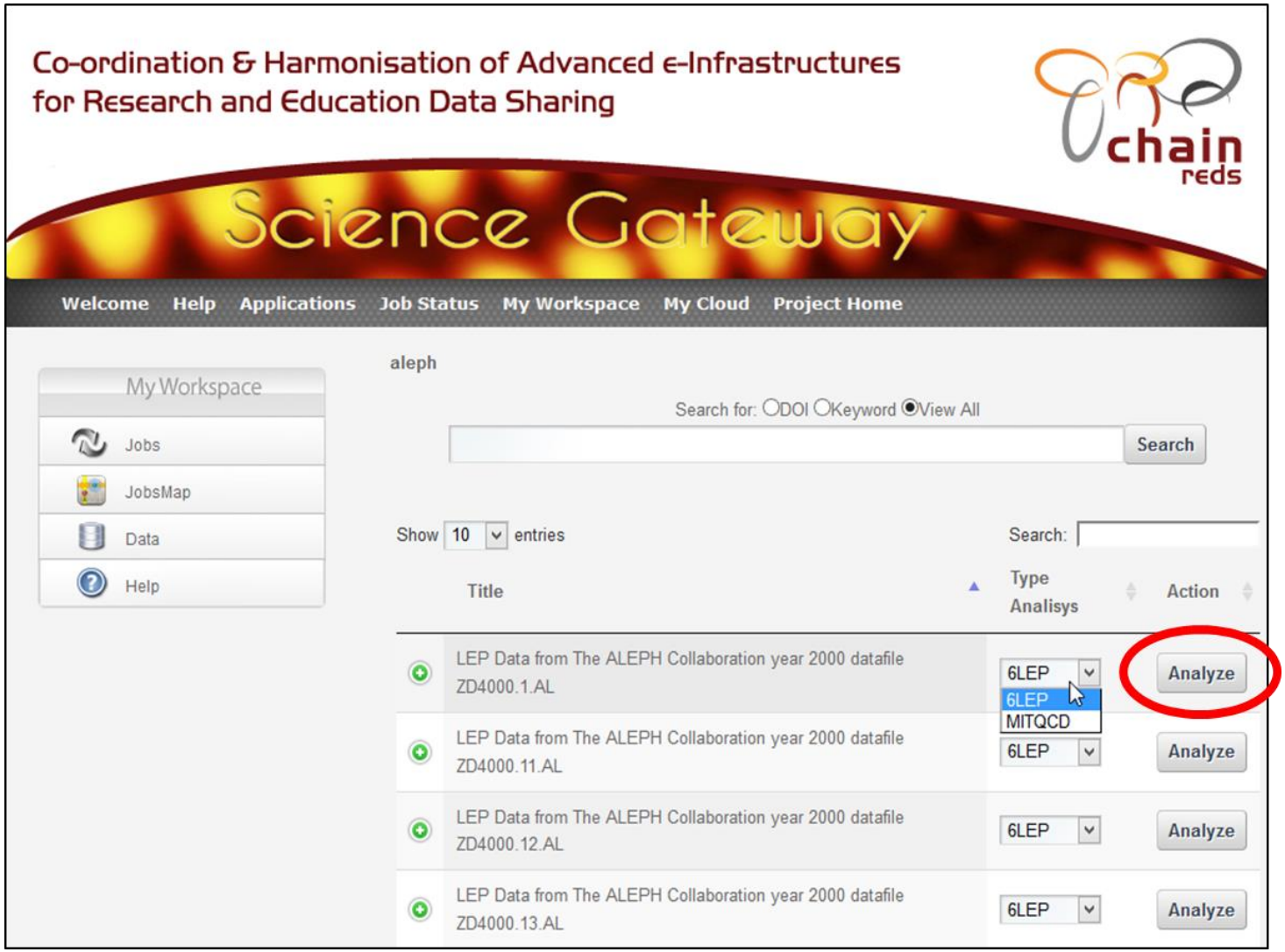
This application perform the search for the production and non-standard decay of a scalar Higgs boson into four tau leptons through the intermediation of the neutral pseudo-scalars Higgs particle.

