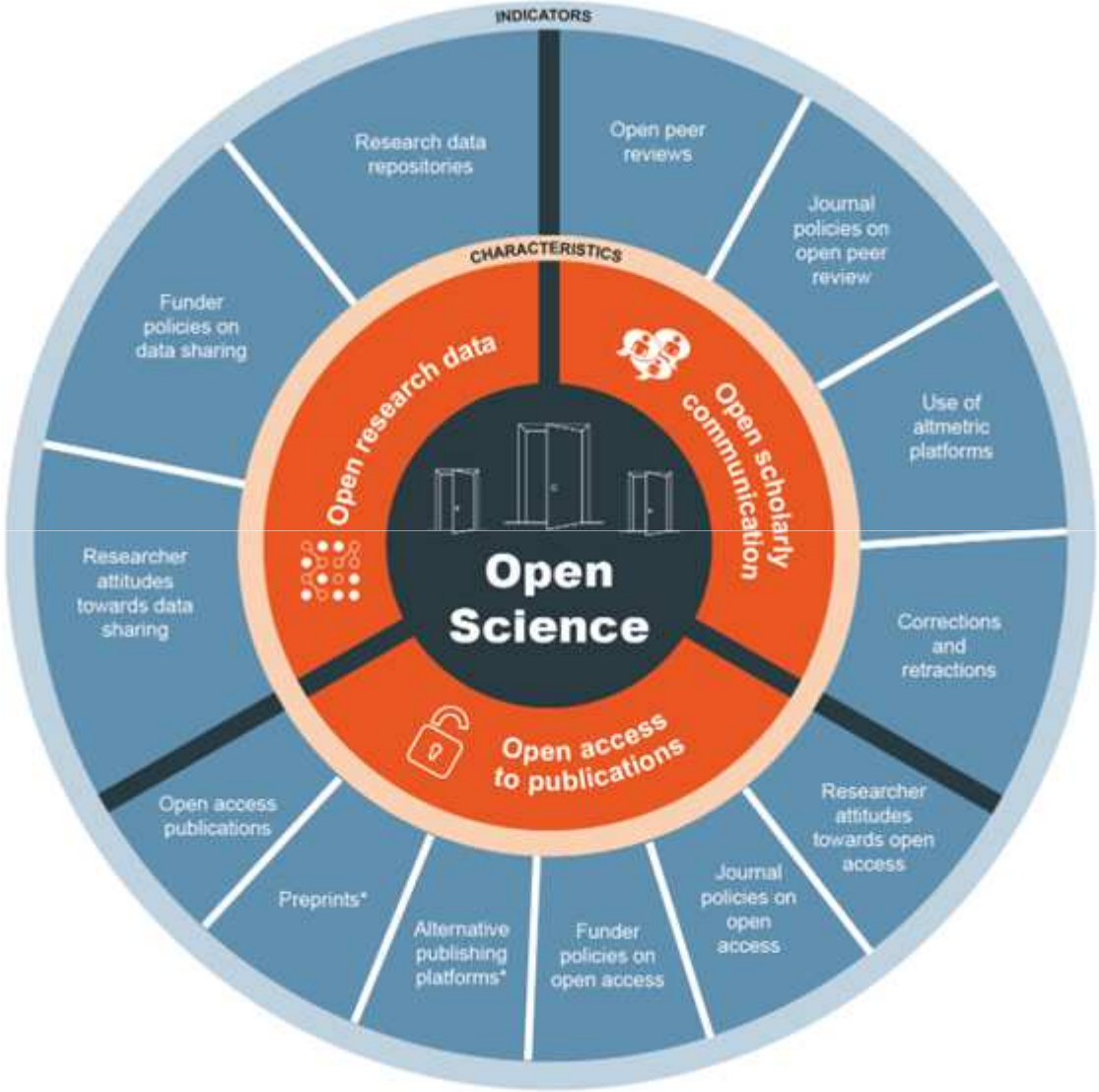
Corso di dottorato in Scienze Farmacologiche
Information literacy in Pharmacological Sciences 2017

Open research data
&
data archives

Emanuela Casson

26 may 2017

Corso di dottorato in Scienze Farmacologiche
mation literacy in Pharmacological Sciences 2017



Open Science Monitor – European Commission
http://ec.europa.eu/research/openscience/index.cfm?pg=home&section=monitor

