Open Science and Research Data Management - Challenges and Perspectives

Claudia Bauzer Medeiros
Institute of Computing – Unicamp

ADVICE(S) – 3 STORIES

ROCKS, THEOREMS AND PIGS, ONE FISH

ADVICE(S)

- OPEN YOUR SENSES (TO “DATA”)
- TALK TO ENTHUSIASTIC SCIENTISTS
- COST VS VALUE
- RESEARCH

= **WHAT IF**

= PERTURBATION ON STATUS QUO

= one object, countless opportunities

Open Science and Research Data Management - Challenges and Perspectives

Claudia Bauzer Medeiros
Institute of Computing – Unicamp

Main take-aways

- What is open science?
- What does OPEN mean?
- What is data?
- Why should I care?

Main take-aways

- What is open science?
- What does OPEN mean?
- What is data?
- Why should I care?

- DATA = concentrate of everything!!!

Outline

Open Science

Definitions

Challenges

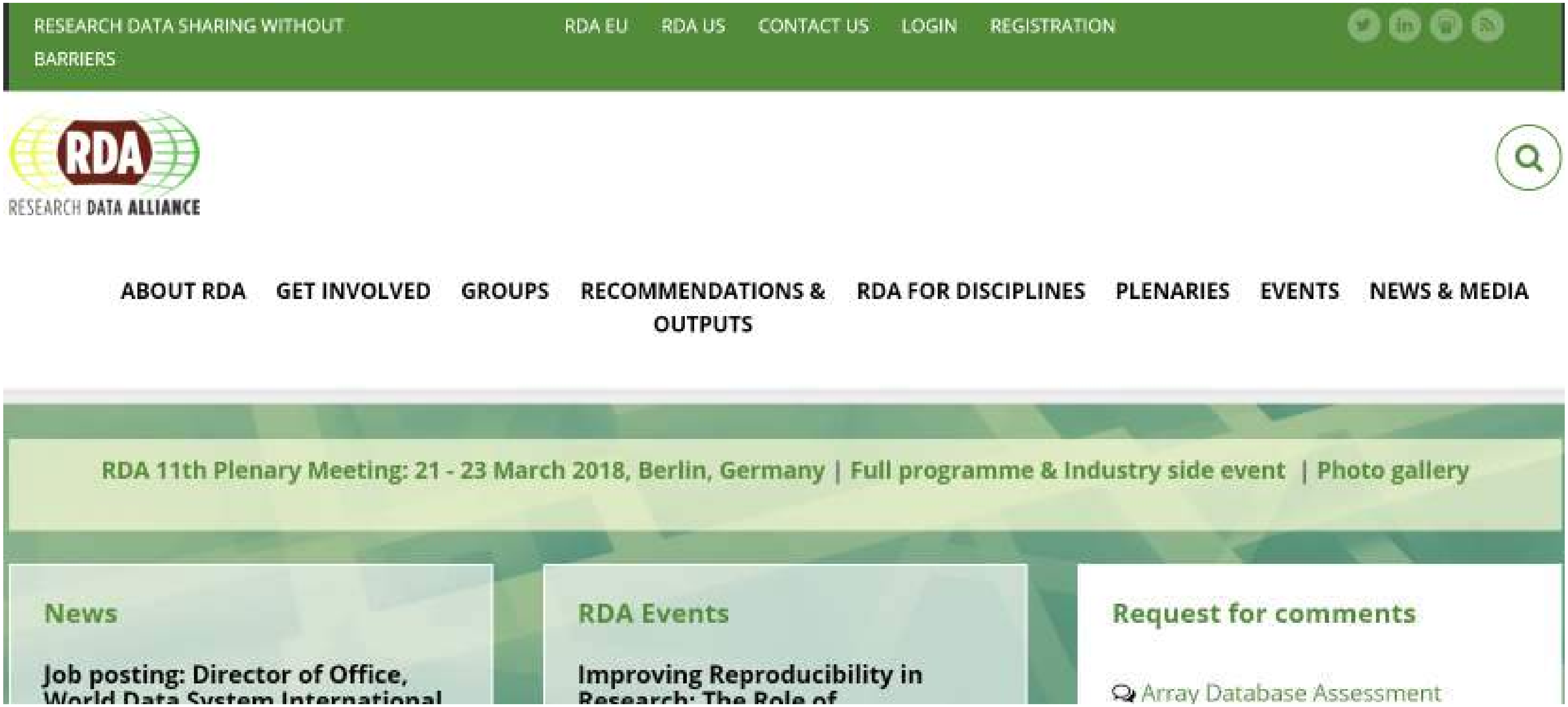
4 Challenges in (Research Data) Sharing

Open Data at FAPESP

Data Management Plans

Network of open research data repositories

RDA - https://www.rd-alliance.org



RESEARCH DATA SHARING WITHOUT BARRIERS

RDA EU | RDA US | CONTACT US | LOGIN | REGISTRATION

RESEARCH DATA ALLIANCE

ABOUT RDA | GET INVOLVED | GROUPS | RECOMMENDATIONS & OUTPUTS | RDA FOR DISCIPLINES | PLENARIES | EVENTS | NEWS & MEDIA

RDA 11th Plenary Meeting: 21 - 23 March 2018, Berlin, Germany | Full programme & Industry side event | Photo gallery

News

Job posting: Director of Office, World Data System International

RDA Events

Improving Reproducibility in Research: The Role of

Request for comments

Array Database Assessment

@ Claudia Bauzer Medeiros

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

OPEN SCIENCE BY DESIGN

Realizing a Vision for 21st Century Research

National Academies of Sciences, Engineering, Medicine

July 2018

Open science =

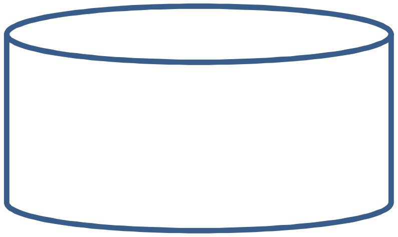
- **Open access = papers**
- **Open data**
- **Open methods = open source**

@ Claudia Bauzer Medeiros

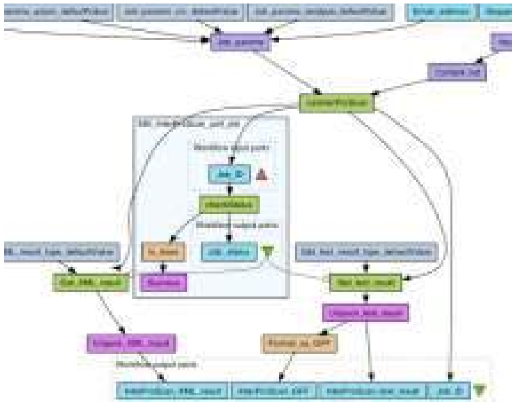
OPEN SCIENCE?

- OPEN ACCESS
  - Papers

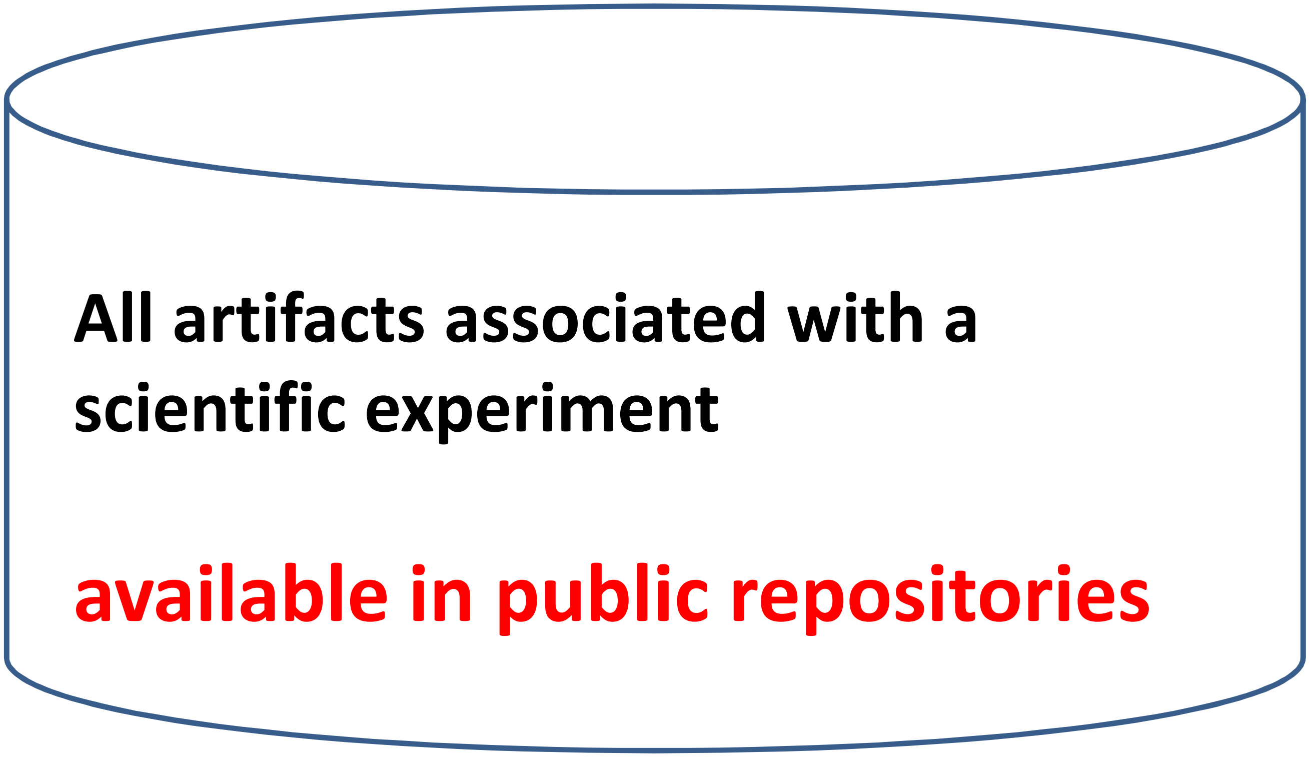
- OPEN DATA
  - Static



- OPEN PROCESSES
  - Dynamic



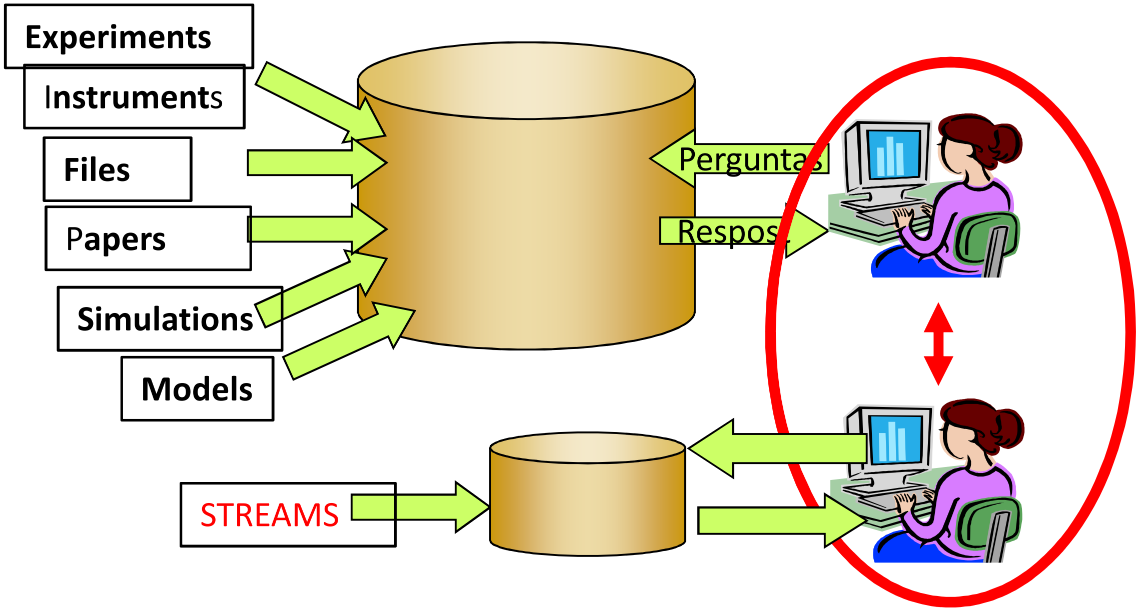
OPEN????



