Open Data:
motivation, challenges and solutions

Eric Chassande-Mottin
CNRS/IN2P3 AstroParticule et Cosmologie

Motivations (1)

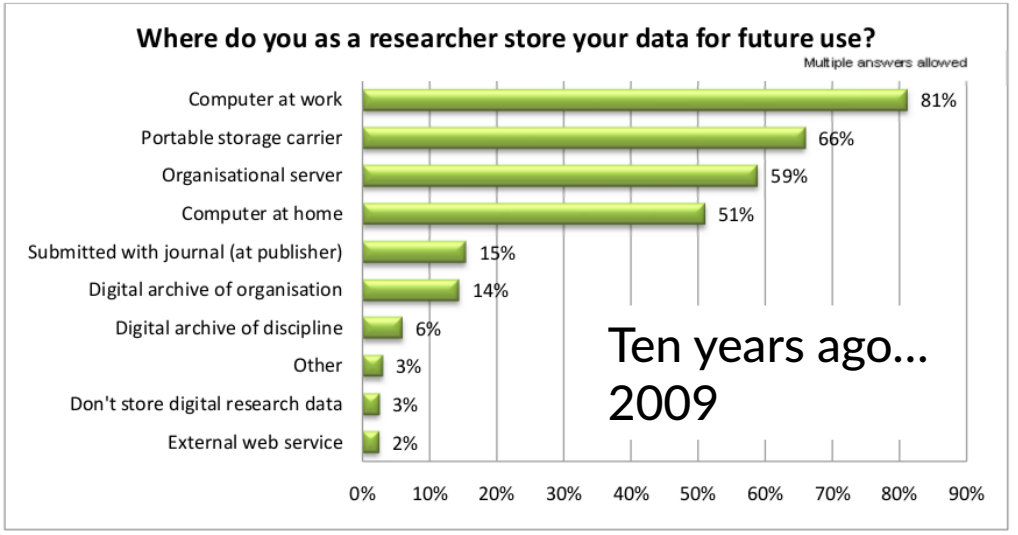


Figure 17: where researchers keep their data for future use, n = 1202

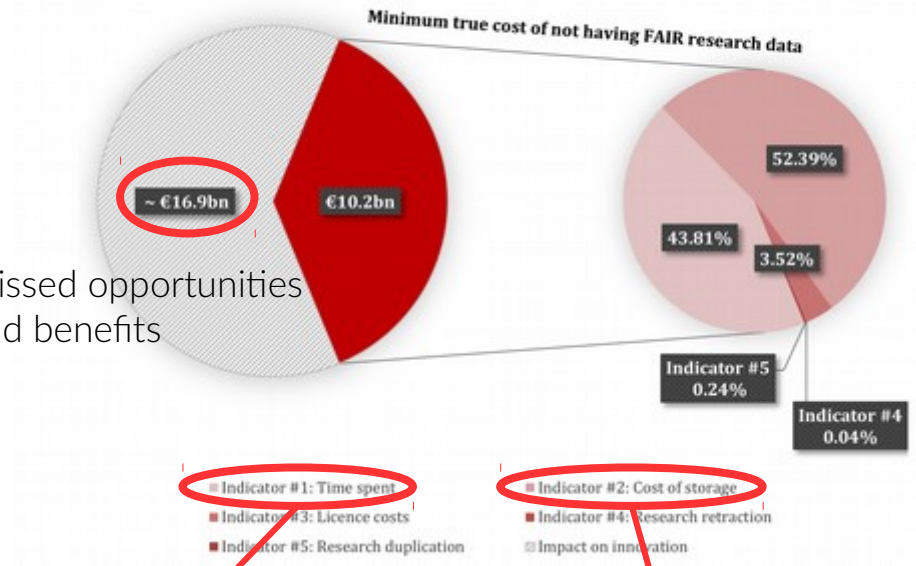
https://libereurope.eu/wp-content/uploads/2010/01/PARSE.Insight.-Deliverable-D3.4-Survey-Report.-of-research-output-Europe-Title-of-Deliverable-Survey-Report.pdf

https://arxiv.org/abs/0906.0485

Motivations (2)