Horizon 2020,
Work Programme 2016-2017,
Science with and for Society -
European Commission Decision
C(2017)2468 of 24 April 2017
http://ec.europa.eu/research/participants/data/ref/h2020/wp/2016_2017/main/h2020-wp1617-swfs_en.pdf

Openness

- Knowledge is open if **anyone** is free to **access**, **use**, **modify**, and **share** it - subject, at most, to measures that preserve provenance and openness
  http://opendefinition.org/od/

- A piece of data is open if **anyone** is free to **use**, **reuse**, and **redistribute** it - subject only, at most, to the requirement to attribute and/or share-alike
  http://en.wikipedia.org/wiki/Open_data

- By open data in science we mean that it is **freely** available on the **public internet** permitting any user to **download**, **copy**, **analyse**, **re-process**, **pass them to software** or **use them for any other purpose** without financial, legal, or technical barriers
  pantonprinciples.org/

Open research data - how

FAIR principles http://www.datafairport.org

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.

Open research data - why

Reasons for open research data:

- Moral: closed can be unjust
- Ethical: community norms expect it
- Utilitarian: greater communal good
- Personal: greater pesonal benefit
- Most scientific data is lost; costs many billions

… mandates?

What is “research data”?

Research data is data that is **collected**, **observed**, or **created**, for purposes of analysis to produce original research results.

A data classification based on:

- the mode of **production** or **generation** (observational, experimental, simulation, derived or compiled, reference)
- **formats** (text, numeric, multimedia, models, software languages...); es. chemistry data - preferred file format JCAMP
Es. UK Data Archive - file formats table
http://www.data-archive.ac.uk/create-manage/format/formats-table
- **content** (sociological surveys, DNA nucleotide sequences, algorithms...)

Each data category may require a different type of **data management plan (DMP)**

Data management plan (DMP)

"A formal statement describing how research data will be managed and documented throughout a research project and the terms regarding the subsequent deposit of the data with a data repository for long-term management and preservation"

RDC Glossary - Research Data Canada

DMP should include information on:

- the handling of research data **during** and **after** the end of the project
- what data will be **collected**, **processed** and/or **generated**
- which **methodology** and **standards** will be applied
- whether data will be **shared/made open access**
- how data will be **curated** and **preserved**
- data **protection** - insuring that sensitive data is not tampered with or stolen
- data **ownership** - the legal rights pertaining to your data.

DMP templates, exemples &tools

- Data Management Plan Templates and Examples / Thomas Harrod (George Washington University, Himmelfarb Health Sciences Library)
http://libguides.gwumc.edu/c.php?g=27812&p=170533

- Example DMPs and guidance / Digital Curation Centre
http://www.dcc.ac.uk/resources/data-management-plans/guidance-examples

- How to create a DMP Plan / OpenAire
https://www.openaire.eu/opendatapilot-dmp

- Horizon 2020 FAIR Data Management Plan (DMP) template
(Version 26 July 2016) Annex1, p. 6-12
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

- ERC Data Management Plan template (Version 1.0, 21 April 2017) https://erc.europa.eu/sites/default/files/document/file/ERC_DataManagementPlan_template.docx

