Training young researchers for an open future

Trieste, 8th of July 2015

Introduction to open research data

Ignasi Labastida Juan
Oficina de Difusió del Coneixement
CRAI Universitat de Barcelona

context

how we arrived here

Data deluge

Examples of big data:

The LHC experiments produce about 15 petabytes of raw data each year

200 Gb for the human genome sequencer

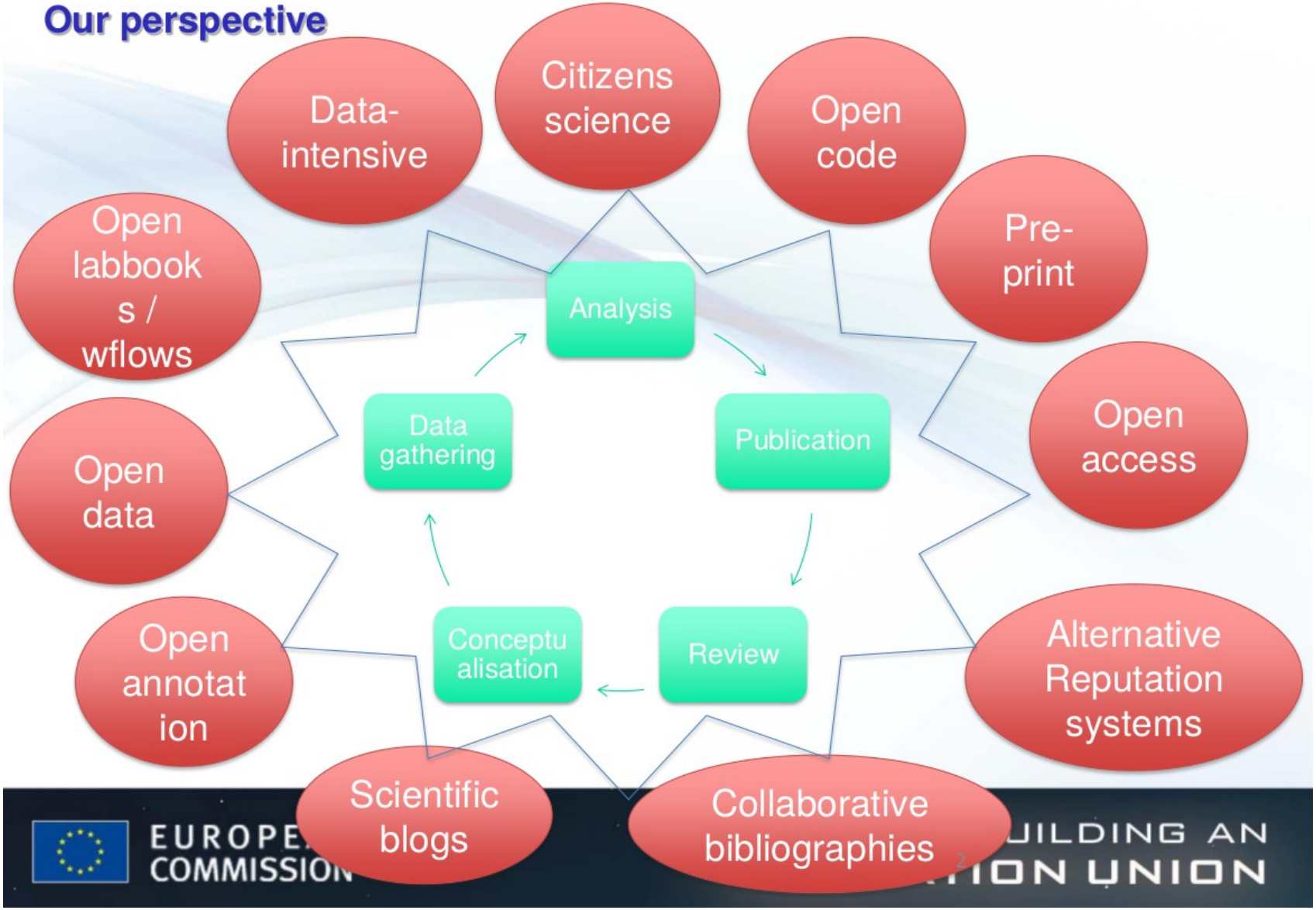
Any researcher collects or generate data at any level

FOSTER

open movement

- open access declarations
- open content licenses
- open educational resources
- open data
- open government
- open science
- open research

FOSTER



J. C. Burgelman, European Commission, Science 2.0

FOSTER

openness

looking for a definition

Defining Openness

"Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)."