All artifacts associated with a scientific experiment

available in public repositories

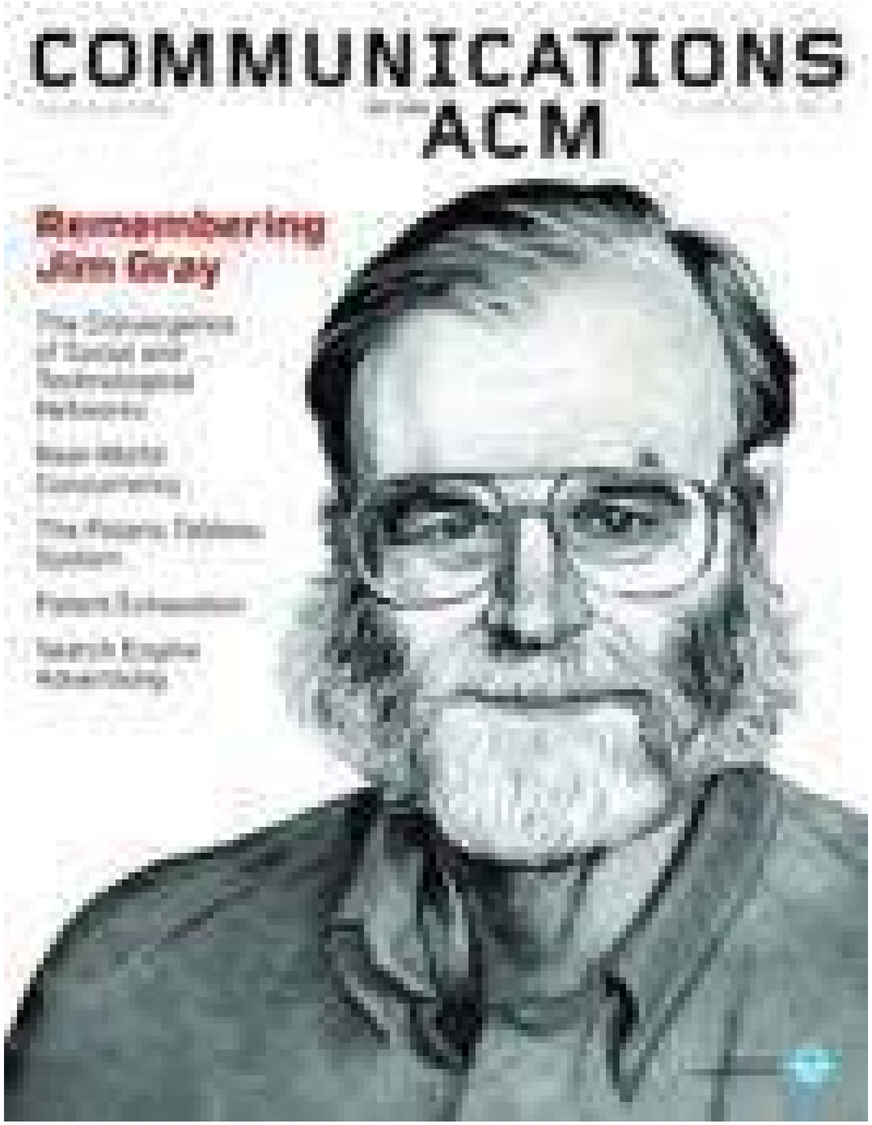
How to Share – slide adapted from Jim Gray



Data driven-science

PARENTHESIS – JIM GRAY?

CACM Nov 2008



COMMUNICATIONS OF THE ACM

Remembering Jim Gray

The Convergence of Social and Technological Networks

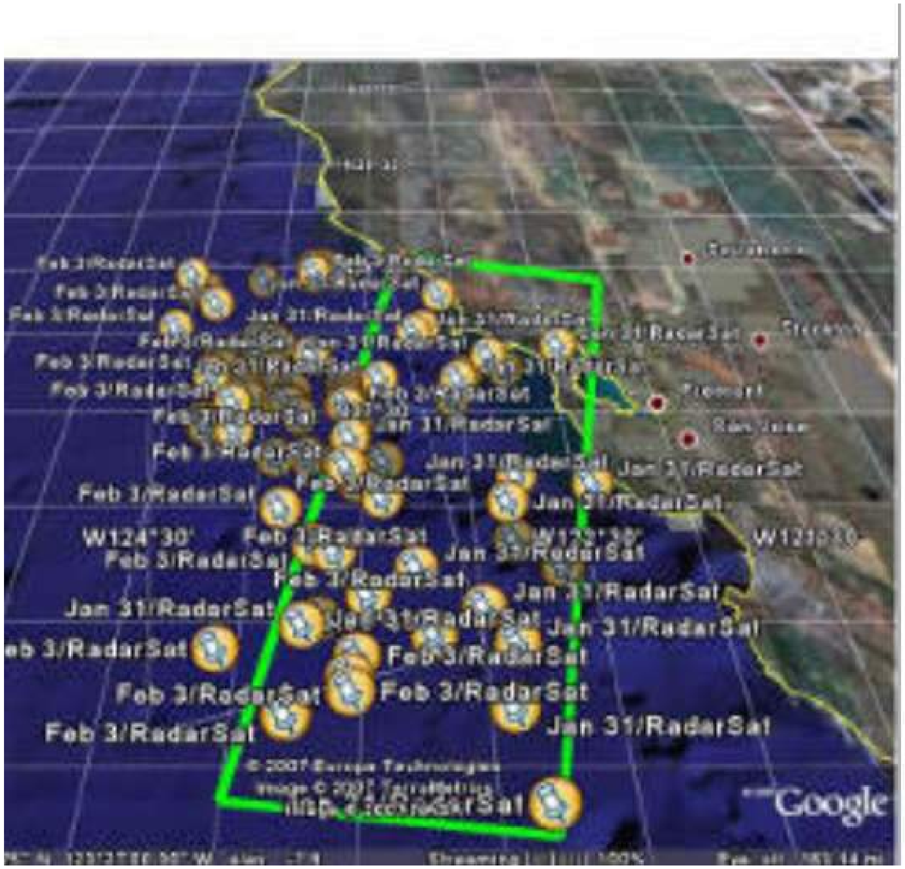
[illegible]

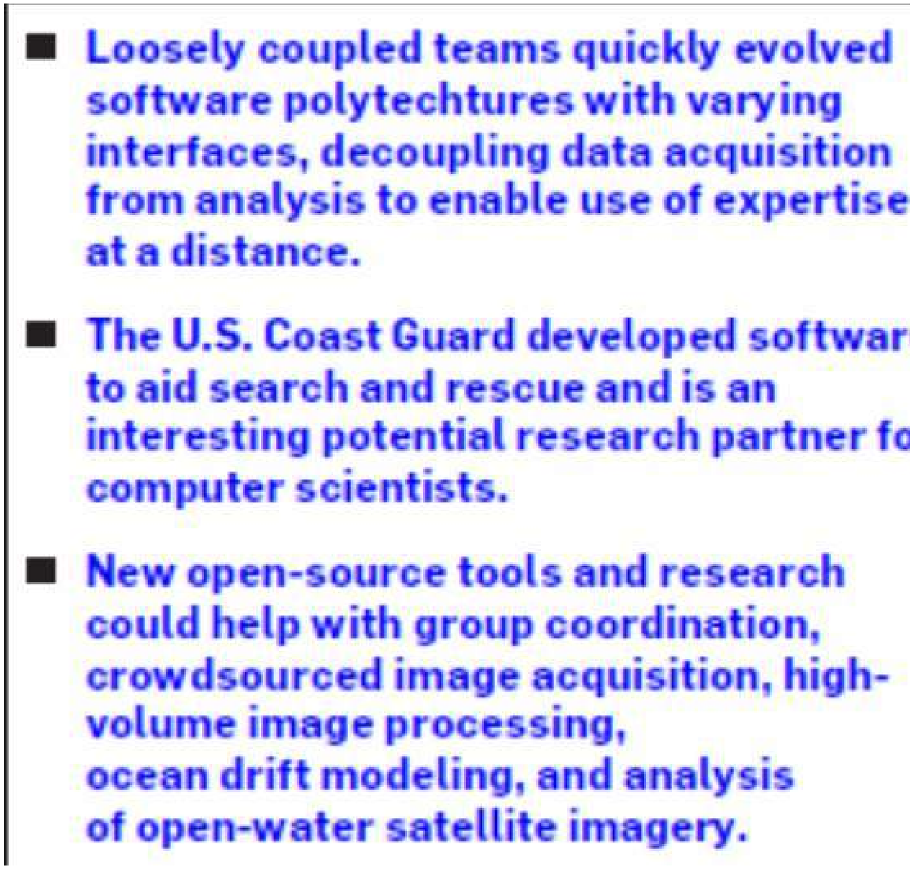
[illegible]

[illegible]

Search Engine Advertising

CACM 54(7):77-87, 2011 (Hellerstein, Tennenhouse)

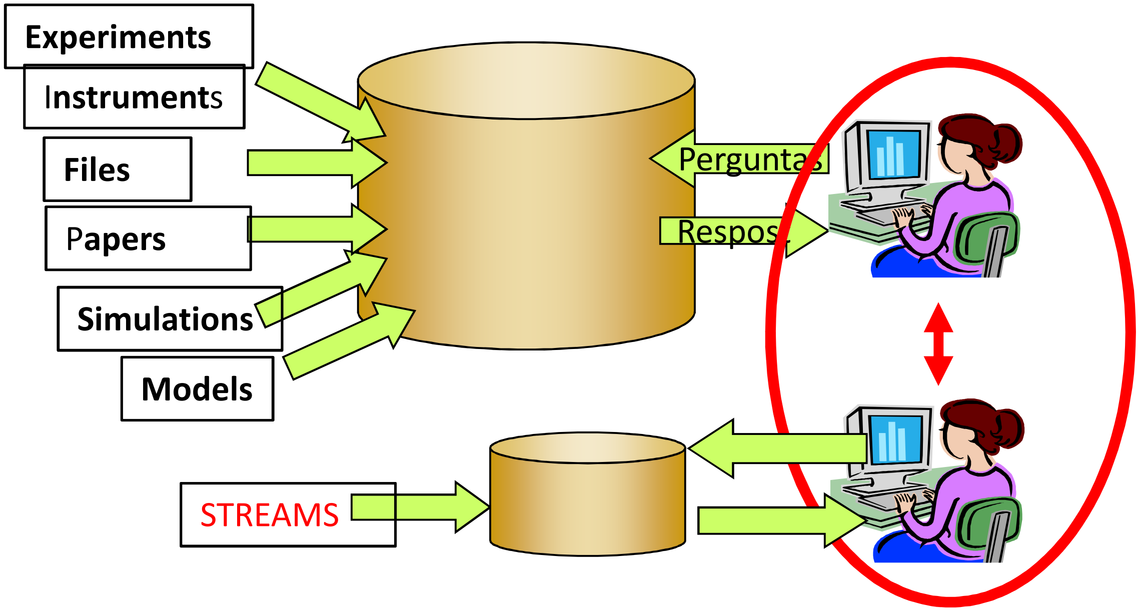




- Loosely coupled teams quickly evolved software polytechtures with varying interfaces, decoupling data acquisition from analysis to enable use of expertise at a distance.
- The U.S. Coast Guard developed softwar to aid search and rescue and is an interesting potential research partner fo computer scientists.
- New open-source tools and research could help with group coordination, crowdsourced image acquisition, high-volume image processing, ocean drift modeling, and analysis of open-water satellite imagery.

CLOSE PARENTHESIS

How to Share – slide adapted from Jim Gray



Data driven-science

What is Open Data?

- “What is OPEN **DIGITAL** DATA”
  - Share “everything”? Not necessarily

- Everyone can
  - Discover if data exist
  - Discover how to obtain them

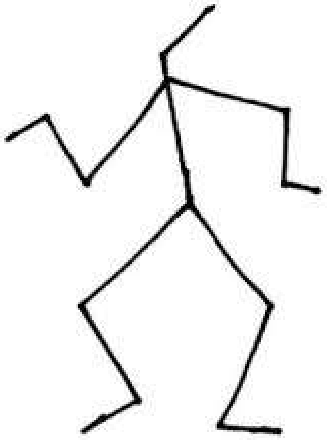
Under constraints – security, confidentiality, ethics, intellectual property

OPEN SCIENCE – OPEN METADATA

WHAT IS METADATA

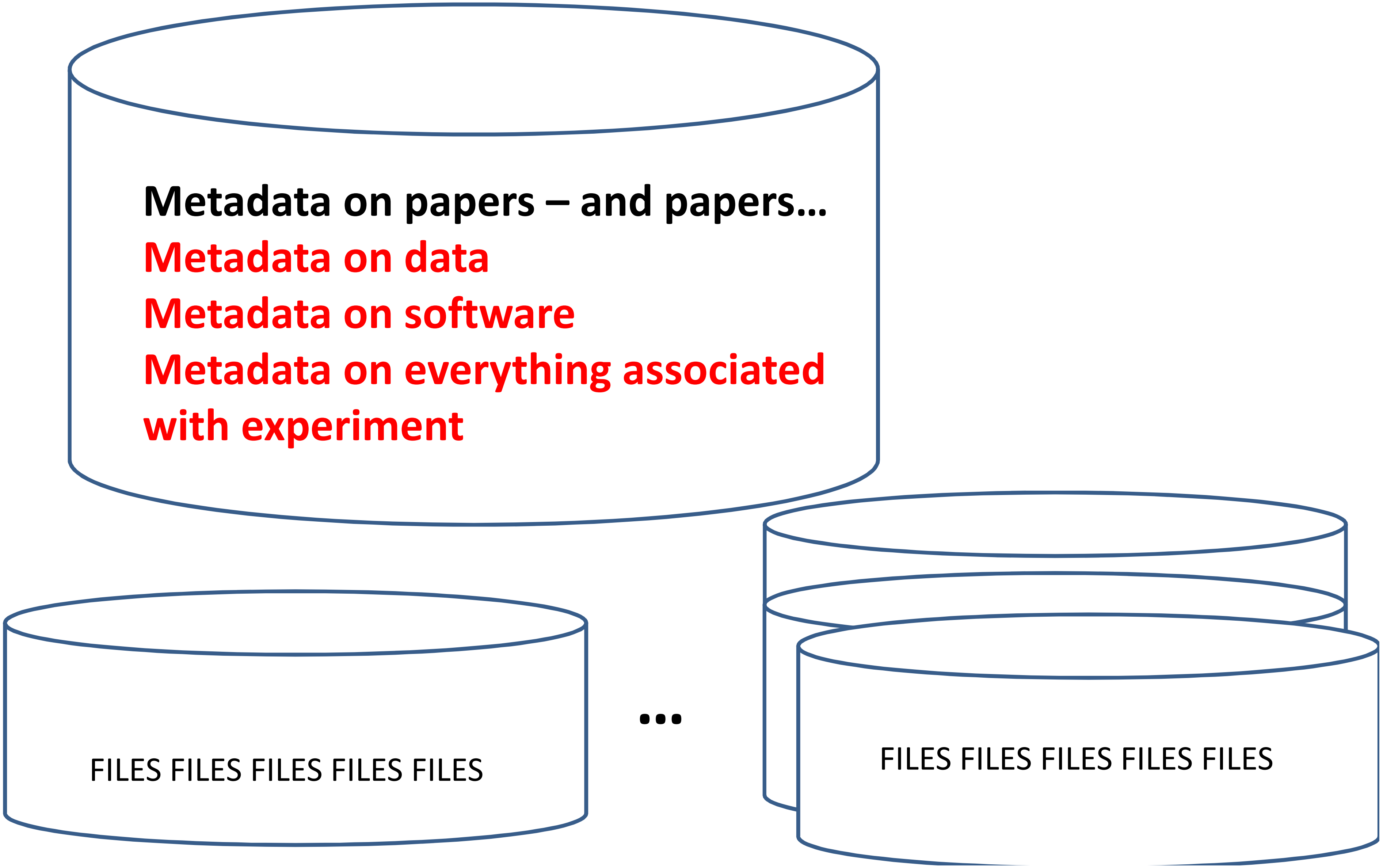
METADATA

Data Science
São Paulo School of Advanced Science
on Learning from Data
USP



OPEN SCIENCE – OPEN METADATA

OPEN SCIENCE



Metadata on papers – and papers…

Metadata on data

Metadata on software

Metadata on everything associated with experiment

FILES FILES FILES FILES FILES

…

FILES FILES FILES FILES FILES

International Scenario – Open Science

- Official policy in North America, Australia and New Zealand
- Compulsory for European financing after 2021
- Japan, South Korea
- Brazil – Federal government plans
- Brazil – FAPESP policies (Open access, open data)