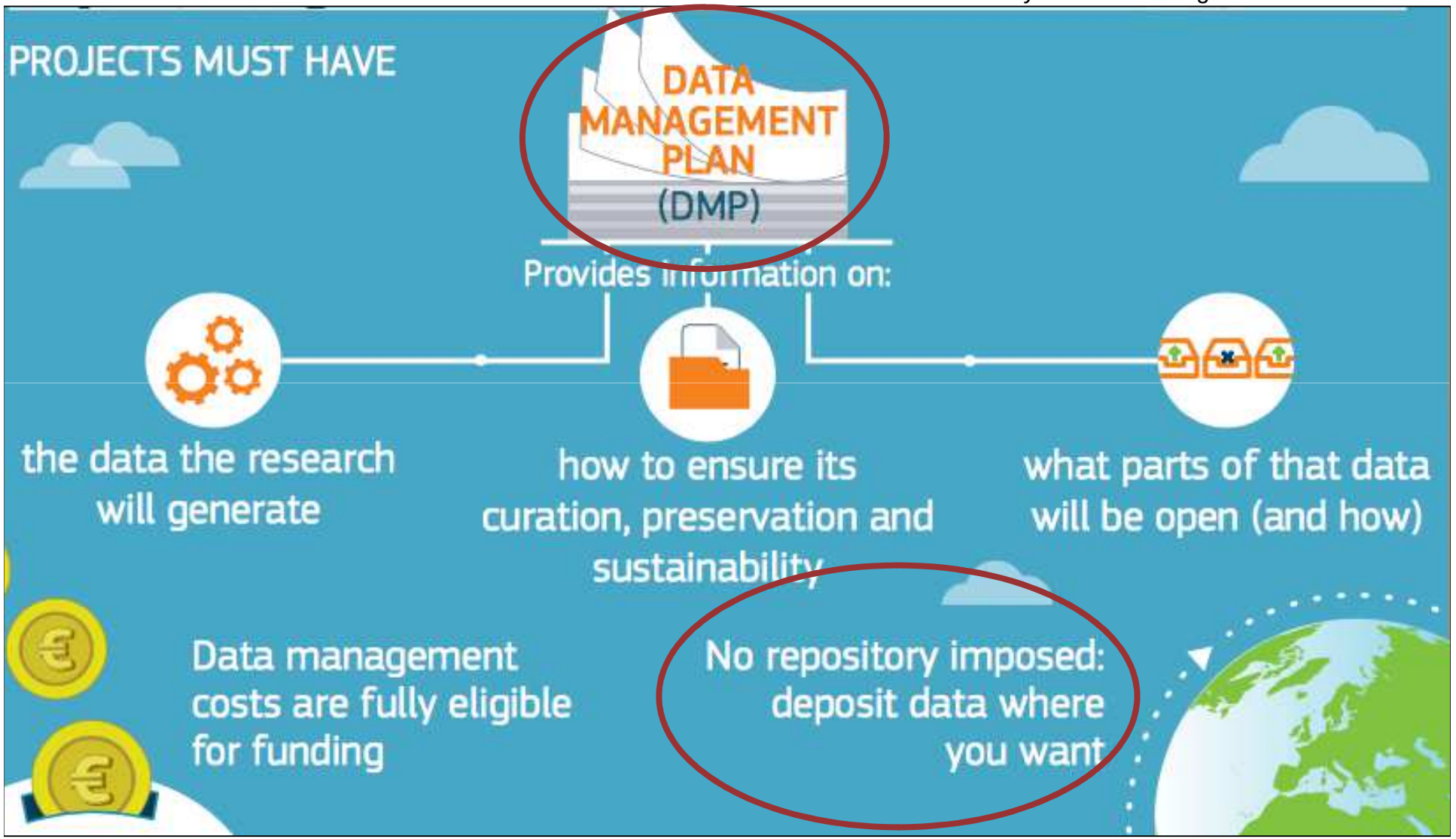
- DMPonline https://dmponline.dcc.ac.uk has been developed by the DCC - Digital Curation Centre to help write a DMP

ORD – Horizon2020 (1)