The analysis was conducted by the ALEPH collaboration with the data collected at centre-of-mass energies from 183 to 209 GeV.

Sign-in to RUN

Results are published in JHEP 1005 (2010) 049 DOI: 10.1007/JHEP05(2010)049

Reproducibility of ALEPH data with the CHAIN-REDS Science Gateway (3/3)

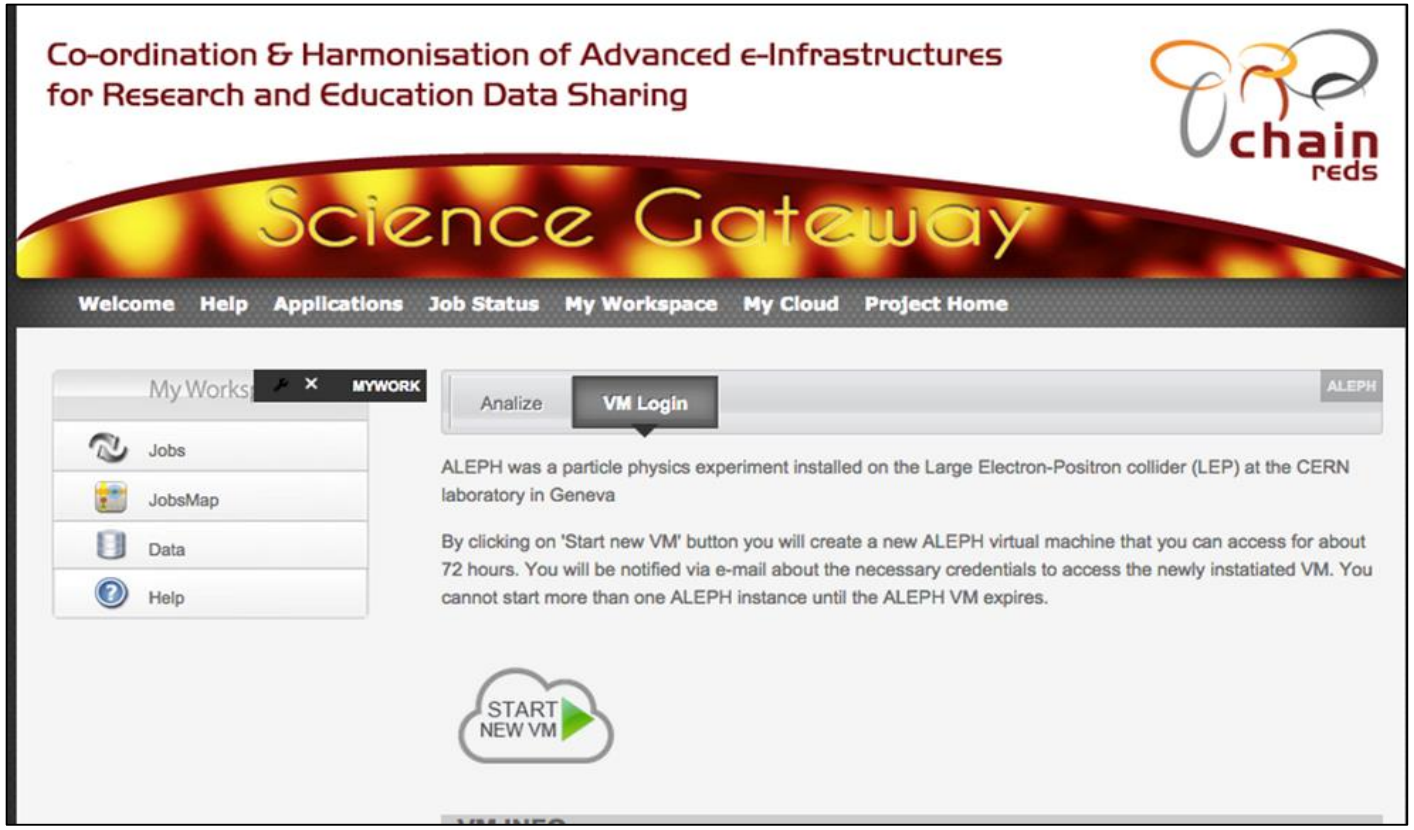


Remember: repeatability and reproducibility are not all

Reusability and «extensibility» matter!