Open Science – G7 Priority

1. **Human Capital Formation – research and innovation**
2. **Financing – inclusive science, research and innovation**
3. **Global Research Infrastructures**

---→ Open Science

(Canada, USA, France, Germany, Japan, Italy, UK)

+ Representative from EU

G7 2017
ITALIA

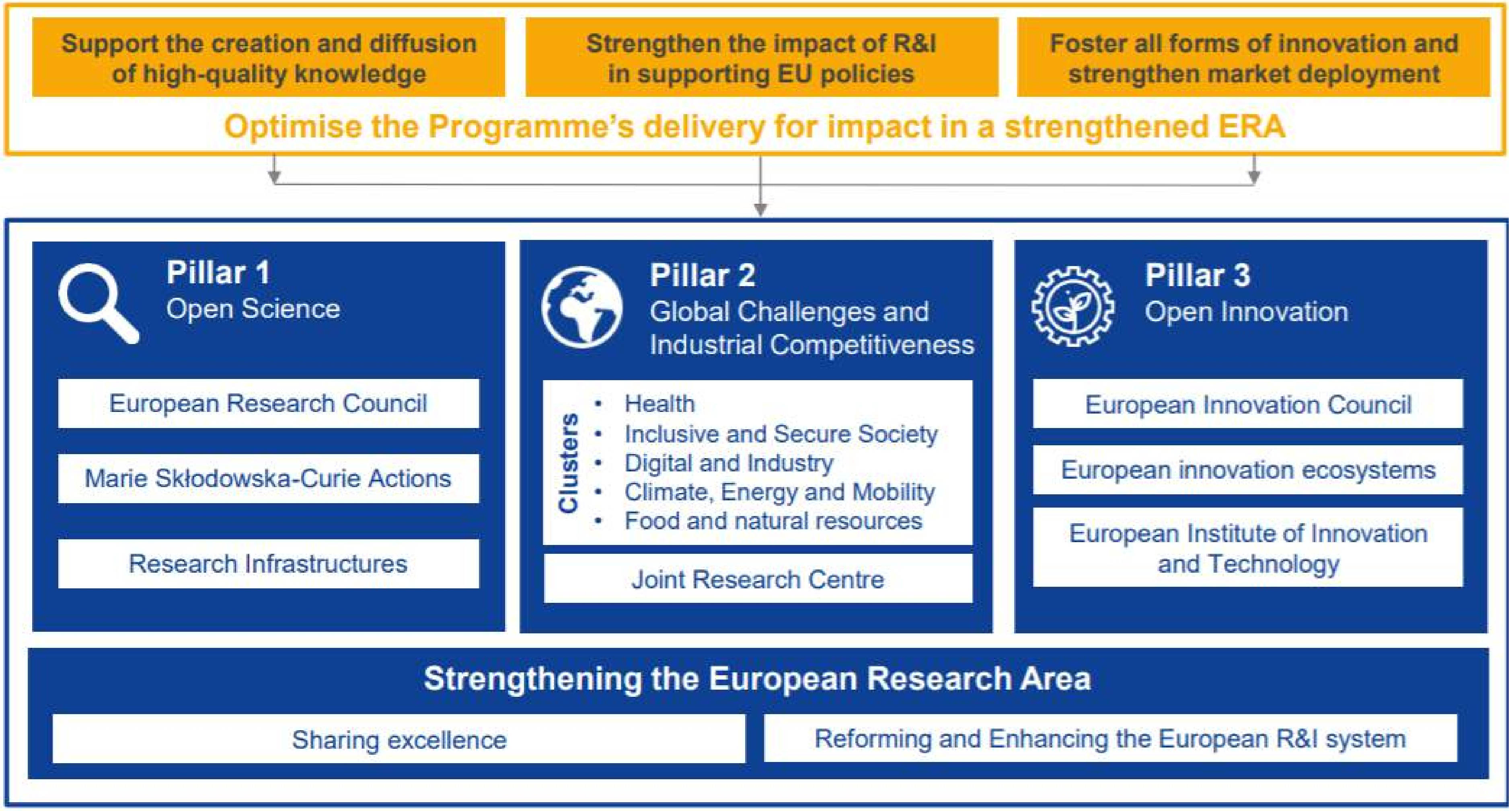
G7 SCIENCE MINISTERS' COMMUNIQUÉ

Turin, 27 – 28 September

@ Claudia Bauzer Medeiros

Horizon Europe: evolution not revolution

Specific objectives of the Programme



Support the creation and diffusion of high-quality knowledge

Strengthen the impact of R&I in supporting EU policies

Foster all forms of innovation and strengthen market deployment

Optimise the Programme's delivery for impact in a strengthened ERA

Pillar 1
Open Science

- European Research Council
- Marie Skłodowska-Curie Actions
- Research Infrastructures

Pillar 2
Global Challenges and Industrial Competitiveness

Clusters
- Health
- Inclusive and Secure Society
- Digital and Industry
- Climate, Energy and Mobility
- Food and natural resources

Joint Research Centre

Pillar 3
Open Innovation

- European Innovation Council
- European innovation ecosystems
- European Institute of Innovation and Technology

Strengthening the European Research Area

- Sharing excellence
- Reforming and Enhancing the European R&I system

European Commission

JUNE 2018

“The European Union could save €10.2 billion per year by using FAIR* and free open access research data.”

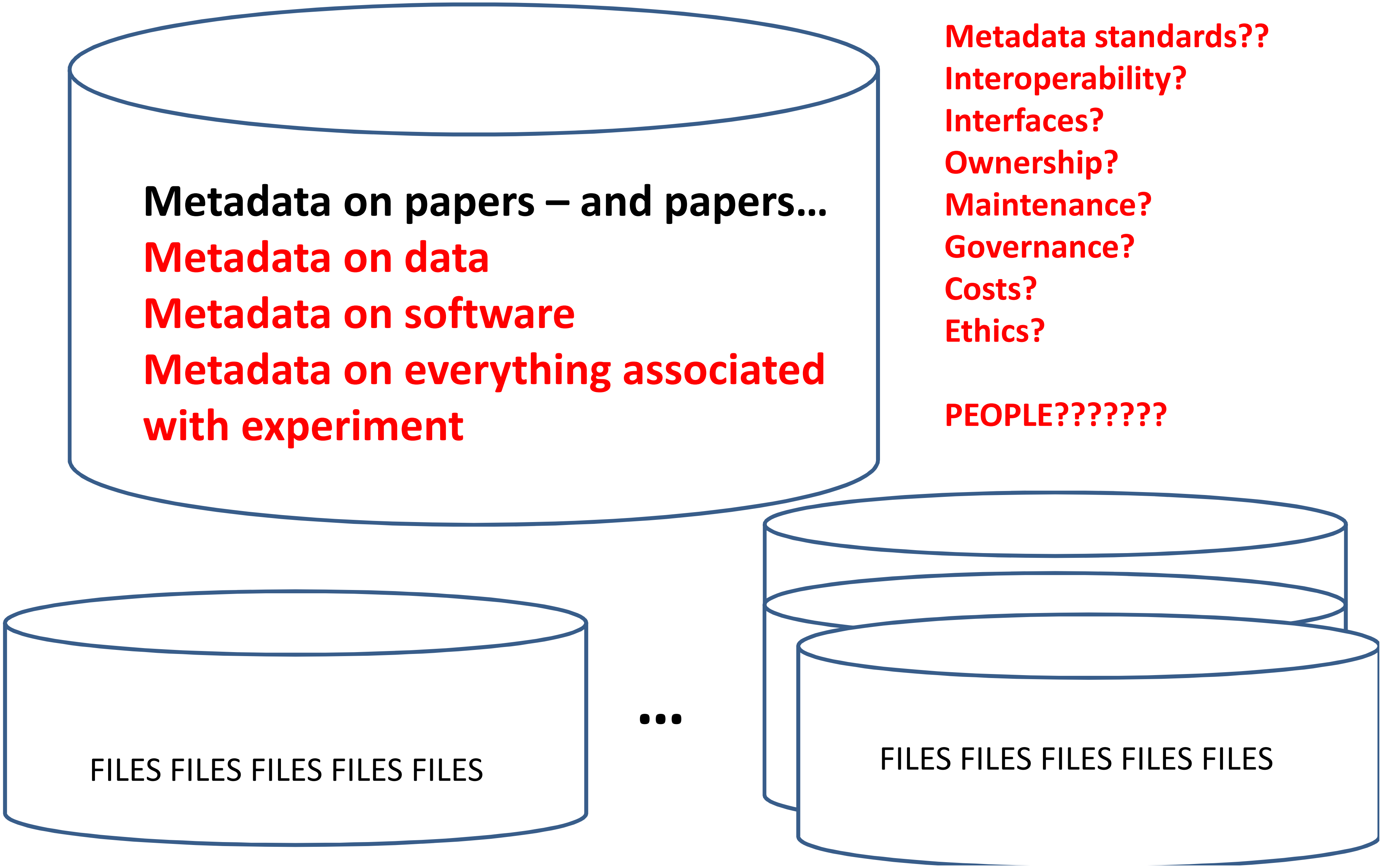
PwC EU Services (March 2018). *Cost-Benefit analysis for FAIR research data - Cost of not having FAIR research data*. Directorate General for Research and Innovation (European Commission).

*FAIR: Data meeting standards of Findability, Accessibility, Interoperability and Usability.

Why – The need for Open Science

- Validate research and advance science
- (Re)use = save resources
  - Data
  - Processes
  - People(?)
- (Re)use = improve, modify, accelerate scientific research
- Avoid fraud - transparency

OPEN SCIENCE - Challenges



Metadata on papers – and papers…

Metadata on data

Metadata on software

Metadata on everything associated with experiment

Metadata standards??
Interoperability?
Interfaces?
Ownership?
Maintenance?
Governance?
Costs?
Ethics?

PEOPLE???????

FILES FILES FILES FILES FILES

…

FILES FILES FILES FILES FILES

Open Science challenges

- Metadata standards
- Interoperability
  - Data
  - Processes
  - People(?)
- Interfaces, ownership, maintenance
- Preservation
- Ethics

OUTLINE

Open Science

- Definitions
- Challenges

4 Challenges in (Research Data) Sharing

Open Data at FAPESP

- Data Management Plans
- Network of open research data repositories

(Four) Challenges in Sharing Data

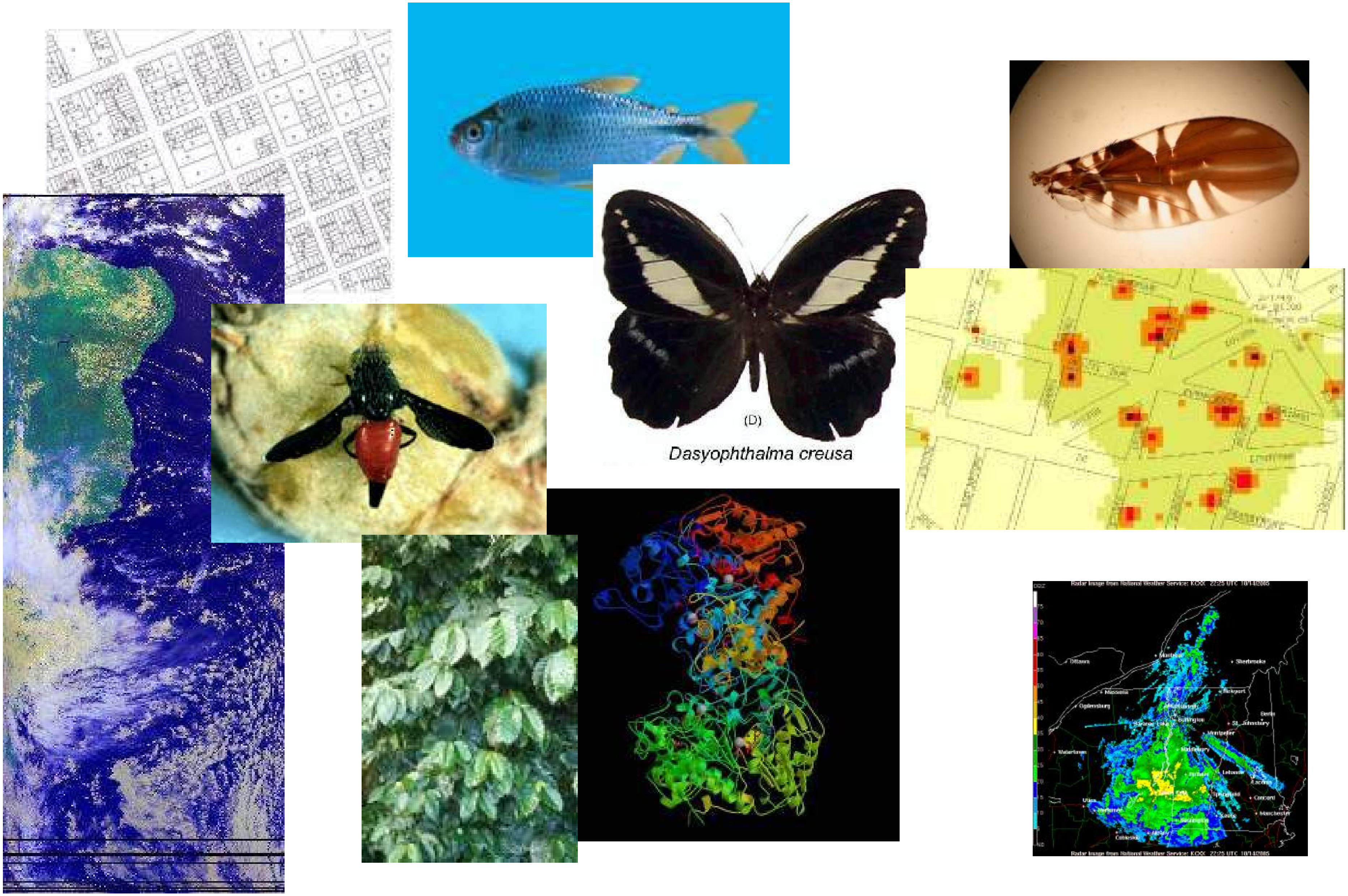
(1). What is data ?