From 2017, research data is open by default, with possibilities to opt out

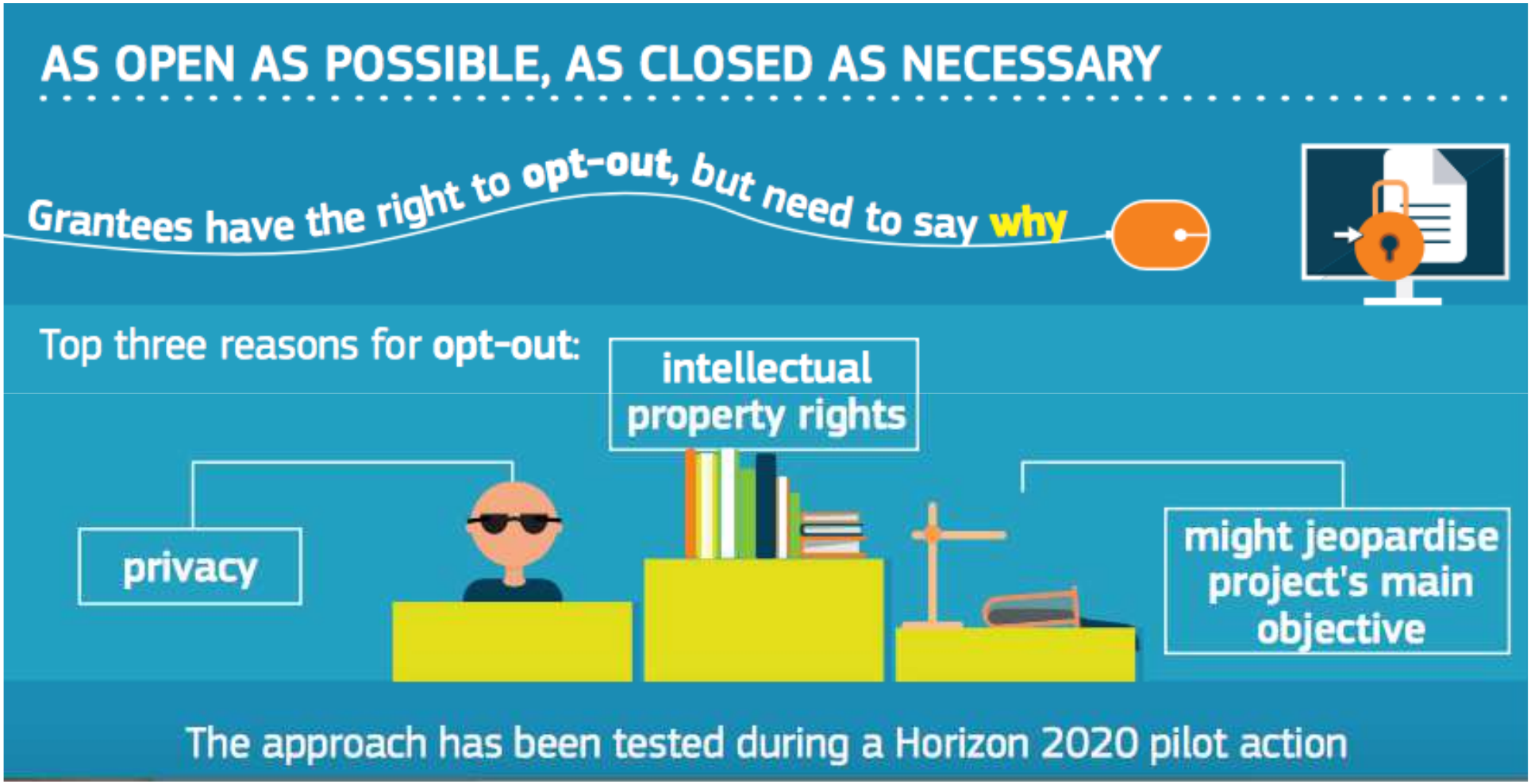
http://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf#view=fit&pagemode

ORD − Horizon2020 (2)



http://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf#view=fit&pagemode

ORD − Horizon2020 (3)



http://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf#view=fit&pagemode

Paola Gargiulo, OpenAIRE : aggiornamento sull'infrastruttura e strumenti a supporto della gestione dei dati della ricerca, www.slideshare.net/ marzo 2017

Research data repository (RDR)