Duplication of efforts has a cost

Source: https://publications.europa.eu/s/naPT



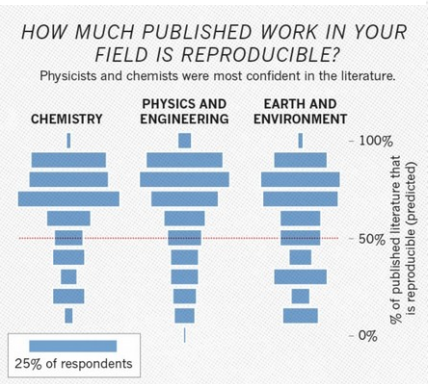
Missed opportunities and benefits

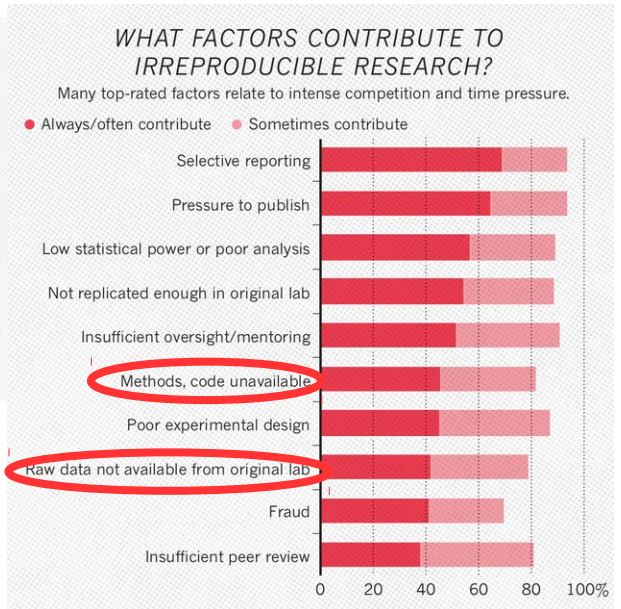
Time spent in reproducing past results after data loss or in duplicating what others have done

Duplication of data storage due to missing open archives

Reproducibility crisis?

Source: https://www.nature.com/news/1.19970





Raw data not available
Methods, code unavailable

1576 researchers surveyed

"**Reproducibility** is like brushing your teeth.
It is good for you, but it takes time and effort.
Once you learn it, it becomes a habit."

Open science initiatives – Worldwide

- 2003 – Berlin declaration on Open Access to Knowledge in the Sciences and Humanities
- 2004 – OECD: Declaration on Access to Research Data from Public Funding
- 2007 – OECD: Principles and Guidelines for Access to Research Data from Public Funding
- 2013 – G8 Science minister statement
  - publicly funded scientific research data should be open
  - discoverable, accessible, assessable, intelligible, useable, and [...] interoperable
  - recognition of researchers fulfilling these principles, and appropriate digital 2infrastructure
- 2016 – OECD: Open Science statement

OECD (2015), "Making Open Science a Reality", OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, Paris, https://doi.org/10.1787/5jrs2f963zs1-en.

At the European level

- Science Europe – http://www.scienceeurope.org
  Good practises, e.g., standardisation of research data management, etc.
  CNRS is *not* a member of Science Europe
- Open science policy platform – EU commission
  https://ec.europa.eu/research/openscience
  https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-policy-platform
- Open Access Infrastructure for Research in Europe
  https://www.openaire.eu – support open access/open data mandates in Europe
- European Open Science Cloud (EOSC) – Nov 2018
  https://www.eosc-portal.eu – Major initiative provide a public data repository which conforms to open science values
- Plan S – Towards open access to scientific publications

At the national level

- Loi pour une République numérique – oct 2016
- Plan pour la Science ouverte – juil 2018

  Généraliser l'accès ouvert aux publications

  Structurer et ouvrir les données de la recherche

  S'inscrire dans une dynamique durable, européenne et internationale

  cache.media.enseignementsup-recherche.gouv.fr/file/Actus/67/2/PLAN_NATIONAL_SCIENCE_OUVERTE_978672.pdf

- Comité pour la science ouverte – www.ouvrirlascience.fr

  Coordonne l'action national pour la mise en place du plan

  Présidé par Bernard Larrouturou (DGRI)

- Consortium Couperin – www.couperin.org

From abstract policies to real life... New legal obligations

- Mandatory to **publish articles and books in open access** resulting from publicly funded research
  - Obligation for projects supported by the ANR, Horizon 2020 and ERC
- Mandatory to **openly disseminate research data** from publicly funded programs
  - Requires data management plans in calls for research projects

    H2020: Open Research Data pilot for open access to research data

    ANR: “Les coordinateurs des projets financés à partir de 2019, devront fournir un Plan de Gestion des Données”

At the CNRS level

- Raising awareness within CNRS management

  Recent info meeting on open science, 8 oct 2019 – http://www.cnrs.fr/en/node/4133