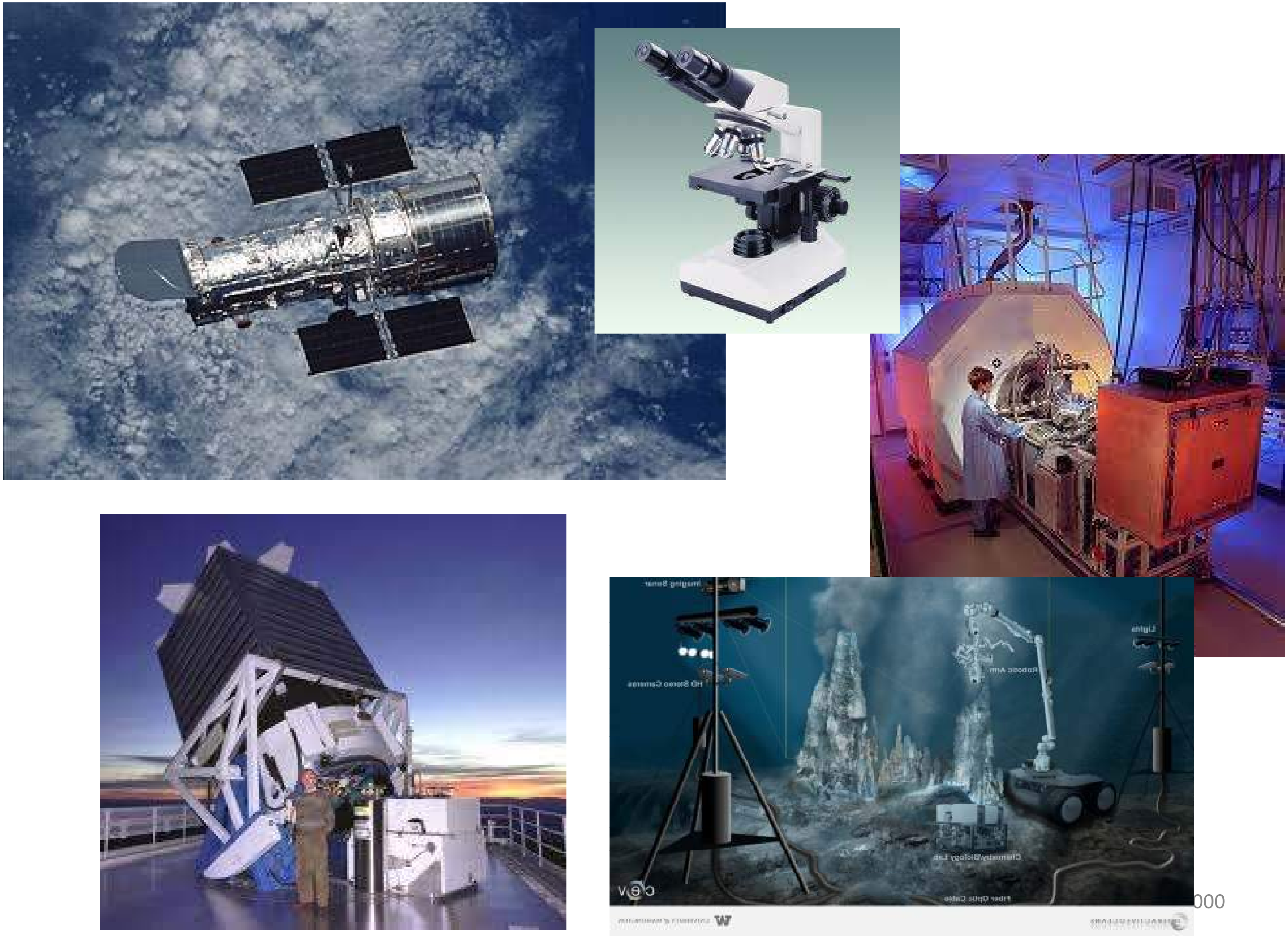
(2). Lack of common/consensual infrastructures

(3). Finding and identifying

(4). Understanding

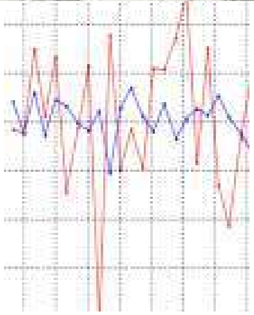
Data sources?





Research Data?

- Direct and indirect observations



Big questions

WHAT IS (RESEARCH) DATA????

- [Digital only]
- **Any file associated with research**

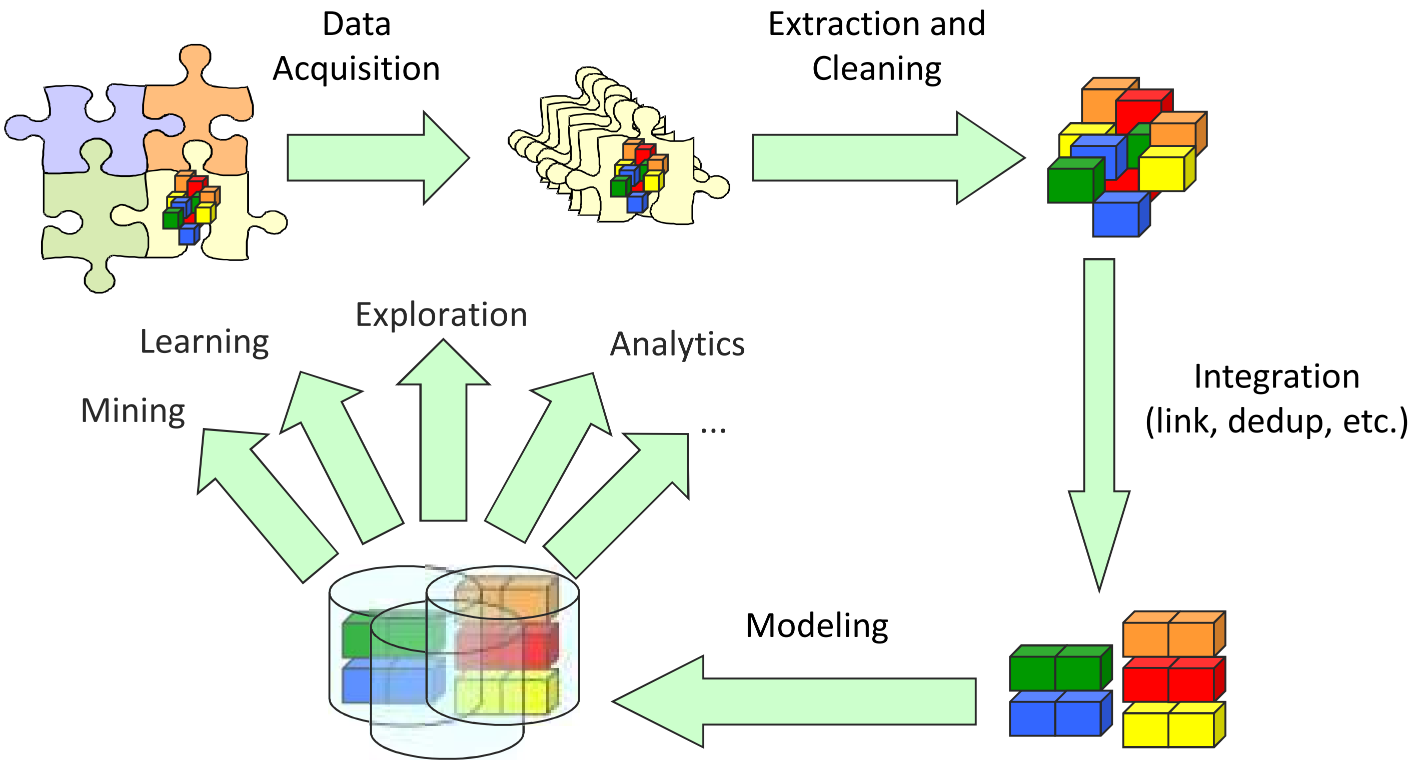
HOW TO SHARE????

- **Via research data repositories**

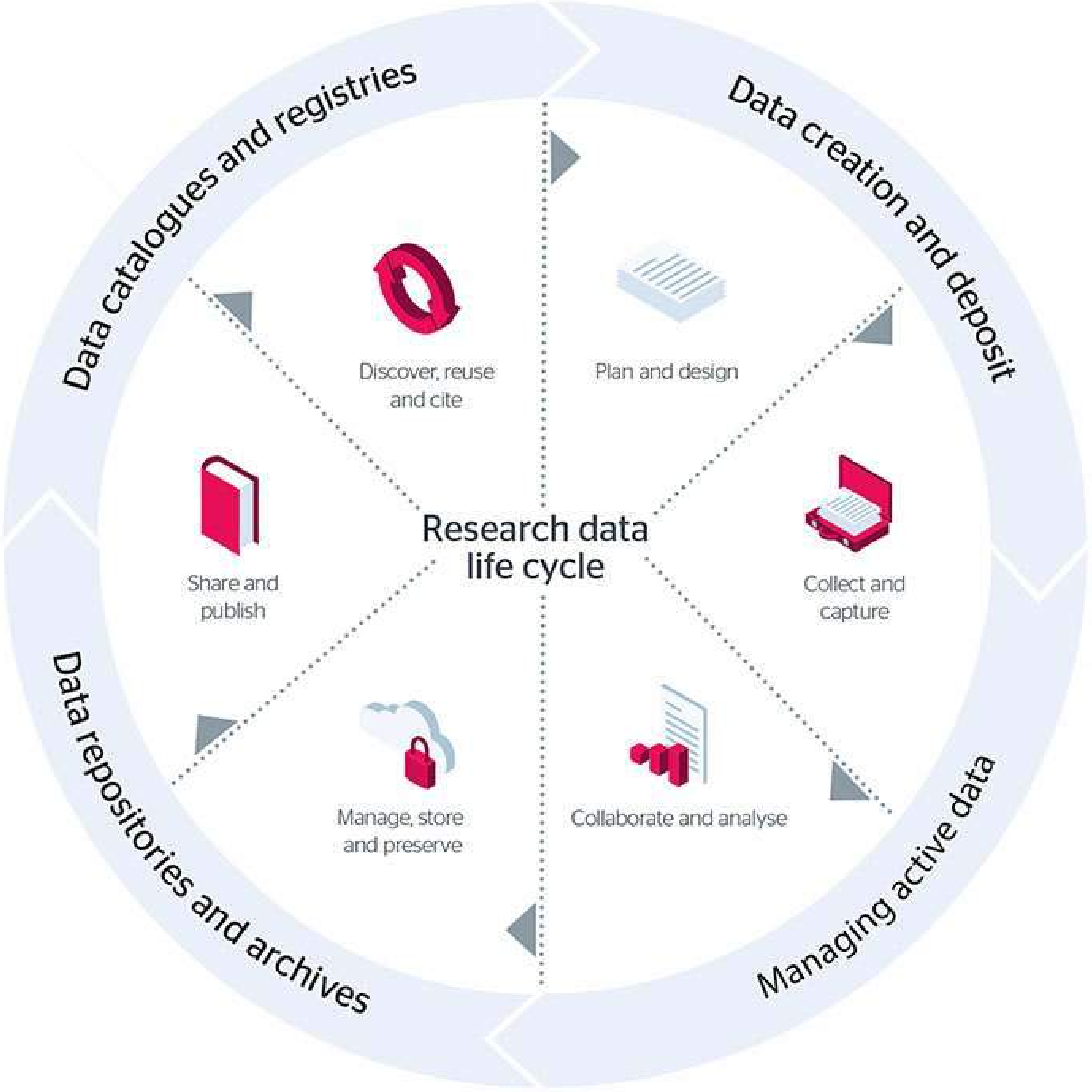
FOR WHAT????

