

Standardizing life sciences datasets to improve reproducibility in the EOSC

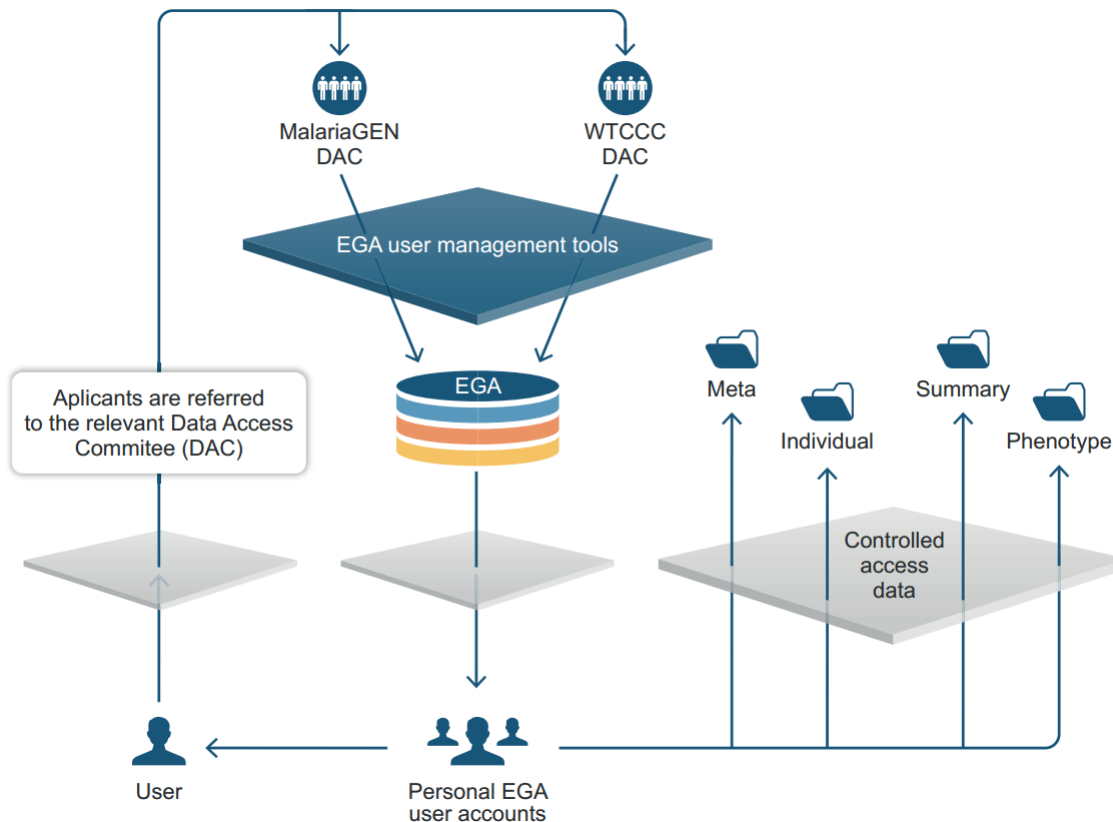
Jordi Rambla

September, 15th 2017



What is the EGA?

The EGA is a resource for permanent secure archiving and sharing of all types of potentially identifiable genetic and phenotypic data resulting from biomedical research projects.



Data is provided by research centers and health care institutions.

Access is controlled by Data Access Committees.

Data requesters are researchers from other research or health care institutions.

<https://ega-archive.org>

Project goal

The EGA was created by the EBI, in 2007, as an extension of the ENA...

Project goal:

To transform the EGA to a joint project (*in the context of ELIXIR Europe*) to have a real impact in the development of personalized medicine