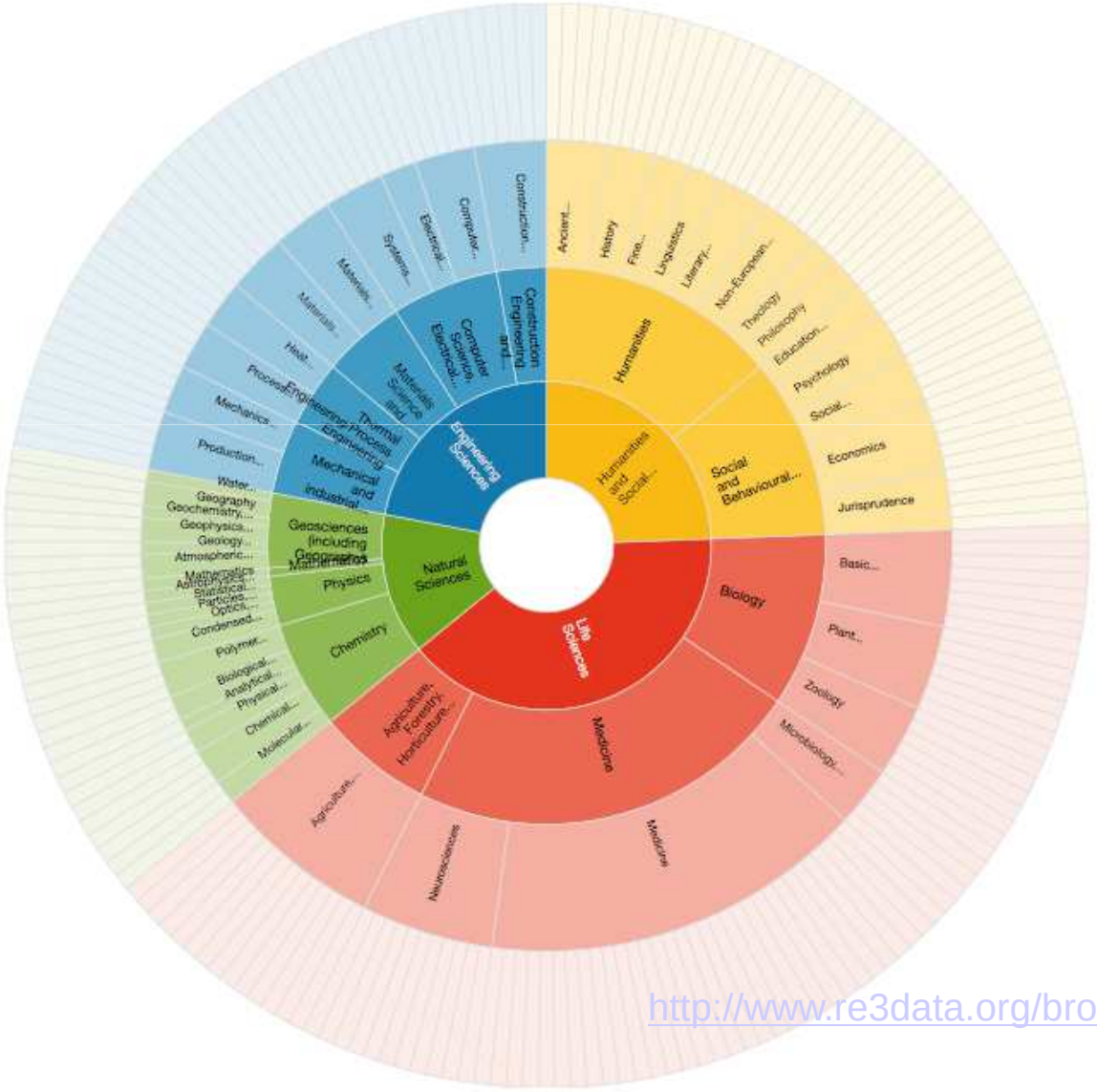
Research data must be deposited, preferably in a research data repository.

re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES **registry of research data repositories** http://www.re3data.org to help find a suitable discipline specific repository.
Re3data listes 1.500 research data repositories, making it the largest and most comprehensive registry of data repositories available on the web.

Vierkant, P., Spier, S., Ruecknagel … (2013). Schema for the description of research data repositories : version 2.1. re3data.org. https://doi.org/10.2312/re3.004

Re3data

Corso di dottorato in Scienze Farmacologiche
Information literacy in Pharmacological Sciences 2017



http://www.re3data.org/browse/by-subject/

RDR - Zenodo

Zenodo www.zenodo.org is a repository that:

- shares research results in a wide variety of formats including text, spreadsheets, audio, video, and images across all fields of science.
- displays the research results and gets credited by making the research results citable and integrating them into existing reporting lines to funding agencies like the European Commission.
- integrates their research outputs with the OpenAIRE portal
- a digital object identifier (DOI) is automatically assigned to all files
- integration between ResearcherID (WoS) and ORCID

RDR - PHAIDRA

Corso di dottorato in Scienze Farmacologiche
Information literacy in Pharmacological Sciences 2017

PHAIDRA
COLLEZIONI DIGITALI

PHAIDRA is the acronym for Permanent hosting, archiving and indexing of digital resources and assets. It is a digital asset management system with long-term archiving functions.
https://phaidra.cab.unipd.it

but also…

research data archive − FAIR Principles

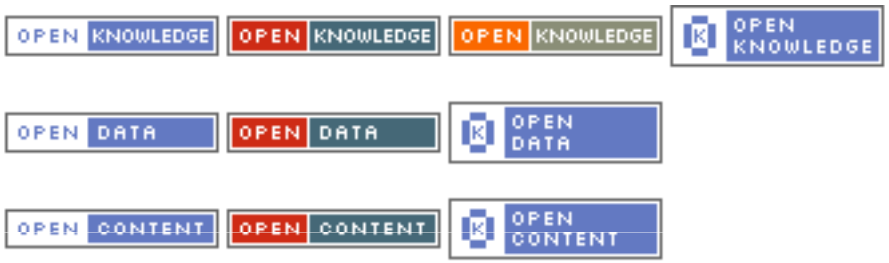
ORD Licenses

- Use a **recognized waiver** or **license** that is appropriate for data es. CC0 and PDDL. Non-commercial and other **restrictive clauses** should **not be used**

- CC0 – Creative Commons zero http://creativecommons.org/publicdomain/zero/1.0/
- PDDL - Public Domain Dedication and License http://opendatacommons.org/licenses/pddl/1.0/

ORD for users

- How do I know that data is Open Data?