"Open data and content can be freely used, modified, and shared by anyone for any purpose"

http://opendefinition.org/

FOSTER

According to the BOAI definition

"By *open access* to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited"

FOSTER

research data

looking for a definition

Definitions of research data

the recorded factual material commonly accepted in the scientific community as necessary to validate research findings (NCSU)

data in the form of facts, observations, images, computer program results, recordings, measurements or experiences on which an argument, theory, test or hypothesis, or another research output is based (ANDS)

That which is collected, observed, or created in a digital form, for purposes of analysing to produce original research results (UoE)

FOSTER

Types of research data

- collected
- observed
- created

- raw
- cleaned
- processed

FOSTER

research data policies

recommendations and requirements

research data policies

- Institutional
- National
- International

- where to store data
- which data
- how to share data

FOSTER

UCL research data policy

UCL considers the research data generated by its members as a valuable research output, an asset to the institution and a critical contribution to the knowledge economy.

The purpose of this Policy is to provide a framework to define the responsibilities of all UCL members and to guide researchers and students in how to manage the data, enabling research data to be maintained and preserved as a first class research object and made available to widest possible audience for the highest possible impact.

FOSTER

UCL research data policy

This policy is intended to ensure that research data created as part of the research process are:

- Accurate, complete, authentic and reliable;
- Attributable and citable;
- Identifiable, retrievable and available with minimal barriers;
- Secure from loss and degradation;
- Retained for a minimum of ten years after publication or public release;
- Compliant with legal obligations, ethical responsibilities and the rules of funding bodies.

FOSTER

RCUK common principles on data policy

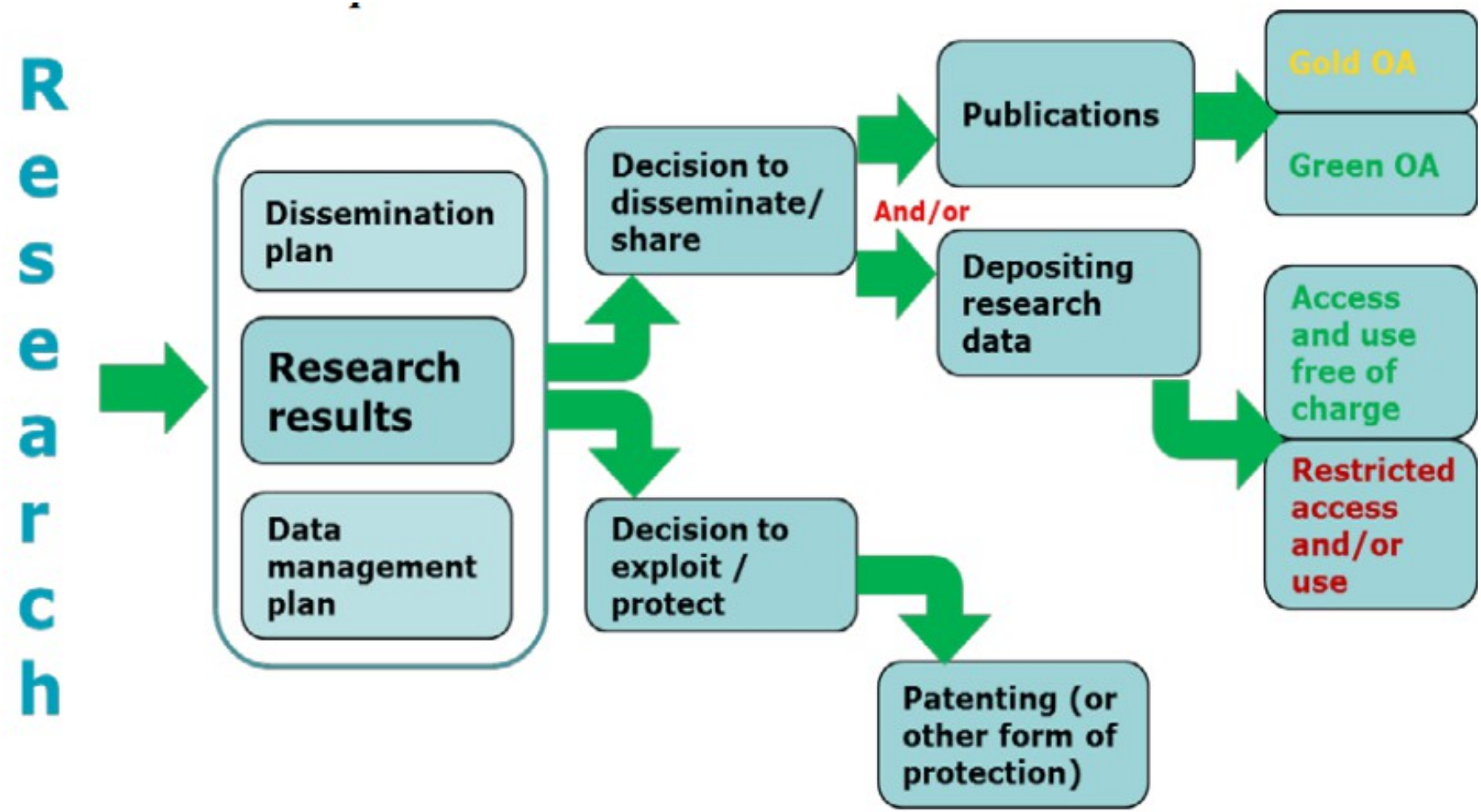
- Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner.
- Institutional and project specific data management policies and plans should be in accordance with relevant standards and community best practice. Data with acknowledged long-term value should be preserved and remain accessible and usable for future research.
- To enable research data to be discoverable and effectively re-used by others, sufficient metadata should be recorded and made openly available to enable other researchers to understand the research and re-use potential of the data. Published results should always include information on how to access the supporting data.
- RCUK recognises that there are legal, ethical and commercial constraints on release of research data. To ensure that the research process is not damaged by inappropriate release of data, research organisation policies and practices should ensure that these are considered at all stages in the research process.