- **Advance research -> scientific, economic, social, cultural benefits**

The Big Data Pipeline



H. V. Jagadish ACM SIGMOD Blog - 06/2012


Data catalogues and registries
Data creation and deposit
Managing active data
Data repositories and archives
Research data life cycle
Discover, reuse and cite
Plan and design
Collect and capture
Collaborate and analyse
Manage, store and preserve
Share and publish

Open, Sharing, Challenges few Mention

Updating and versioning

Curation and long term preservation

Visualization

Uses, reuses and mis-uses/ethics

Doors that are open and closed via choice of data to collect

- For xxx science to work, interpretation is needed (who are the “appropriate” experts?)

(Four) Challenges in Open Science

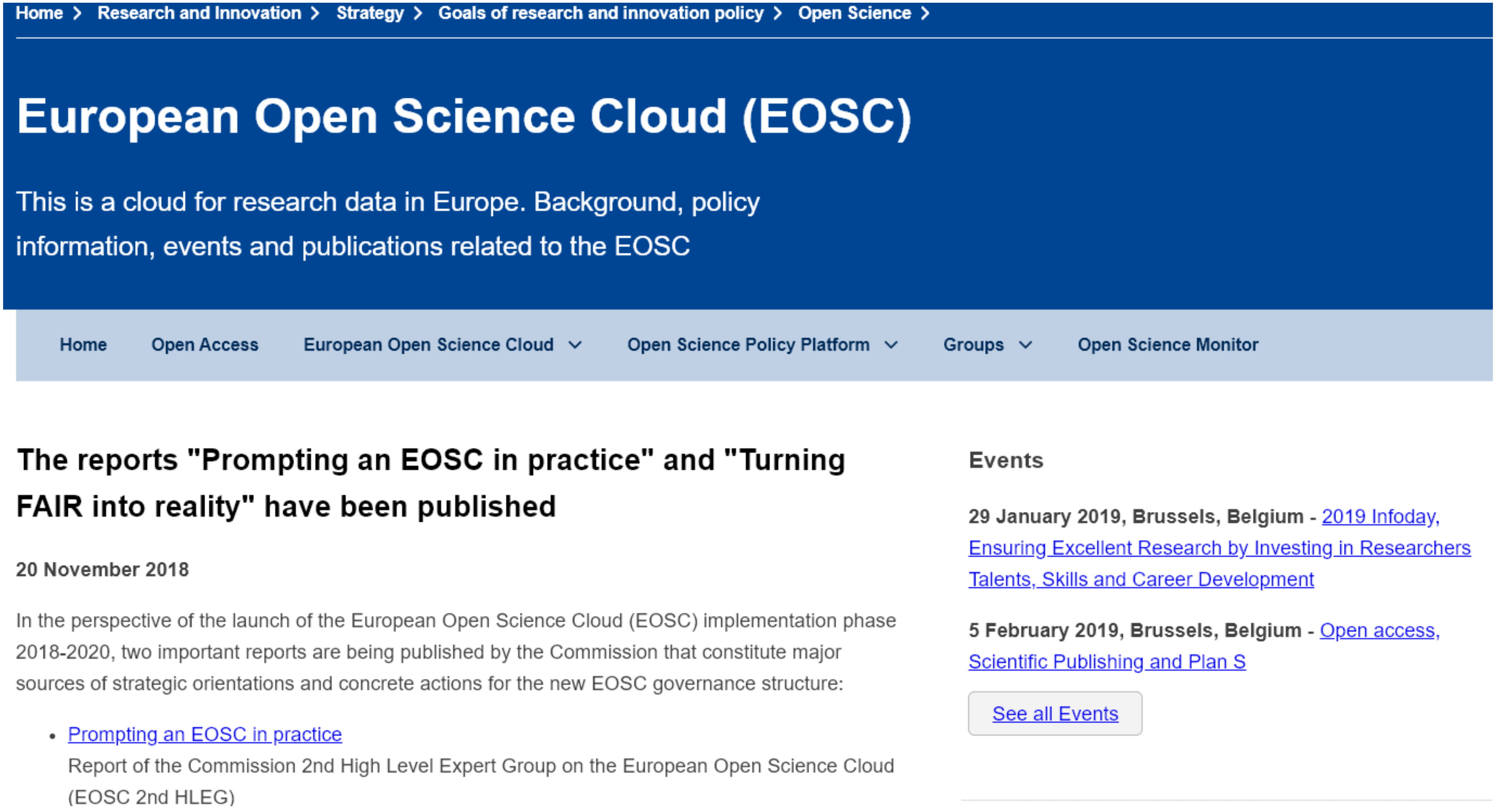
(1). What is data ?

(2). Lack of common/consensual infrastructures

(3). Finding and identifying

(4). Understanding

Infrastructure(s) = 260M + 12B euros/yr


Home › Research and Innovation › Strategy › Goals of research and innovation policy › Open Science ›
Home
Open Access
European Open Science Cloud
Open Science Policy Platform
Groups
Open Science Monitor
See all Events

European Open Science Cloud (EOSC)

This is a cloud for research data in Europe. Background, policy information, events and publications related to the EOSC

The reports "Prompting an EOSC in practice" and "Turning FAIR into reality" have been published

20 November 2018

In the perspective of the launch of the European Open Science Cloud (EOSC) implementation phase 2018-2020, two important reports are being published by the Commission that constitute major sources of strategic orientations and concrete actions for the new EOSC governance structure:

- Prompting an EOSC in practice
  Report of the Commission 2nd High Level Expert Group on the European Open Science Cloud (EOSC 2nd HLEG)

Events

29 January 2019, Brussels, Belgium - 2019 Infoday, Ensuring Excellent Research by Investing in Researchers Talents, Skills and Career Development

5 February 2019, Brussels, Belgium - Open access, Scientific Publishing and Plan S

Finding and Identifying

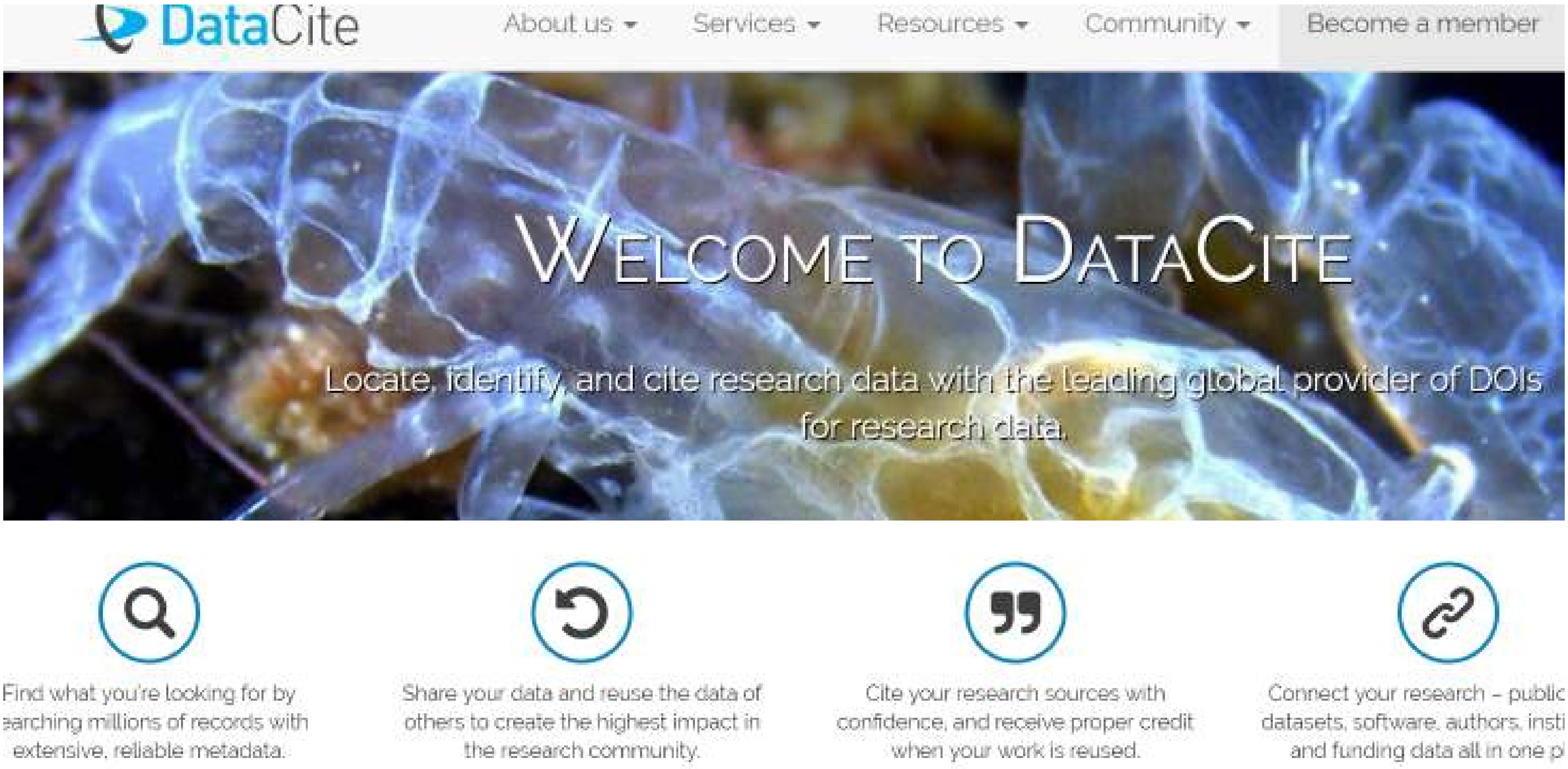
How to search/find (specification ?)

Where?

What to publish (everything vs ethics/privacy)

Unique id?

HOW??? Datacite.org (Find, share, cite, connect)



WELCOME TO DATACITE

Locate, identify, and cite research data with the leading global provider of DOIs for research data.

Find what you're looking for by earching millions of records with extensive, reliable metadata.

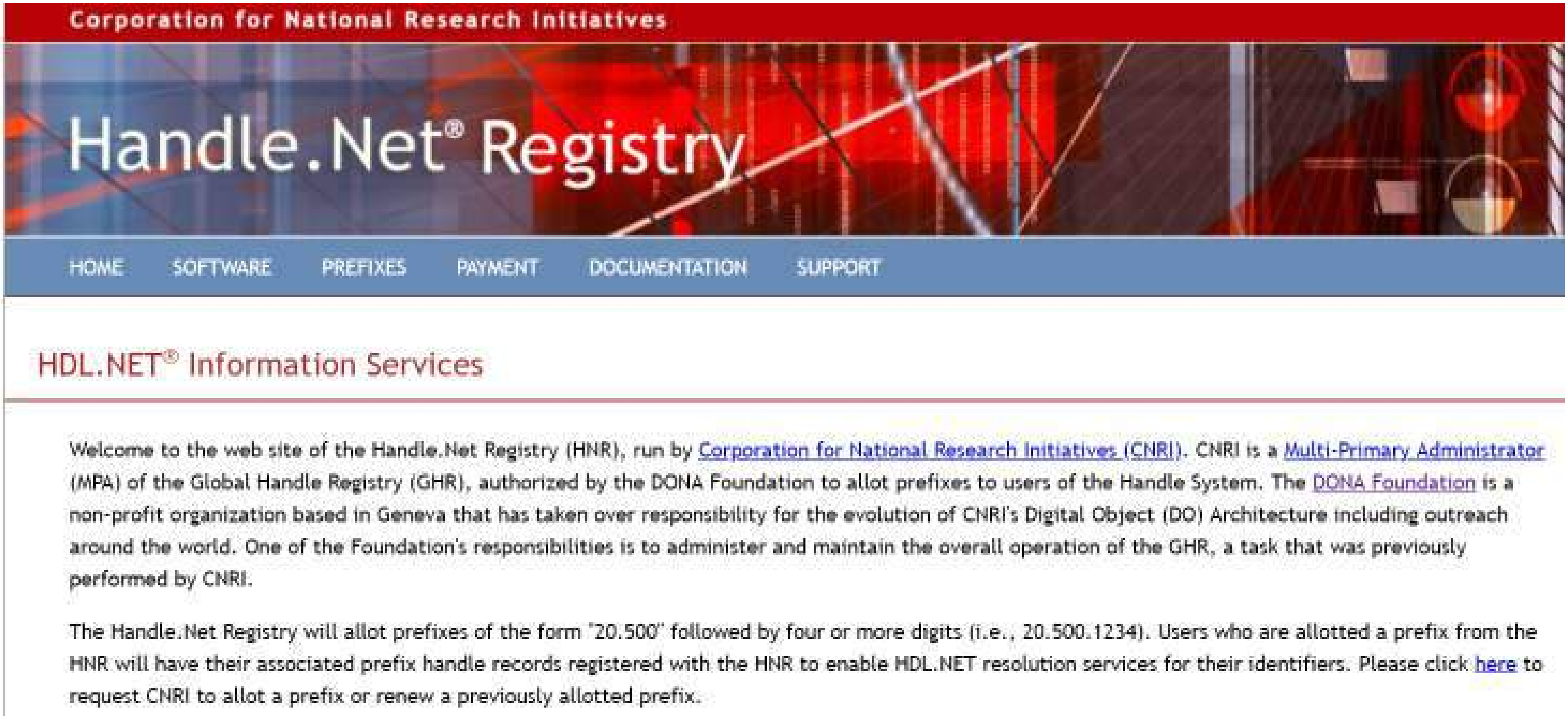
Share your data and reuse the data of others to create the highest impact in the research community.

Cite your research sources with confidence, and receive proper credit when your work is reused.