Reusability of ALEPH data with the CHAIN-REDS Science Gateway (1/2)

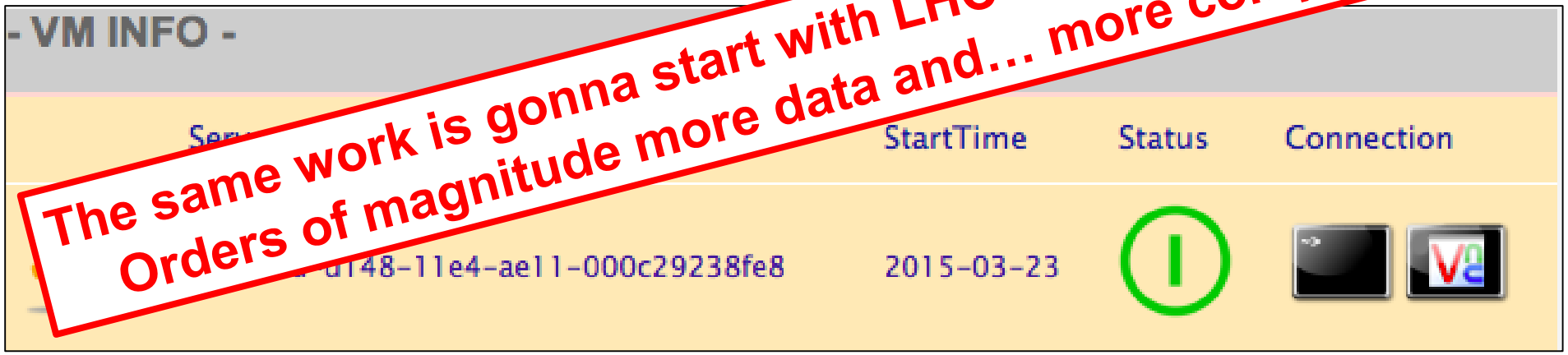
1. From within the CHAIN-REDS Science Gateway an entitled researcher can start a VM already configured to re-use/extend a given scientific analysis
2. The VM was previously distributed both on the CHAIN-REDS Cloud Testbed and on the EGI Federated Cloud using the EGI AppDB



Reusability of ALEPH data with the CHAIN-REDS Science Gateway (2/2)

1. The VM available tor a customizable amount of time during which the user has full access to the dataset(s) and analysis algorithm(s) of the experiment
2. The user receives via email the credentials to access the VM using di[illegible] protocols (e.g., SSH, VNC, etc.). Clicking on the SSH or VNC i[illegible] directly access the cloud-located VM from within the [illegible]

The same work is gonna start with LHC's ALICE Experiment
Orders of magnitude more data and… more complexity



New stable analyses (and their results), generated running the VM, may be registered in the OAR (with a DOI) to further extend the analysis catalogue shared across the Virtual Research Community

Summary and conclusions

- Open Science vision can be implemented only if the "openness" paradigm becomes pervasive in research
- Science outputs' reproducibility, but also re-usability and extensibility, are key to walk through the "knowledge path" in both directions
- The INFN Open Access Repository is a pilot data preservation repository of science products meant to serve both researchers and citizen scientists; what makes OAR different from other repositories is its capability to connect to Science Gateways and exploit cloud resources worldwide to easily reproduce/extend scientific analyses
- The feasibility of the OAR Knowledge Workflow has successfully been tested with ALEPH datasets and will now benchmarked with ALICE ones
- For the new tests we plan to explore CERNVM as well as Containers + Docker

Thank you !

Co-authors

- S. Bianco (INFN LNF)
- T. Boccali (INFN Pisa)
- R. Bruno (INFN Catania)
- C. Carrubba (INFN Catania)
- G. Inserra (INFN Catania)
- M. Maggi (INFN Bari)
- D. Menasce (INFN Milano Bicocca)
- R. Ricceri (INFN Catania)