FOSTER

RCUK common principles on data policy

- To ensure that research teams get appropriate recognition for the effort involved in collecting and analysing data, those who undertake Research Council funded work may be entitled to a limited period of privileged use of the data they have collected to enable them to publish the results of their research. The length of this period varies by research discipline and, where appropriate, is discussed further in the published policies of individual Research Councils.
- In order to recognise the intellectual contributions of researchers who generate, preserve and share key research datasets, all users of research data should acknowledge the sources of their data and abide by the terms and conditions under which they are accessed.
- It is appropriate to use public funds to support the management and sharing of publicly-funded research data. To maximise the research benefit which can be gained from limited budgets, the mechanisms for these activities should be both efficient and cost-effective in the use of public funds.

horizon 2020



Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

FOSTER

data pilot at horizon 2020

- the data, including associated metadata, needed to validate the results presented in scientific publications;
- other data, including associated metadata, as specified and within the deadlines laid down in the data management plan.

Data management plan:

- early deliverable within six months of the project

Two steps: to deposit the research data into a research data repository and to enable reuse by using licenses (CC BY/CC0)

Costs relating to the implementation of the pilot will be eligible

FOSTER

data pilot adressed at

- Future and Emerging Technologies
- Research infrastructures – part e-Infrastructures
- Leadership in enabling and industrial technologies – Information and Communication Technologies
- Societal Challenge: 'Secure, Clean and Efficient Energy' – part Smart cities and communities
- Societal Challenge: 'Climate Action, Environment, Resource Efficiency and Raw materials' – except raw materials
- Societal Challenge: 'Europe in a changing world – inclusive, innovative and reflective Societies'
- Science with and for Society

FOSTER

Reasons to opt out

- incompatible with the Horizon2020 obligation to protect results if they can reasonably be expected to be commercially or industrially exploited;
- incompatible with the need for confidentiality in connection with security issues;
- incompatible with existing rules concerning the protection of personal data;
- would jeopardise the achievement of the main aim of the action;
- will not generate / collect any research data;
- other legitimate reason to not take part in the Pilot

FOSTER

Publishers data policy

- PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception.
- Copernicus Publications recommends depositing data that correspond to journal articles in reliable data repositories, assigning digital object identifiers, and properly citing data sets as individual contributions.

FOSTER

PLOS policy

- All data and related metadata underlying the findings reported in a submitted manuscript should be deposited in an appropriate public repository, unless already provided as part of the submitted article.

http://journals.plos.org/plosone/s/data-availability

FOSTER

managing research data

what to do with data

to take into account

- how to describe research data
- where to store data
- how to cite data

- planning

FOSTER

describing research data

- Title
- Description
- Creator/Collector
- Geographic location
- Time reference
- Terms of use

FOSTER

storing research data

- locally
- externally

and posting?

http://www.re3data.org/

FOSTER

Trusted repositories?

Five principles from the Data Seal of Approval

- Data can be found on the Internet
- Data are accessible, while taking into account relevant legislation with regard to personal information and intellectual property of the data
- Data are available in a usable format
- Data are reliable
- Data can be referred to

  http://datasealofapproval.org/en/information/guidelines/

FOSTER

Liability for managing and sharing research data

who is liable for what

Recommendations and Roadpmaps

- RECODE

  http://recodeproject.eu/

- Roadmap for Research Data from the League of European Research Universities (LERU)

  http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf

FOSTER

Stakeholdres

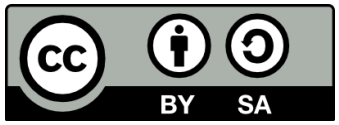
- Funders
- Higher Education/Research Institutions
- Data managers
- Publishers
- Researchers