- Are there any restrictions on what I can do with Open Data?

Citation of datasets

A data citation should include the following elements:

- **Author(s):** the creator(s) of the dataset. If extant, the creator should include a "nameIdentifier," such as an Open Researcher and Contributor ID (ORCID) or International Standard Name Identifier (ISNI)

- **Publication/Release date:** he date the dataset was made available or the date all quality assurance procedures were completed

- **Title:** the formal title of the data set

- **Version**: the precise version of the data used.

- **Publisher/Archive/Distributor:** the organization distributing or hosting the data, ideally over the long term

- **Identifier**: a unique string that identifies the resource; should be a persistent scheme such as a DOI (10.1234/8675309), handle,
or ARK(www.example.org/ark:/12345/lucky777).

- **Access Date** because data can changeable in ways that are not always reflected in release dates and versions, it is important to indicate when on-line data were accessed

Data, policies and journal publishers

Many journal publishers are asking authors to make the data underpinning a journal article available

- Nature Publishing Group (2009)
www.nature.com/authors/policies/availability.html

- PLoS (2014)
http://journals.plos.org/plosone/s/data-availability

- PNAS - Proceedings of the National Academy of Sciences (2009)
http://www.pnas.org/site/authors/editorialpolicies.xhtml#xi

- RS - Royal Society
https://royalsociety.org/journals/ethics-policies/data-sharing-mining/

- RSC – Royal Society of Chemistry
http://www.rsc.org/journals-books-databases/journal-authors-reviewers/prepare-your-article/experimental-data/

Data archives

Figshare and Dryad (multidisciplinary repositories)

- http://figshare.com

Figshare is an online digital repository where researchers can preserve and share their research outputs, including figures, datasets, images, and videos

- http://datadryad.org

Dryad hosts research data underlying scientific and medical publications. Historically, the repository has been strongest in the life sciences. All material in Dryad is associated with a scholarly publication.

Disciplinary data repositories

- **1000 Genomes**

http://www.internationalgenome.org

The 1000 Genomes Project is an public catalog of human genetic variation, including SNPs and structural variants, and their haplotype contexts. The genomes of about 2500 unidentified people from about 25 populations around the world will be sequenced using next-generation sequencing technologies.

Content type(s): Standard office documents, images, structured graphics, scientific and statistical data formats, raw data, plain text

- **freeBIRD - free Bank of Injury and Emergency Research Data**

https://ctu-app.lshtm.ac.uk/freebird/index.php/contact/

The FREEBIRD website aims to facilitate data sharing in the area of injury and emergency research in a timely and responsible manner. It has been launched by providing open access to anonymised data on over 30,000 injured patients (the CRASH-1 and CRASH-2 trials).

Content type(s): Standard office documents, scientific and statistical data formats, raw data

- **Global Initiative on Sharing Avian Influenza Data - GISAID**

EpiFlu database

http://platform.gisaid.org/epi3/frontend#

This platform is designed and maintained by scientists for scientists from various disciplines in influenza research, including veterinary and human virology, bioinformatics, epidemiology, immunology and clinical analysis etc.

Content type(s): Raw data

- **Hazardous Substance Data Bank - HSDB**

http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB

HSDB is a toxicology database that focuses on the toxicology of potentially hazardous chemicals. It provides information on human exposure, industrial hygiene, emergency handling procedures, environmental fate, regulatory requirements, nanomaterials, and related areas.