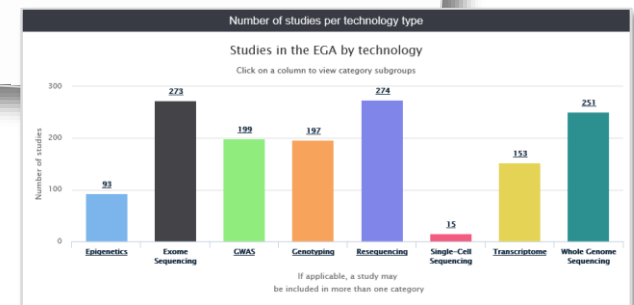
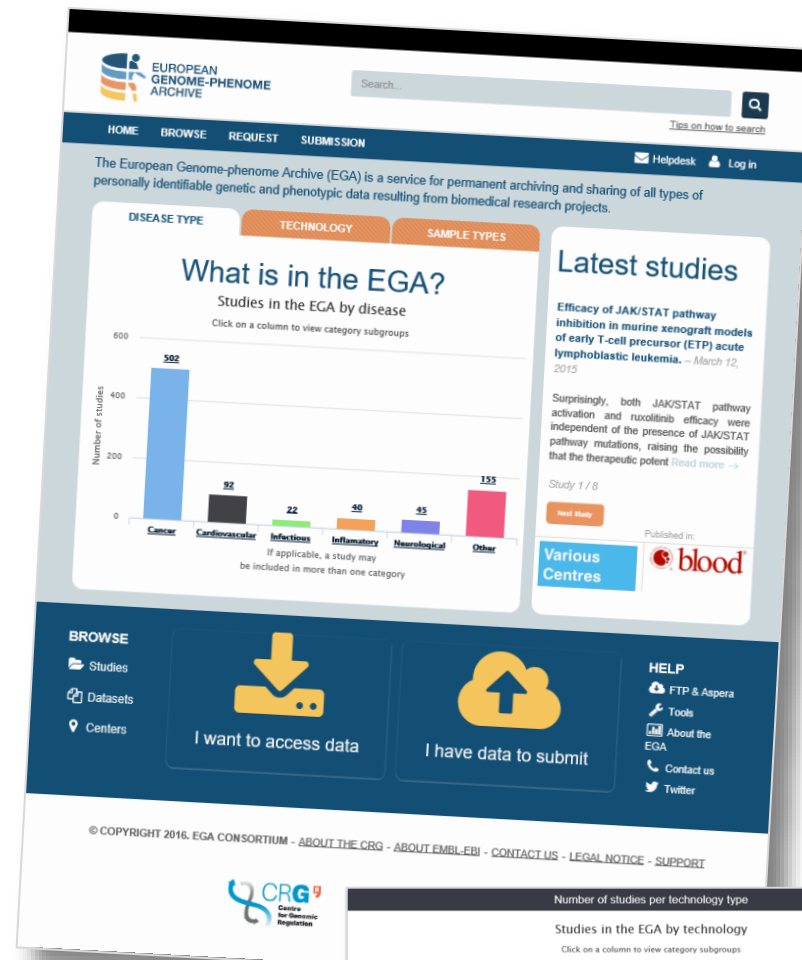
The EGA contains a variety of data

The EGA in numbers

- > 1,300 Studies
- 3,400 Datasets
- >800 Data providers
- >9,000 Data Requesters

The EGA in Volume

- >4 Petabytes



* Updated Sept, 8th 2017

The EGA is part of many international projects



EOSC and EOSCpilot



EOSCpilot
The European Open Science
Cloud for Research Pilot Project
www.eoscpiot.eu



**Science & Technology
Facilities Council**

The European Open Science Cloud Pilot

Brian Matthews

Science and Technology Facilities Council



EOSCpilot
The European Open Science
Cloud for Research Pilot Project
www.eoscpiot.eu



EOSCpilot
The European Open Science
Cloud for Research Pilot Project
www.eoscpilot.eu



EUROPEAN
COMMISSION

Brussels, 19.4.2016
COM(2016) 178 final

**COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN
PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL
COMMITTEE AND THE COMMITTEE OF THE REGIONS**

**European Cloud Initiative - Building a competitive data and knowledge economy in
Europe**

{SWD(2016) 106 final}
{SWD(2016) 107 final}

Why Europe is not fully tapping into the potential of data:

- 🔗 **Data not always open** and **lack of incentives and rewards** for data sharing
- 🔗 **Lack of interoperability** required for data sharing ... noting deep-rooted walls between disciplines.
- 🔗 **Fragmentation between data infrastructures** that are split by scientific and economic domains, countries and governance models
- 🔗 Surging demand for **High Performance Computing** at a scale above single member state resources
- 🔗 **Data reuse employing advance analysis techniques** adequate protection of personal data considering forthcoming revision of Copyright legislation.