Connect your research – public datasets, software, authors, insti and funding data all in one p

@ Claudia Bauzer Medeiros

(HOW???) Handle.Net



HDL.NET® Information Services

Welcome to the web site of the Handle.Net Registry (HNR), run by Corporation for National Research Initiatives (CNRI). CNRI is a Multi-Primary Administrator (MPA) of the Global Handle Registry (GHR), authorized by the DONA Foundation to allot prefixes to users of the Handle System. The DONA Foundation is a non-profit organization based in Geneva that has taken over responsibility for the evolution of CNRI's Digital Object (DO) Architecture including outreach around the world. One of the Foundation's responsibilities is to administer and maintain the overall operation of the GHR, a task that was previously performed by CNRI.

The Handle.Net Registry will allot prefixes of the form "20.500" followed by four or more digits (i.e., 20.500.1234). Users who are allotted a prefix from the HNR will have their associated prefix handle records registered with the HNR to enable HDL.NET resolution services for their identifiers. Please click here to request CNRI to allot a prefix or renew a previously allotted prefix.

Understanding

How to understand what you find?

How to reuse it?

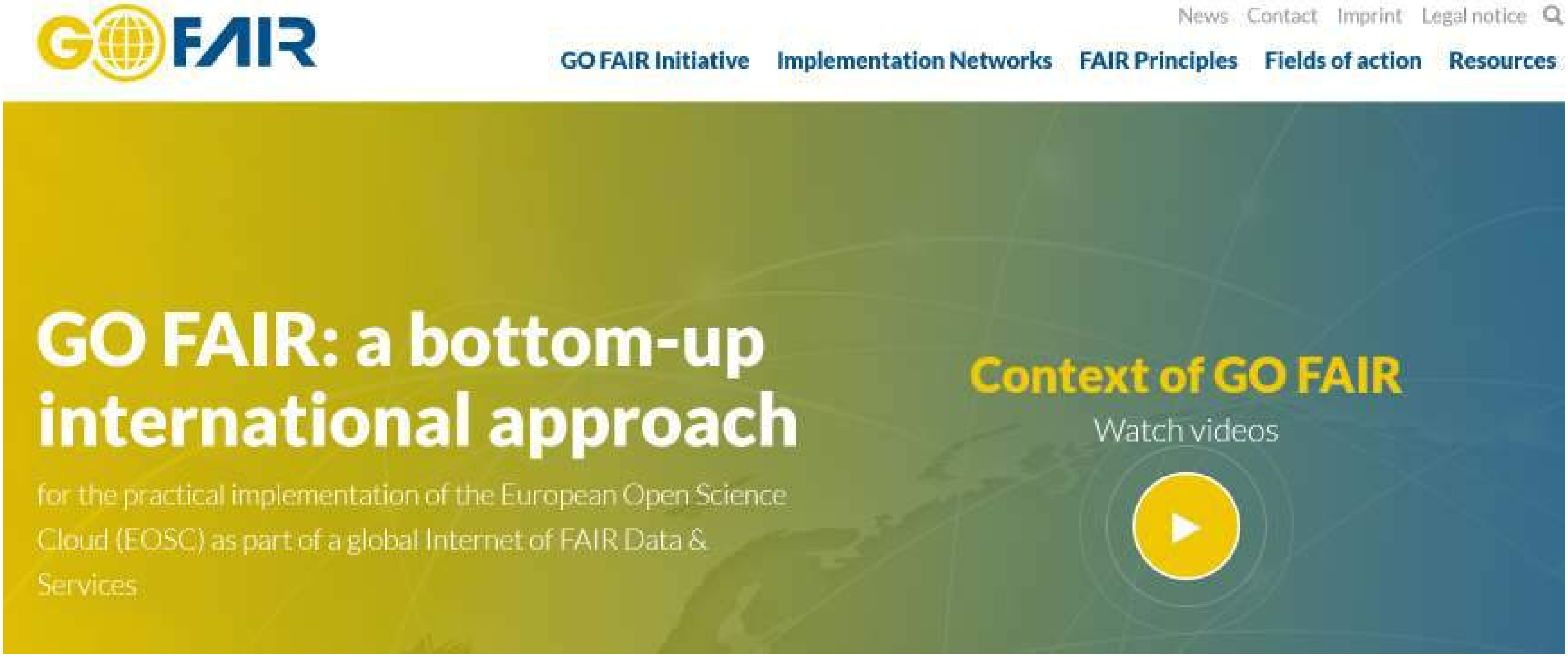
Everything is domain dependent
project dependent

Open Science requires FAIR Data

- Findable
- Accessible
- Interoperable
- Reusable

- **??? Have you fairicized your data???**

Open Science requires FAIR data



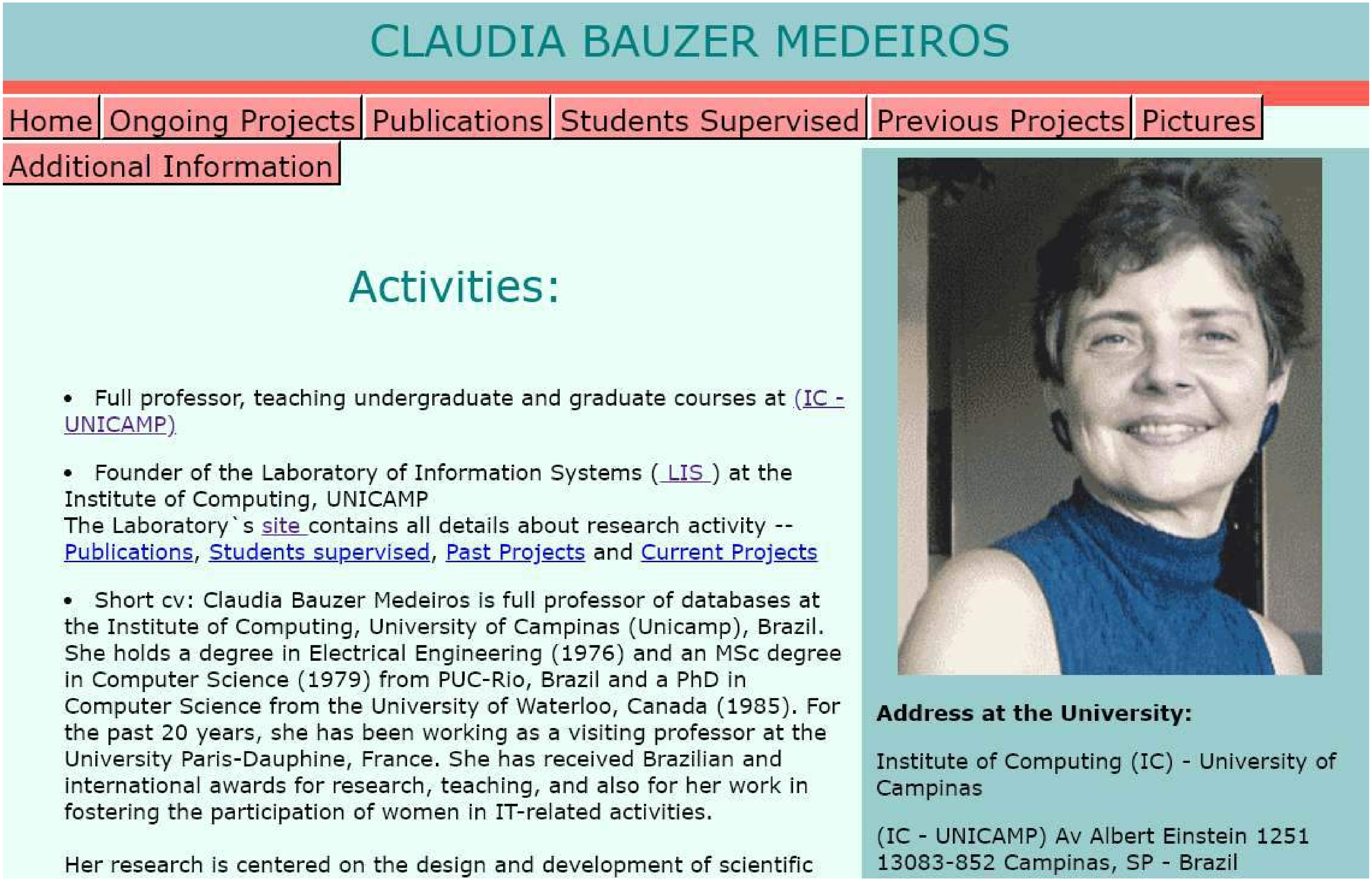
GO FAIR: a bottom-up international approach

for the practical implementation of the European Open Science Cloud (EOSC) as part of a global Internet of FAIR Data & Services

Context of GO FAIR

Watch videos

@ Claudia Bauzer Medeiros



CLAUDIA BAUZER MEDEIROS

Activities:

- Full professor, teaching undergraduate and graduate courses at (IC - UNICAMP)
- Founder of the Laboratory of Information Systems (LIS) at the Institute of Computing, UNICAMP
  The Laboratory`s site contains all details about research activity -- Publications, Students supervised, Past Projects and Current Projects
- Short cv: Claudia Bauzer Medeiros is full professor of databases at the Institute of Computing, University of Campinas (Unicamp), Brazil. She holds a degree in Electrical Engineering (1976) and an MSc degree in Computer Science (1979) from PUC-Rio, Brazil and a PhD in Computer Science from the University of Waterloo, Canada (1985). For the past 20 years, she has been working as a visiting professor at the University Paris-Dauphine, France. She has received Brazilian and international awards for research, teaching, and also for her work in fostering the participation of women in IT-related activities.

  Her research is centered on the design and development of scientific

Address at the University:

Institute of Computing (IC) - University of Campinas

(IC - UNICAMP) Av Albert Einstein 1251
13083-852 Campinas, SP - Brazil

Outline

Open Science

Definitions

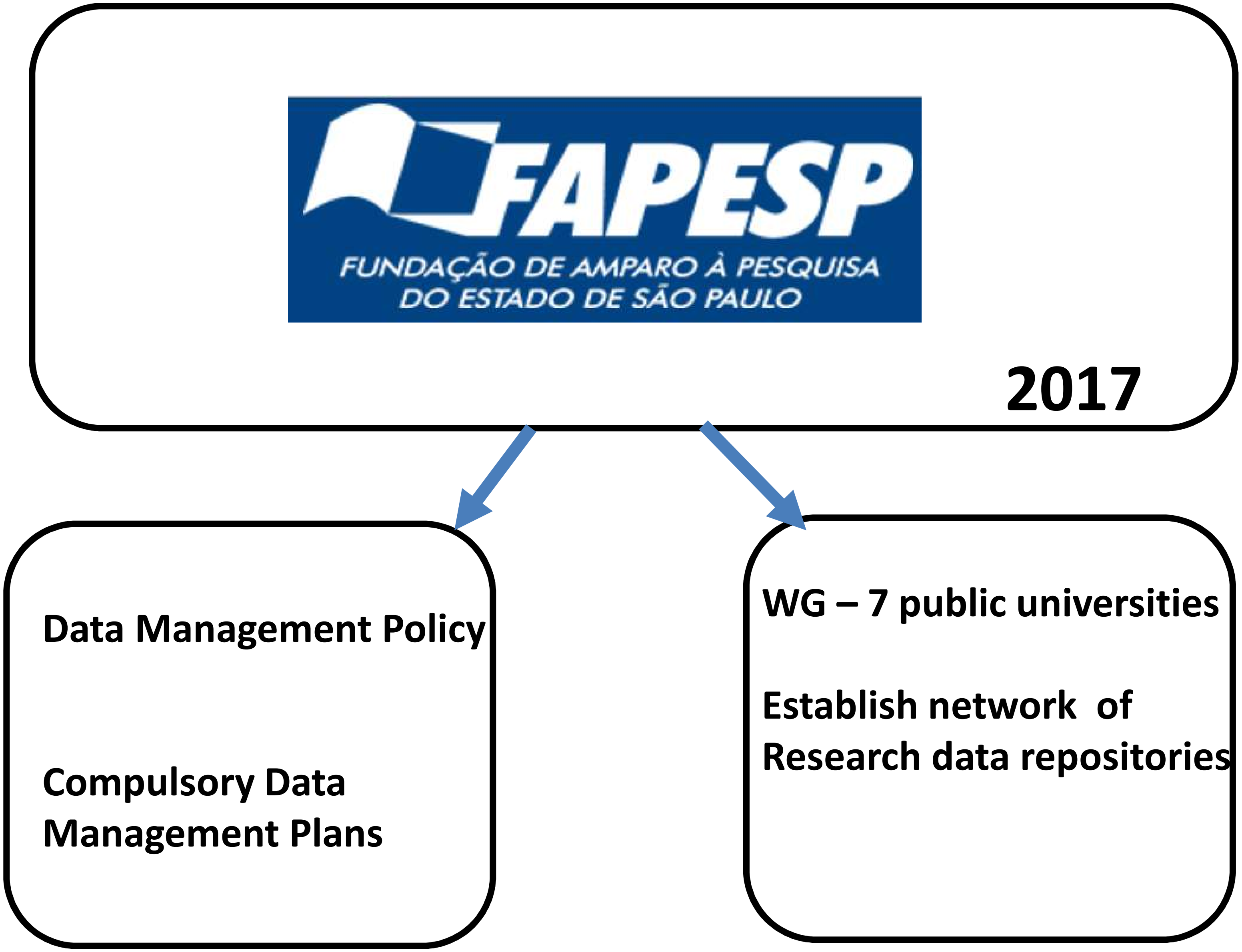
Challenges

4 Challenges in (Research Data) Sharing

Open Data at FAPESP

Data Management Plans

Network of open research data repositories



2017

Data Management Policy

Compulsory Data
Management Plans

WG – 7 public universities

Establish network of
Research data repositories

@ Claudia Bauzer Medeiros

Data Management Plan

- **WHICH** data will produce
- **WHERE** will store
- For how long **TIME**
- **HOW**