FOSTER

opening research data

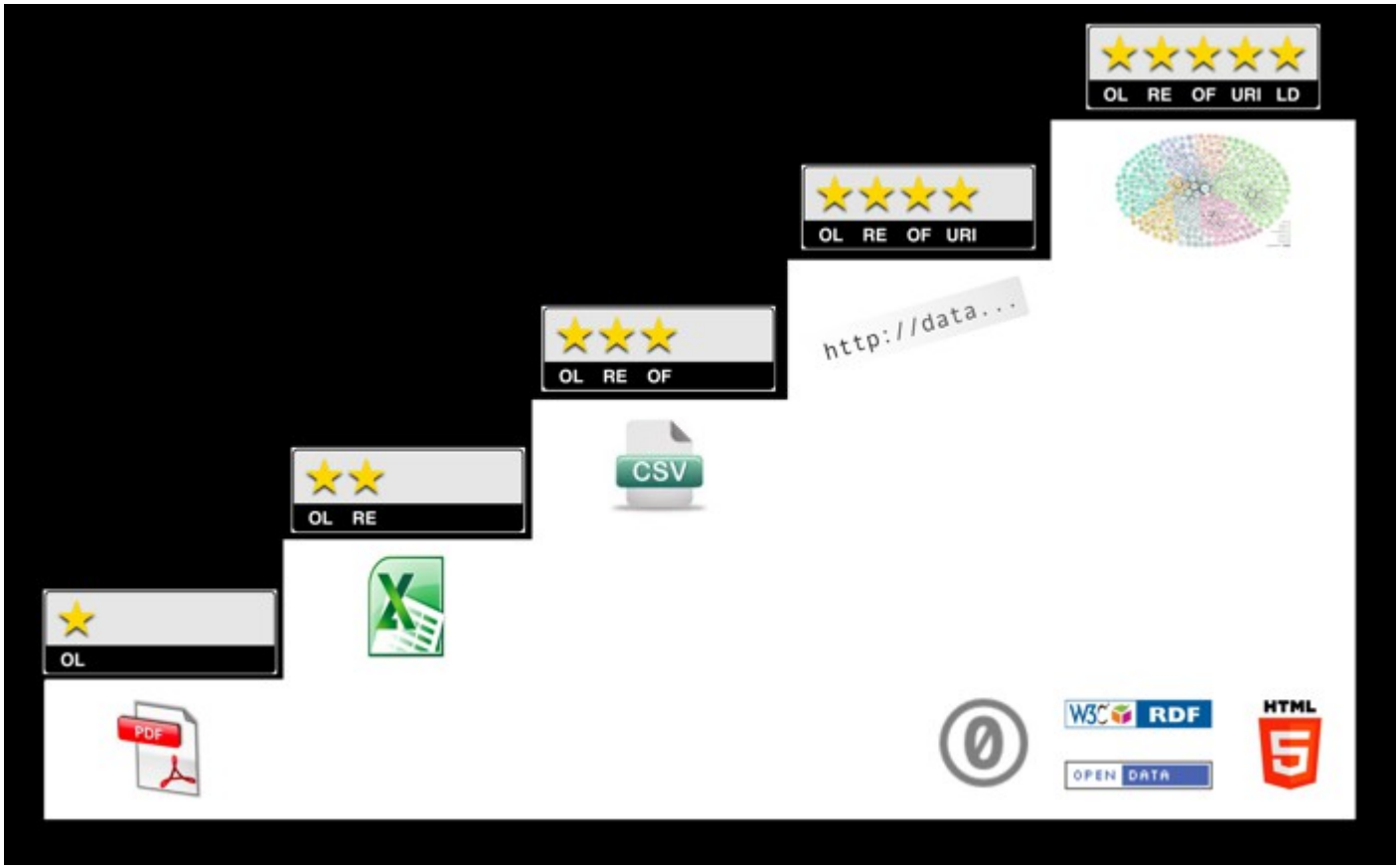
how to open it

opening research data



FOSTER

real open research data



http://5stardata.info/

FOSTER

"Open Data Excuse" Bingo #openDataExcuses

Lawyers want a custom License	Data Protection	Poor Quality	People may misinterpret the data
It's too big	There's no API	It's too complicated	Thieves will use it
It's not very interesting	There's already a project to...	We will get too many enquiries	What if we want to sell it later
We'll get spam	We might want to use it in a paper	I don't mind, but someone else might	Terrorists will use it

For open data teams; print out a copy and put it on your office wall. Cross out each excuse people give you. There are no prizes, but you can tweet "bingo! #openDataExcuses" if you think it might make you feel better*.

* it won't

Generate your own bingo grids at http://data.dev8d.org/devbingo/

some answers at http://gbonanome.github.io/opendatabingo/

FOSTER

Ignasi Labastida Juan
ilabastida@ub.edu
@ignasi