  Direction de l'Information Scientifique et Technique du CNRS – Sylvie Rousset

- First steps towards a policy implementation

  Publications: requirement for open access to HAL

  Objective: all CNRS publications in open access by 2023

  Data: no global plan yet. Initial discussions at institute level

- Large infrastructure – TGIR

  So far, no regulation or binding policy coming from CNRS or ministry

  Data release left to project appreciation

Data sharing = culture change

- Sharing data requires a change in the mind state
  - Private data has been the prevailing model in HEP collaborations for many years
  - Tensions with the way HEP experiments operates currently
- Sharing data requires new expertise and additional ressources
  - Learn good practises for:

    Data curation, documentation, provenance tracking, review
    Data release and dissemination, DOI
    Publish data paper, supporting software, tutorials
  - Requires additional manpower and dedicated training – New type of job
    Requires new services and web infrastructure to allow data distribution

    Financial support should be integrated in the experiment budget since its inception

Open data ‘how-to’ (1)

Intergovernmental initiative
France, Germany and the Netherlands
https://www.go-fair.org

https://www.rd-alliance.org

- Good practises and basic principles

  FAIR – Findable, Accessible, Interoperable, Reusable
  http://www.nature.com/articles/sdata201618

  ‘FAIRification’ process well documented

- Get help and experience sharing

  Research data alliance provides recommendations and organizes forums and workshops

  Journées de la science ouverte – https://jnso2019.sciencesconf.org

- Get support

  ANR Appel Flash 2019 – Call in 2020 ?

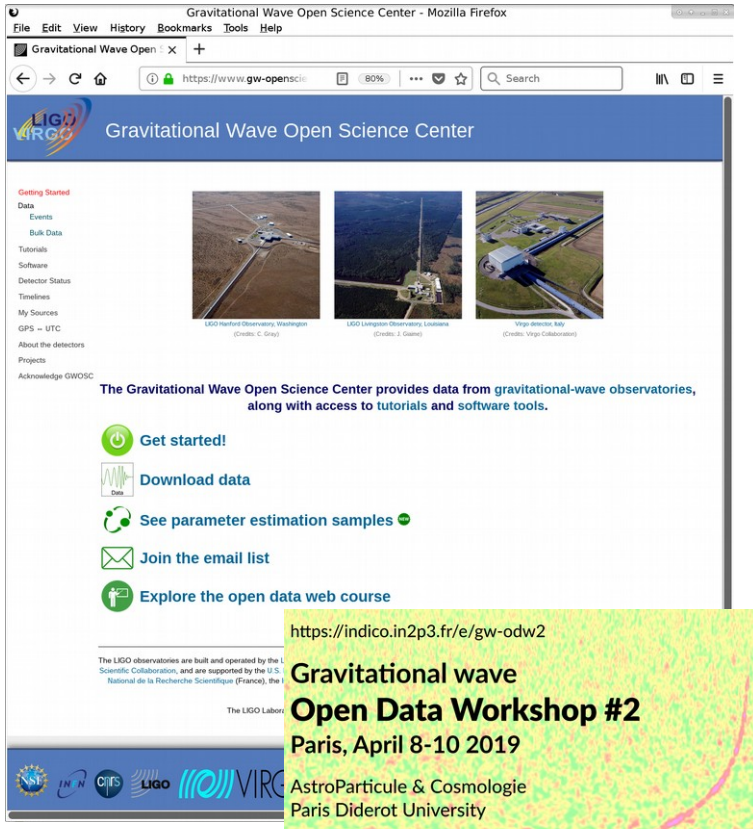
Open data ‘how-to’ (2)

- Open science platforms
  - **Dataverse (Harvard) https://dataverse.harvard.edu**

    Open source web app to share, preserve, cite, explore, and analyze research data
  - **Zenodo (CERN) https://zenodo.org**

    CERN Data Centre-backed research data repository
    Provides citeable discoverable links to data – Link to GitHub
- EU and national platforms in progress

gw-openscience.org



"Only days after they announced the first observation of gravitational waves, the LIGO and Virgo collaborations made public their data."

- Whole science-run data and GW event catalogs
  - ✔ Downloads: 60 TB/week peak
  - ✔ 80+ papers using open data
- Documentation, usage recommendation
- Online training: video tutorials and Jupyter notebooks

1.5 FTE (Virgo contribution)

Take-home messages

"Digital information lasts forever, or for the next five years, whichever comes first."
Jeff Rothenberg (RAND)

Open science is happening: **paradigm shift**

- Significant change in the way we do science
- Big push from EU and government leading to **legal obligations**

Opening research data

- Open by default – As open as possible, as closed as necessary

Good practises and tools available

CNRS/IN2P3 is a major stakeholder

- Preserve large and precious datasets from large-scale expensive experiments

Main limitations today

- **No institute-wide strategy**
- **No funding scheme** integrated with experiment design
- Requires **trained staff** with a career perspective: new type of job with specific know-how