- Given ethical, privacy, IP aspects etc

WG – Data repository network

ufscar UNIFESP USP unesp ITA UFABC UNICAMP

Embrapa

Seven public universities, approx. 48 campi

11,5 thousand faculty

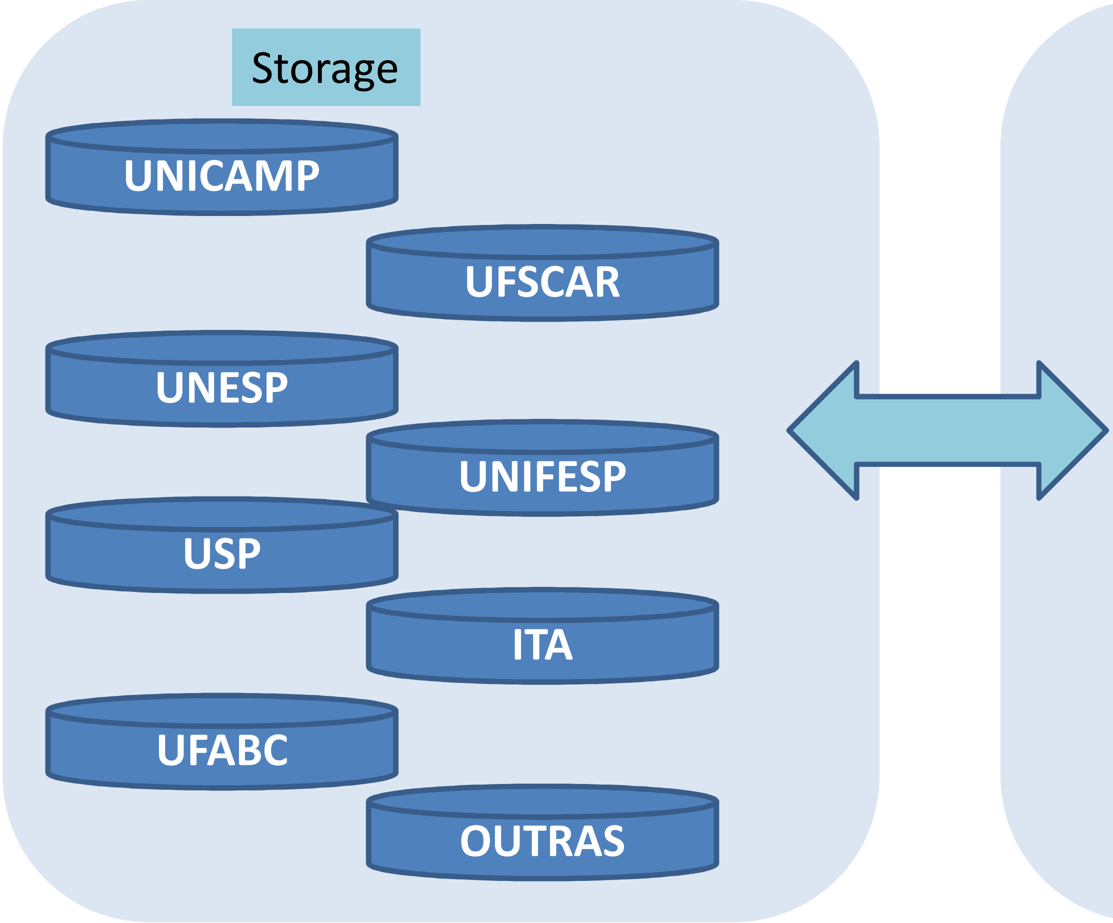
170 thousand students

+ researchers in (informatics in) agriculture

Mission – establish network

WG – Data Repository Network

- Each participant has its own system
- Single search (metadata harvester) interface



RETRIEVAL

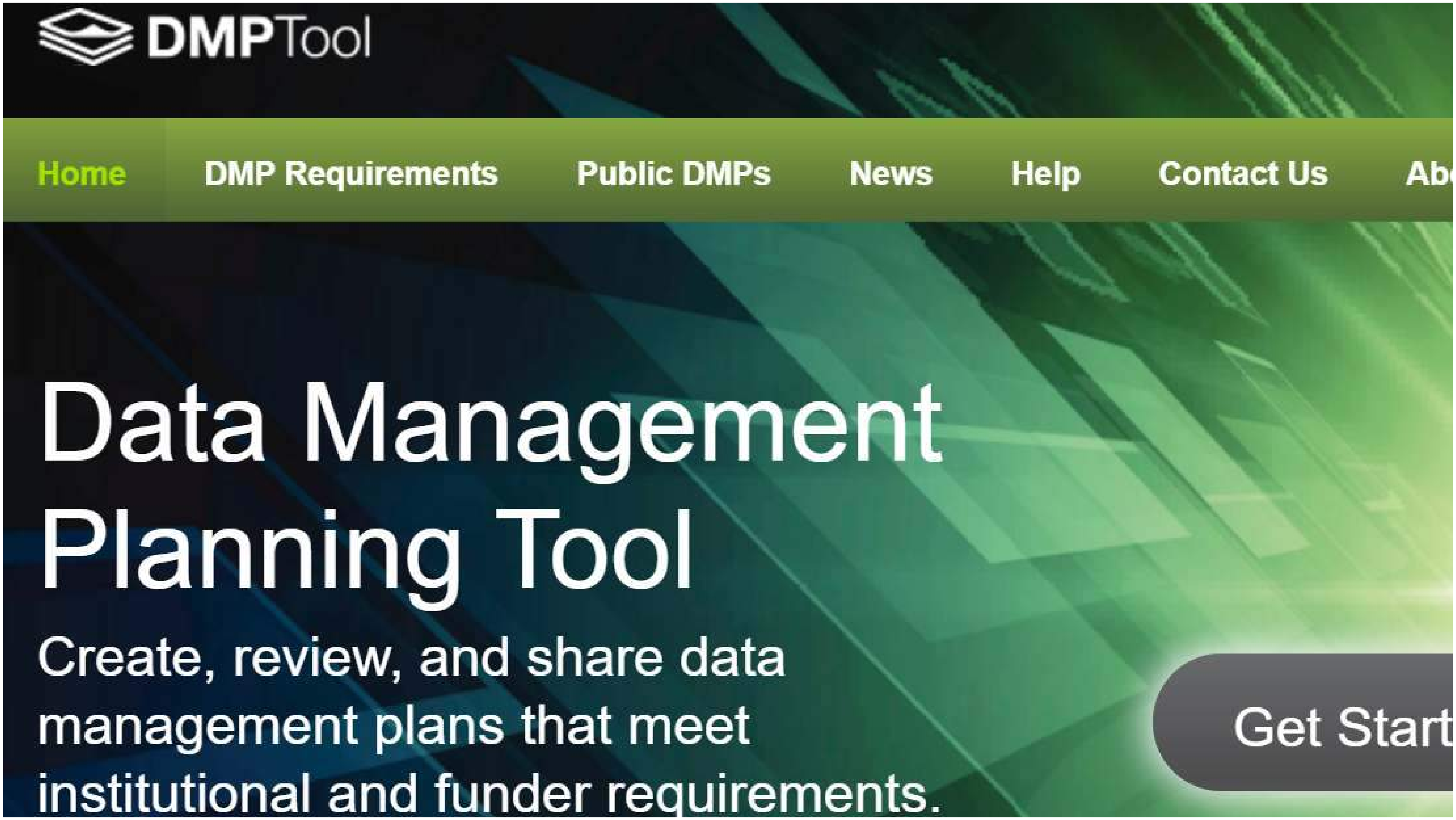
FAPESP Metabuscador de dados de pesquisa

NINE compulsory metadata fields

ID	Type	Description
1	dc.title	Project title
2	dc.subject	Keywords
3	dc.description	Abstract
4	dc.contributor.author	Author (ORCID)
5	dc.identifier.uri	File id
6	dc.description.sponsorship	Funding agencies
7	dc.description.sponsorshipId	Project numbers
8	dc.type	File type (software, others
9	dc.identifier	File id (handle)

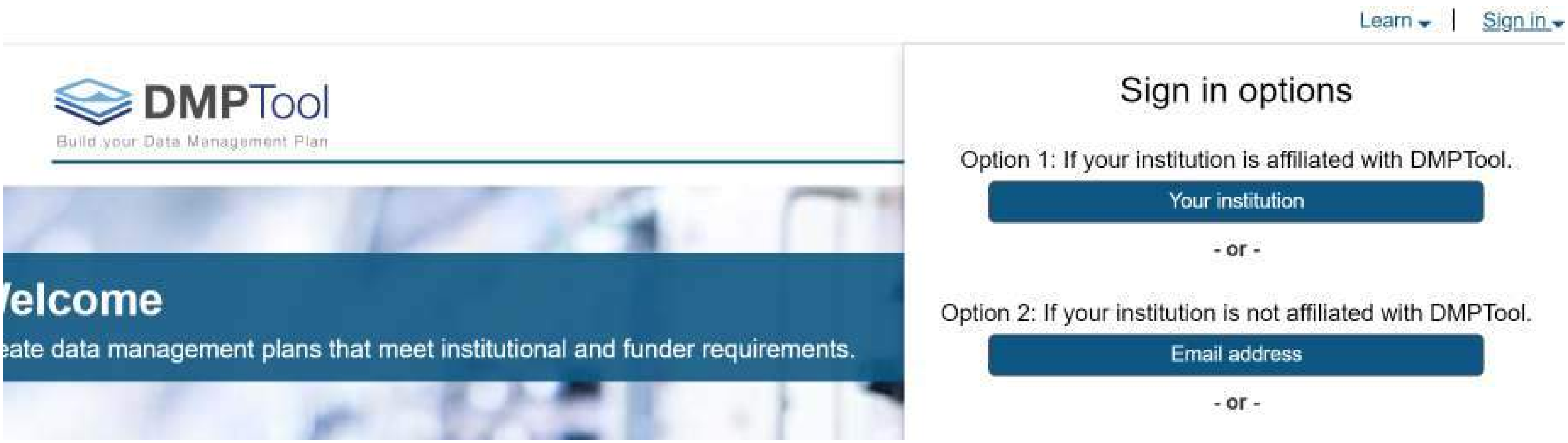
HOW TO UNDERSTAND DATA MANAGEMENT REQUIREMENTS !!!!!!!!!!!!!!!!!

Preparing DMP - dmptool.org



Data Management Planning Tool

Create, review, and share data management plans that meet institutional and funder requirements.



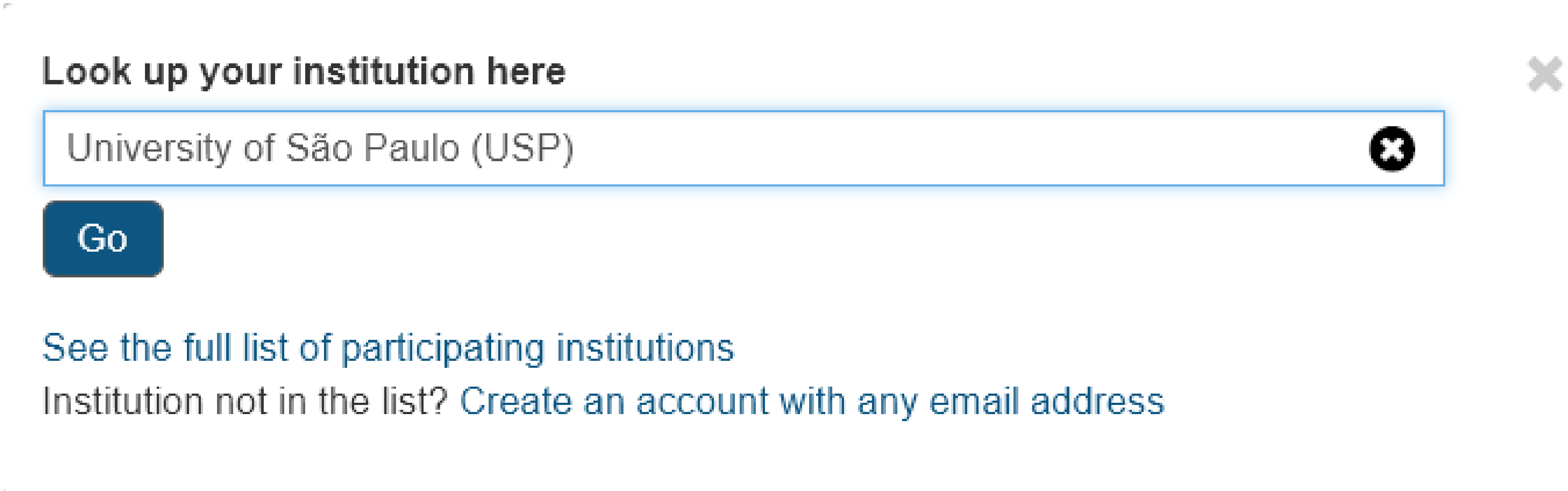
/elcome

ɛate data management plans that meet institutional and funder requirements.

Sign in options

Option 1: If your institution is affiliated with DMPTool.

Option 2: If your institution is not affiliated with DMPTool.