Proposed a European Open Science Cloud

- Make all **scientific data** produced by the Horizon 2020 programme **open by default**.
- Raise awareness and **change incentive structures** for academics industry and public services to share their data.
- Develop **specification for interoperability** and data sharing across disciplines and infrastructures
- Create a fit-for-purpose **pan-European governance structure** to federate scientific data infrastructures and overcome fragmentation.
- **Develop cloud based services** for Open science **supported by** the necessary **data infrastructure**
- **Enlarge the scientific user base** to researchers and innovators from all disciplines.



Final draft

Monday 20 June 2016

A Cloud on the 2020 Horizon

Commission High Level Expert Group on the European Open Science Cloud

Realising the European Open Science Cloud: first report and recommendations

Preface by Barend Mons, Chair





This report aims to lay out a high level, living roadmap for the realisation of the European Open Science Cloud (EOSC). The High Level Expert Group, with ten members from European countries, Japan and Australia, discussed extensively in several meetings, conferences, policy events and met with key stakeholders (30 November 2015) and research funders (15 March 2016). Based on these consultations, on many 'white papers' and on a range of presentations and feed-back at international meetings, we are confident that our recommendations count on a high-level of consensus amongst all stakeholders. This was a solid basis to embark on this challenging journey with the Commission, the Member States and International partners in concert.

The title of this first report may have a slightly threatening ring to it and indeed, if we do not act, there might be a looming crisis on the Horizon. The vast majority of all data in the world (in fact up to 90%) has been generated in the last two years. Computers have long surpassed individuals in their ability to perform pattern recognition over large data sets. Scientific data is in dire need of openness, better handling, careful management, machine actionability and sheer re-use. One of the sobering conclusions of our consultations was that research infrastructure and communication appear to be stuck in the 20th century paradigm of data scarcity. We should see this step-change in science as an enormous opportunity






Definitions




European:

-  research and innovation are global - EOSC cannot be built exclusively in and for Europe
-  Europe, is in a strong position to lead this initiative as already distributed and collaborative



Open:

-  not all data and tools can be open. E.g. confidentiality and privacy.
-  Open is also often confused with 'for free'. Free data and services do not exist.
-  Intelligently open is what we mean,

Science:

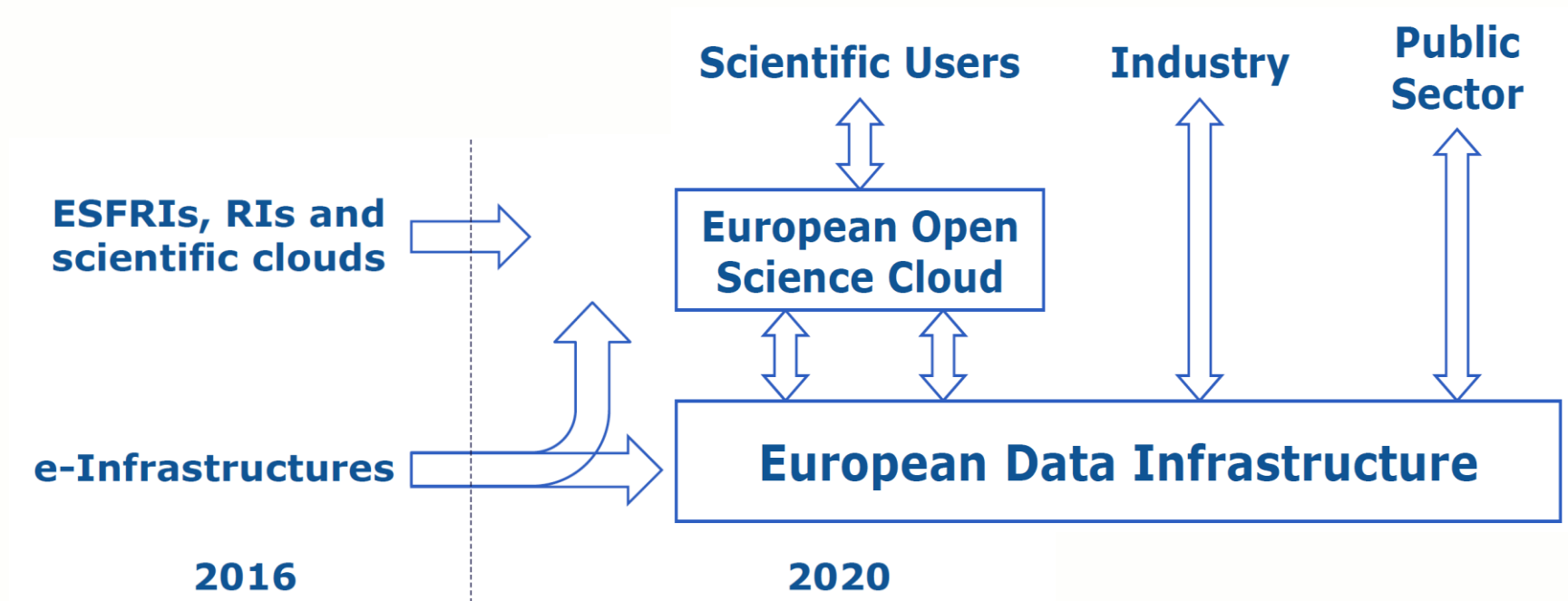
-  explicitly includes all disciplines including the arts and humanities,
-  Also societal innovation and productivity,
-  support broad societal participation in Open Innovation and Open Science.