Content type(s): Images, plain text

- **JCB DataViewer**

http://jcb-dataviewer.rupress.org/jcb/page/contact/

The JCB DataViewer is an image hosting and presentation platform for original image datasets associated with articles published in The Journal of Cell Biology, (Rockefeller University Press)

Content type(s): Images

- **Open Phacts**

http://www.openphacts.org/index.php

The Open PHACTS consortium will reduce the barriers to drug discovery by applying semantic technologies to available data resources, creating an Open Pharmacological Space. Open PHACTS draws together multiple sources of publicly-available pharmacological and physicochemical data, accessible via the Open PHACTS Explorer

Content type(s): Standard office documents, networkbased data, images, structured graphics, audiovisual data, scientific and statistical data formats, raw data, plain text

- **SuperTarget**

http://insilico.charite.de/supertarget/index.php?site=home

SuperTarget is a database developed in the first place to collect informations about drug-target relations.

Content type(s): Scientific and statistical data formats, structured graphics databases, standard office documents, plain text

- **PharmGKB - Pharmacogenomics Knowledgebase**

https://www.pharmgkb.org/index.jsp

PharmGKB is a comprehensive resource that curates knowledge about the impact of genetic variation on drug response for clinicians and researchers. PharmGKB brings together the relevant data in a single place and adds value by combining disparate data on the same relationship, making it easier to search and easier to view the key aspects and by interpreting the data.PharmGKB provide clinical interpretations of this data, curated pathways and VIP summaries
which are not found elsewhere.

Content type(s): Scientific and statistical data formats, raw data, archived data, standard office documents

- **GenBank®**

http://www.ncbi.nlm.nih.gov/genbank/

GenBank is a comprehensive database that contains publicly available nucleotide sequences for almost 260 000 formally described species. NCBI places no restrictions on the use or distribution of the GenBank data.

Content type(s): Scientific and statistical data formats, images, networkbased data, structured graphics, plain text, software applications

- **Clinical data repository**

http://www.ctsi.umn.edu/researcher-resources/clinical-data-repository

The data in the Clinical Data Repository comes from the electronic health records (EHRs) of more than 2 million patients seen at 8 hospitals and more than 40 clinics. For each patient, data is available regarding the patient's demographics, medical history, problem list, allergies, immunizations, outpatient vitals, diagnoses, procedures, medications, lab tests, visit locations, providers, provider specialties, and more.

Content type(s): Standard office documents, databases standard office documents

Clinical trials

- Australian New Zealand Clinical Trials Registry (ANZCTR)
http://www.anzctr.org.au
- Chinese Clinical Trial Registry (ChiCTR)
http://www.chictr.org.cn/index.aspx
- Clinical Research Information Service (CRiS), Republic of Korea
http://cris.nih.go.kr/cris/en/use_guide/cris_introduce.jsp
- Clinical Trials Registry - India (CTRI)
http://ctri.nic.in/Clinicaltrials/login.php
- Cuban Public Registry of Clinical Trials (RPCEC)
http://registroclinico.sld.cu/en/home
- EU Clinical Trials Register (EU-CTR)
https://www.clinicaltrialsregister.eu/ctr-search/search
- German Clinical Trials Register (DRKS)
https://drks-neu.uniklinik-freiburg.de/drks_web/setLocale_EN.do
- Iranian Registry of Clinical Trials (IRCT)
http://irct.ir/
- ISRCTN registry
https://www.isrctn.com/
- Thai Clinical Trials Registry (TCTR)
http://www.clinicaltrials.in.th/
- The Netherlands National Trial Register (NTR)
http://www.trialregister.nl/trialreg/index.asp
- Pan African Clinical Trial Registry (PACTR)
http://www.pactr.org/
- Peruvian Clinical Trial Registry (REPEC)
http://bit.ly/2rzQ8Dm

Chemistry and chemical biology

- caNanoLab

https://cananolab.nci.nih.gov/caNanoLab/

This is a web-based application designed to facilitate data sharing in the research community to expedite and validate the use of nanomaterials in biomedicine

- PubChem

http://pubchem.ncbi.nlm.nih.gov

PubChem is a freely accessible database (FTP) that provides information about small molecules (contains 3 databases: PubChem BioAssay, PubChem Compound and PubChem Substance)

- ChemSpider

http://www.chemspider.com/

ChemSpider is a free chemical structure database providing fast access to over 58 million structures, properties and associated information

Crystallography open database

www.crystallography.net

Open-access collection of crystal structures of organic, inorganic, metal-organic compounds and minerals, excluding biopolymers

But Elsevier suggests:
Cambridge Crystallographic Data Centre (CCDC)
http://www.ccdc.cam.ac.uk/pages/Home.aspx