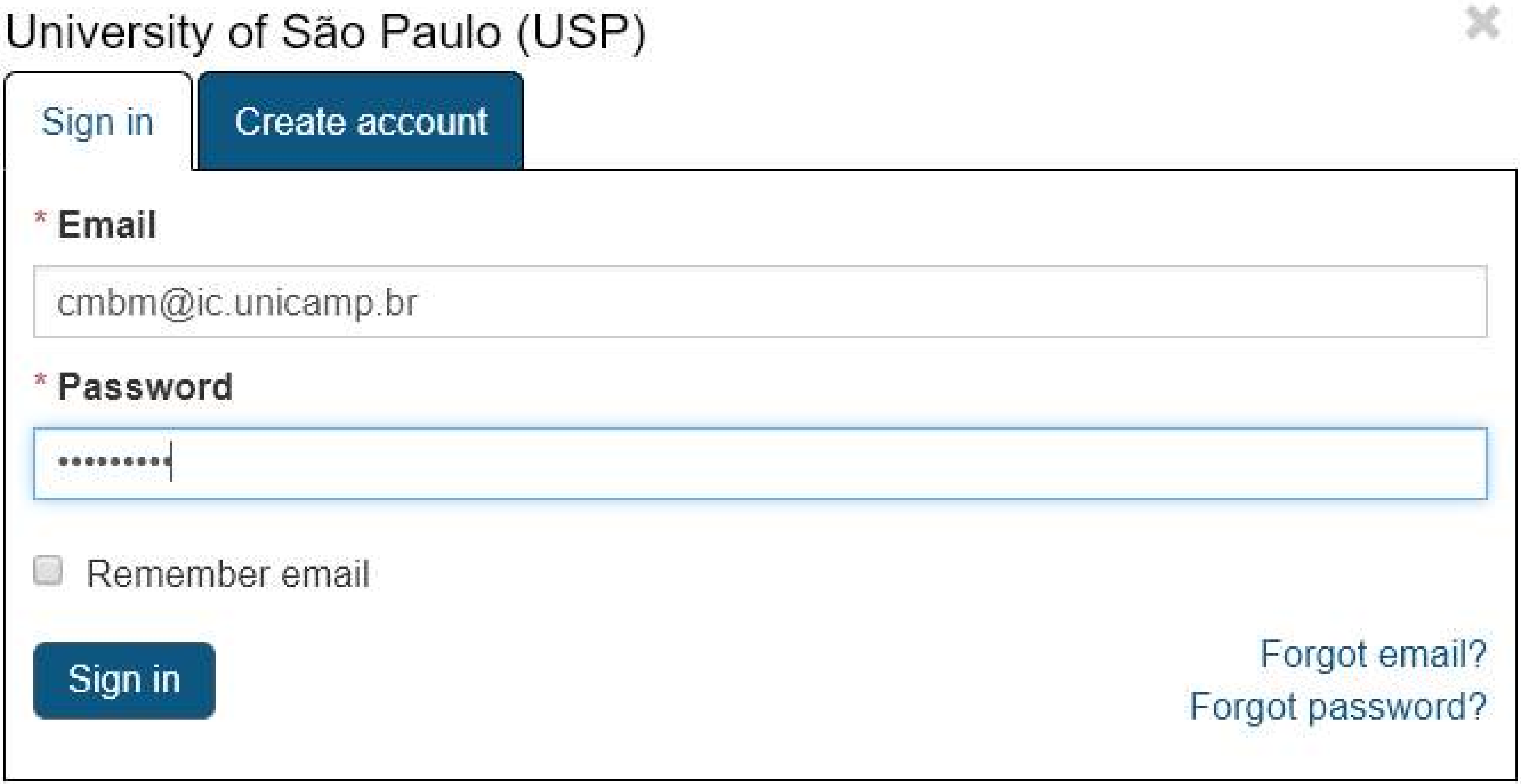


Look up your institution here

University of São Paulo (USP)

See the full list of participating institutions

Institution not in the list? Create an account with any email address



University of São Paulo (USP)

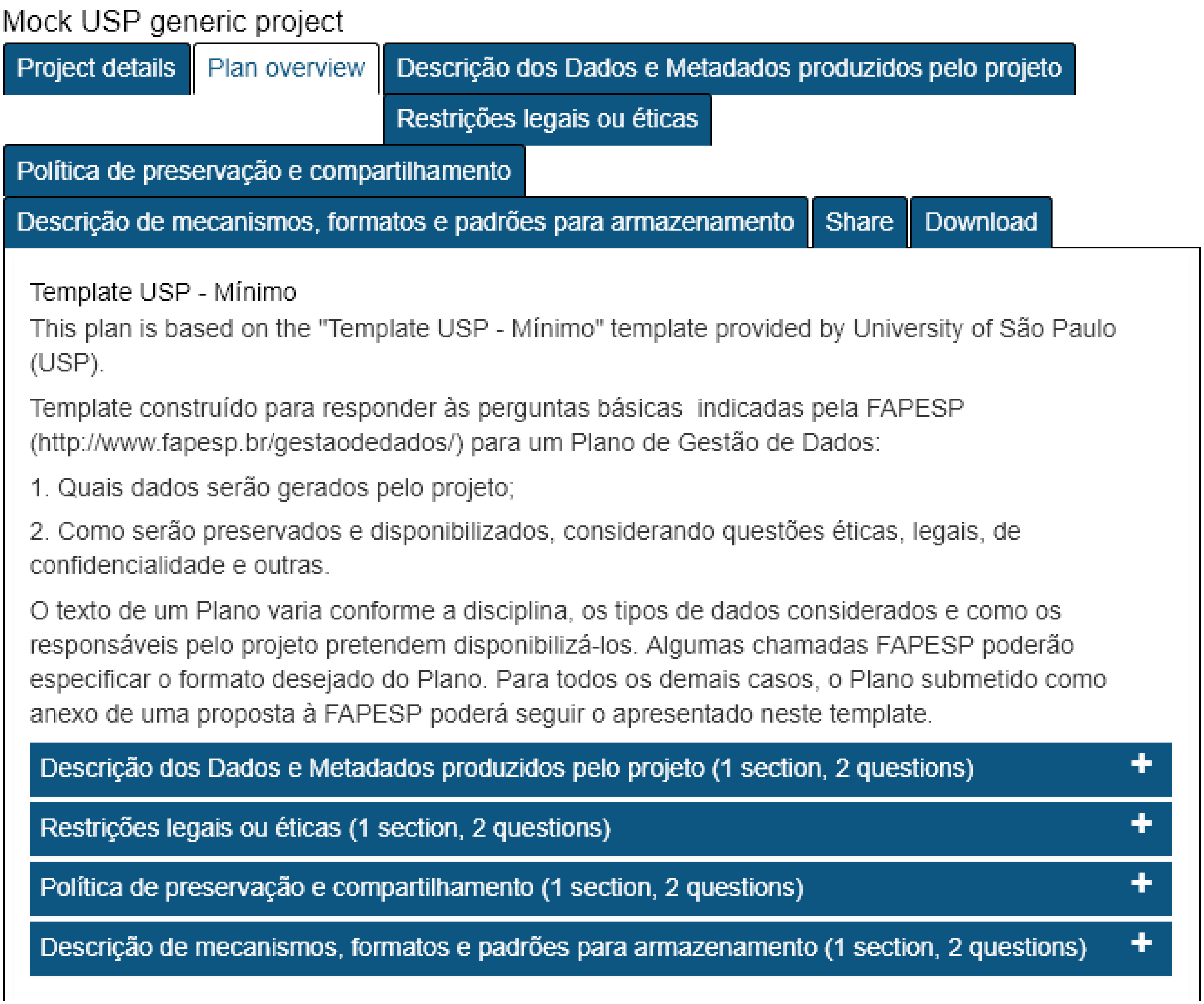
* Email

cmbm@ic.unicamp.br

* Password

•••••••••

□ Remember email


Project details
Plan overview
Descrição dos Dados e Metadados produzidos pelo projeto
Restrições legais ou éticas
Política de preservação e compartilhamento
Descrição de mecanismos, formatos e padrões para armazenamento
Share
Download

Mock USP generic project

Template USP - Mínimo

This plan is based on the "Template USP - Mínimo" template provided by University of São Paulo (USP).

Template construído para responder às perguntas básicas indicadas pela FAPESP (http://www.fapesp.br/gestaodedados/) para um Plano de Gestão de Dados:

1. Quais dados serão gerados pelo projeto;

2. Como serão preservados e disponibilizados, considerando questões éticas, legais, de confidencialidade e outras.

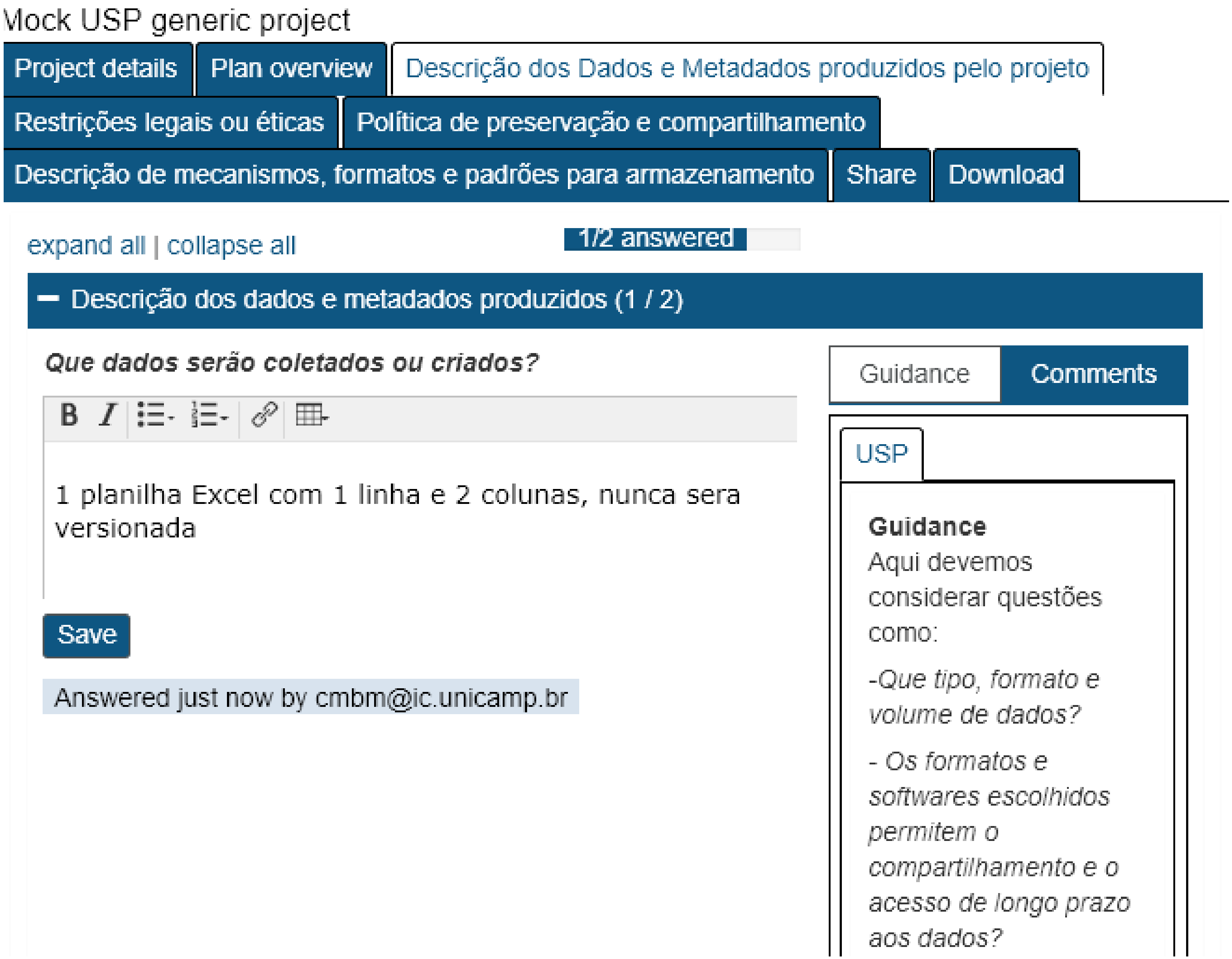
O texto de um Plano varia conforme a disciplina, os tipos de dados considerados e como os responsáveis pelo projeto pretendem disponibilizá-los. Algumas chamadas FAPESP poderão especificar o formato desejado do Plano. Para todos os demais casos, o Plano submetido como anexo de uma proposta à FAPESP poderá seguir o apresentado neste template.

Descrição dos Dados e Metadados produzidos pelo projeto (1 section, 2 questions) +

Restrições legais ou éticas (1 section, 2 questions) +

Política de preservação e compartilhamento (1 section, 2 questions) +

Descrição de mecanismos, formatos e padrões para armazenamento (1 section, 2 questions) +


Mock USP generic project
Project details
Plan overview
Descrição dos Dados e Metadados produzidos pelo projeto
Restrições legais ou éticas
Política de preservação e compartilhamento
Descrição de mecanismos, formatos e padrões para armazenamento
Share
Download
expand all | collapse all
1/2 answered
Guidance
Comments
USP
Save

− Descrição dos dados e metadados produzidos (1 / 2)

Que dados serão coletados ou criados?

1 planilha Excel com 1 linha e 2 colunas, nunca sera versionada

Answered just now by cmbm@ic.unicamp.br

Guidance
Aqui devemos considerar questões como:

-Que tipo, formato e volume de dados?

- Os formatos e softwares escolhidos permitem o compartilhamento e o acesso de longo prazo aos dados?

Main take-aways

- What is open science?
- What does OPEN mean?
- What is data?
- Why should I care?

What is data?

Any object in digital format, static or dynamic

What is the meaning of OPEN?

OPEN metadata in public repository

Why should I care?

Worldwide collaboration
Financing opportunities
Visibility of my work
Accelerating my research

Outline

Open Science -> Open MetaData

Definitions (Open Access, Open Data)

Challenges (Find, Identify, Understand)

(4) Challenges in (Research Data) Sharing

Open Data at FAPESP

Data Management Plans

Network of open research data repositories

ADVICE(S)

- OPEN YOUR SENSES (TO “DATA”)
- TALK TO ENTHUSIASTIC SCIENTISTS
- COST VS VALUE
- RESEARCH

= **WHAT IF**

= PERTURBATION ON STATUS QUO

= one object, countless opportunities

OBRIGADA