Cloud:

-  It can be misinterpreted to indicate that the EOSC is mostly about hard ICT infrastructure
-  But it is much more a **commons of data, software, standards, expertise and policy related to data-driven science and innovation.**



Evolution of infrastructure





*EOSC*pilot: High Level Aims

The *EOSC*pilot project will support the first phase in the development of the EOSC.

- **Establish the governance framework** for the EOSC and contribute to the development of European open science policy and best practice;
- **Develop a number of demonstrators** functioning as high-profile pilots that integrate services and infrastructures to show interoperability and its benefits in a number of scientific domains;
- **Engage with a broad range of stakeholders**, crossing borders and communities, to build the trust and skills required for adoption of an open approach to scientific research.



Science Demonstrators

First 5 Demonstrators

- **Environmental & Earth Sciences** - ENVRI
Radiative Forcing Integration to enable harmonised data access and integration across multiple research communities
- **High Energy Physics** - WLCG: large-scale, long-term preservation and re-use of HEP data in the EOSC open to other researchers
- **Humanities** – TEXTCROWD: Collaborative semantic enrichment of text-based datasets by make new software available on the EOSC.
- **Life Sciences** - Pan-Cancer Analyses & Cloud Computing within the EOSC to accelerate genomic analysis on the EOSC
- **Physics** - The photon-neutron community to improve the community's computing facilities by creating a virtual platform for all users

Second 5 Demonstrators

- **HPCaaS for Fusion** - Culham Science Centre, UK
 - **Life Science Leveraging EOSC** to offload updating and standardizing life sciences datasets and to improve studies reproducibility, reusability and interoperability-CRG, Spain
 - **Seismology**: EPOS Virtual Earthquake and Computational Earth Science e-science environment in Europe- University of Liverpool, UK
 - **CryoEM** Linking distributed data and data analysis resources as workflows in Structural Biology with cryo-Electron Microscopy: Interoperability and reuse CSIC, Spain
 - **Astronomy Open Science Cloud** access to LOFAR data - ASTRON, NL
- 5 more demonstrators to be selected in the autumn.

the EGA EOSCpilot project

Reproducibility crisis

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Volume 496 | Issue 7446 | Editorial | Article

Cornell University Library

arXiv.org > q-bio > arXiv:1508.06715

NATURE | EDITORIAL

Journal List > Bioinformatics > PMC3810853

Announcement: I

24 April 2013

[PDF](#) [Rights & Permissions](#)

Over the past year, *Nature* has and reproducibility of published research. The problems arise in laboratories that do not exert sufficient scrutiny over their information for other research.

Assessing the validity and reproducibility of validation of adenosine deaminase acting on RNA (ADAR)-mediated editing

Lauren A. Sugden,¹ Michael R. Tackett,² Yiannis A. Savitski

[Author information](#) [Article notes](#) [Copyright and License information](#)

Abstract

Motivation: Validation and reproducibility of results from recent embarrassing incidents involving the irreproducibility of this issue and the need for rigorous methods.

Results: Here, we describe an existing statistical method and its utility for assessing the reproducibility of validation of adenosine deaminase acting on RNA (ADAR)-mediated editing, a statistical method for planning validation experiments with confidence limits, which, for a fixed total number of experiments, for the study.

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive | Volume 515 | Issue 7525 | Editorial | Article

NATURE | EDITORIAL

Journals unite for reproducibility

Consensus on reporting principles aims to improve quality control in biomedical research and encourage public trust in science.

05 November 2014

[PDF](#) [Rights & Permissions](#)

Reproducibility, rigour, transparency and independent verification are cornerstones of the scientific method. Of course, just because a result is reproducible does not make it right, and just because it is not reproducible does not make it wrong. A transparent and rigorous approach, however, will almost always shine a light on issues of reproducibility. This light ensures that science moves forward, through independent verifications as well as the course corrections that come from refutations and the objective examination of the resulting data.

The EGA EOSCpilot project: GOALS

1. Make easier to reproduce results archived at EGA
2. Avoid repeated reprocessing of the data with modern tools
3. Make artifacts involved easier to discover (FAIR)

Results reproducibility

- EGA stores both raw and secondary analysis data
- We will like to make very simple to get the published/archived from the raw data
 - Given the reproducibility crisis, ensuring exactitude is very desirable
 - Link data to the pipelines and tools used to analyze them
- Pipeline and tool repositories using stable identifiers are required

Remastered results

- Once raw data is downloaded many users will up to date them by processing against current references and using popular pipelines
 - This means tons of wasted resources to get the same results: *human, computational and time resources*
- We would like to generate reproducible pipelines, run them and get the results back to the EGA
 - Thus users could choose to get the originals, the remastered or both
- We need to actually check the popularity of such “service”
 - Maybe we just need to leverage work done by previous users

FAIR glimpse

- **F**indable, **A**ccessible, **I**nteroperable & **R**e-usable
 - There is more in these words that it looks like
 - Each one is doing its own interpretation

Implementing the FAIR Principles

- **To be Findable:**
 - F1. (meta)data are assigned a globally unique and eternally persistent identifier.
 - F2. data are described with rich metadata.
 - F3. (meta)data are registered or indexed in a searchable resource.
 - F4. metadata specify the data identifier.
- **To be Accessible:**
 - A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
 - A1.1 the protocol is open, free, and universally implementable.
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
 - A2 metadata are accessible, even when the data are no longer available.
- **To be Interoperable:**
 - 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.
- **To be Re-usable:**
 - R1. meta(data) have a plurality of accurate and relevant attributes.
 - R1.1. (meta)data are released with a clear and accessible data usage license.
 - R1.2. (meta)data are associated with their provenance.
 - R1.3. (meta)data meet domain-relevant community standards.

15 Criteria

Findable (defined by metadata (PID included) and documentation)

1. No PID nor metadata/documentation
2. PID without or with insufficient metadata
3. Sufficient/limited metadata without PID
4. PID with sufficient metadata
5. Extensive metadata and rich additional documentation available



Accessible (defined by presence of user license)

1. Metadata nor data are accessible
2. Metadata are accessible but data is not accessible (no clear terms of reuse in license)
3. User restrictions apply (i.e. privacy, commercial interests, embargo period)
4. Public access (after registration)
5. Open access unrestricted

Interoperable (defined by data format)

1. Proprietary (privately owned), non-open format data
2. Proprietary format, accepted by Certified Trustworthy Data Repository
3. Non-proprietary, open format = 'preferred format'
4. As well as in the preferred format, data is standardised using a standard vocabulary format (for the research field to which the data pertain)
5. Data additionally linked to other data to provide context

Make data more discoverable

- EGA is already honoring some FAIR principles
 - Findable, Accessible (\pm), Interoperable (\pm), Re-usable
- As we expand the number of artifacts related to the data archived at EGA, we are increasing the need to describe and link such objects
- We would like to leverage the process of generating the previously described artifacts to gather metadata that would be exposed through the right tools and services.



THANKS!

Core organizations:



Additional sources:



And infrastructure support from the following sources:

