Open Data Publication

How to overcome obstacles to data publication: Issues, requirements, and good practice

Pilsen, 4 September 2013

Guntram Geser

Salzburg Research

ARIADNE is funded by the European Commission's Seventh Framework Programme

Main topics

- Open data – criteria and drivers
- Current non-open behaviours
- Benefits of open data publication

It‘s (not) about…

- It‘s not about data management for its own sake – the objective is making available open data
- It‘s not about data management to comply with policies
- It‘s about benefits of *open data publication*

Open Data – criteria

- Accessible
  - Online, not necessarily without registration
- Reusable
  - not summarized data (i.e. figures, charts, etc.) canned in publications
  - state: raw, cleaned, normalized,…?
  - open format (e.g. not PDF doc)
- Openly licensed (e.g. CC-BY, if other no NoDerivative!)
- For free – yes, but somebody has to pay to ensure sustainability

"Publishing data in a reusable form to support findings must be mandatory" – one of six key areas for action highlighted in the The Royal Society's report *Science as an Open Enterprise* (2012)

Drivers for open data /1

- Expansion from Open Access research publications
  - Initially against rising costs of academic journals
  - Rather well established – "gold": OA journals, "green": self-archiving
- Expansion from „data-intensive" showcase disciplines
  - also „big data" or „data-driven", e.g. astronomy, molecular biology („omics")
  - Cf. High-level Group on Scientific Data *„Riding the wave"* report (2010)

Neelie Kroes, EC Vice-President: *"Taxpayers should not have to pay twice for scientific research and they need seamless access to raw data."*

An argument for e-infrastructures !

What about archaeology ?

Drivers /2

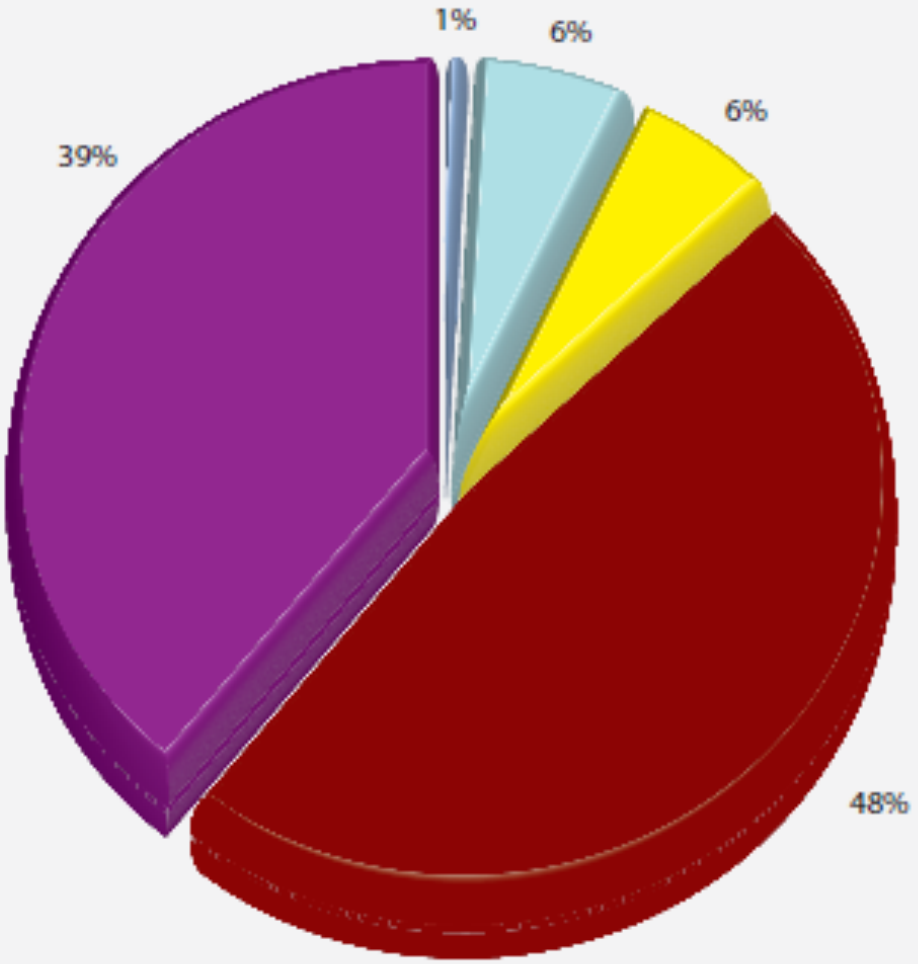
- High-level policies & initiatives
  - OECD: Declaration on Access to Research Data from Public Funding (2004; Principles and Guidelines, 2007)
  - EC Communications: Open data (2011); Towards better access to scientific information (2012)
  - Many others, most recent: Research Data Alliance – international initiative (launched in March 2013), various working & interest groups (archaeology not represented yet)
- Research funding agencies
  - Open Access mandates extended to data
  - Mandatory data management plans

Drivers /3

- Data archiving & access infrastructures put in place
  - Data centres / repositories
    - General: DRYAD, zenodo (related to OpenAIRE), …
    - Archaeology: ADS (UK), eDNA (NL), mappa (IT), tDAR (USA), …
  - Data catalogues, search & access services
  - Data citation standard, e.g. DataCite
- New publication formats
  - „Data Journals“, „Data Papers“ – describe a dataset/DB and its usefulness for research
  - Examples in archaeology
    - Journal of Open Archaeology Data, started 2012
    - Internet Archaeology, started publishing data papers in 2013

EC 2012 survey „Do you agree with the following statement: Generally speaking, there is NO access problem to research data in Europe?"

European Commission: *Online survey on scientific information in the digital age*; Total survey participants: 1140. Germany: 422, France: 120, UK: 127, Italy: 95, NL: 39, Austria: 38, Belgium: 36, Greece: 27, …. (42 countries); N below =?



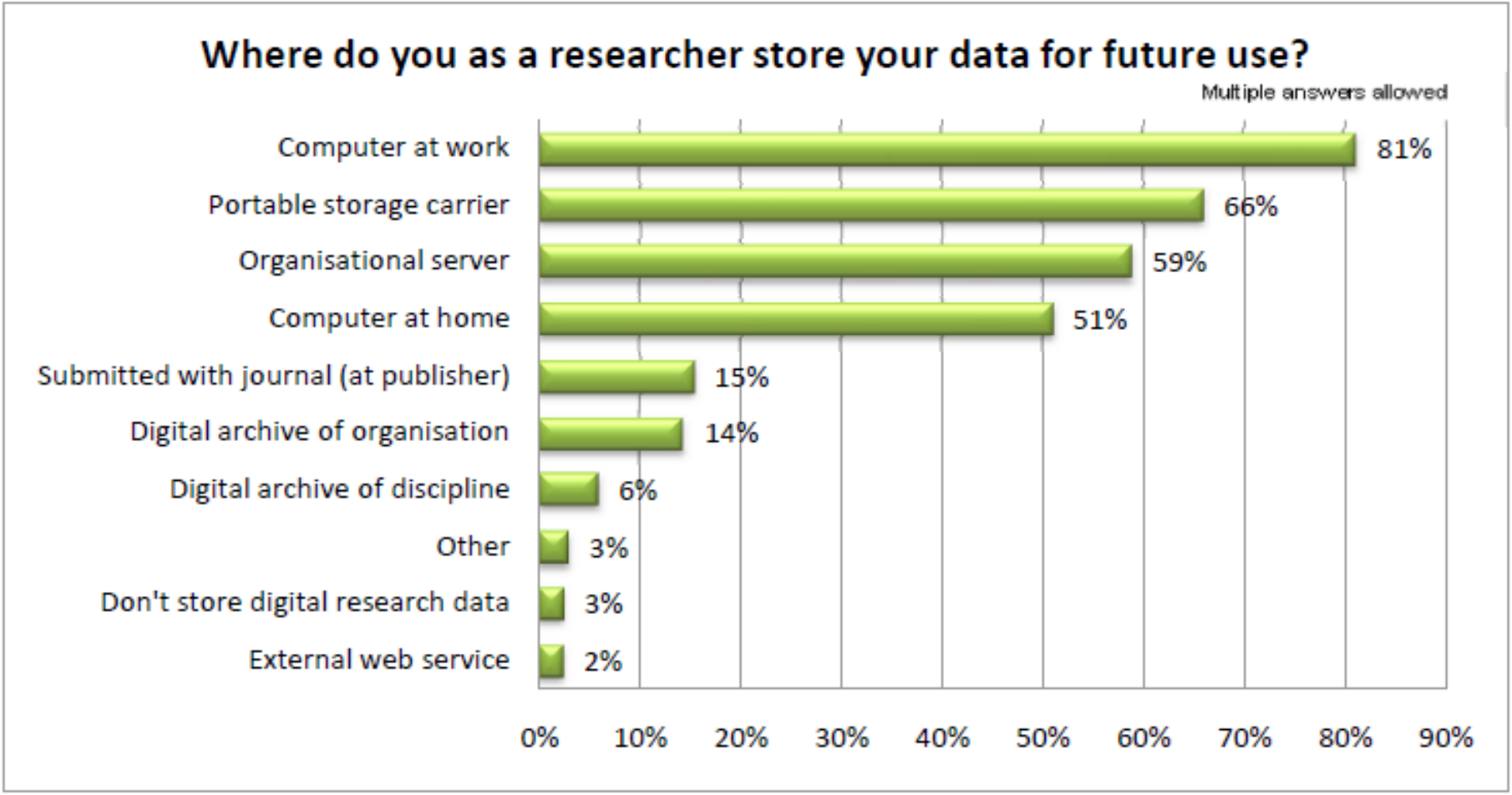
87% „Disagree" or „Disagree strongly"

Why the „access problem“

- Behaviour of researchers contrary to what advocates of proper management and sharing of data would like them to do
- Most re-useable data remains locked away
  - On personal computers
  - Portable storage carriers
  - Restricted access servers
  - …

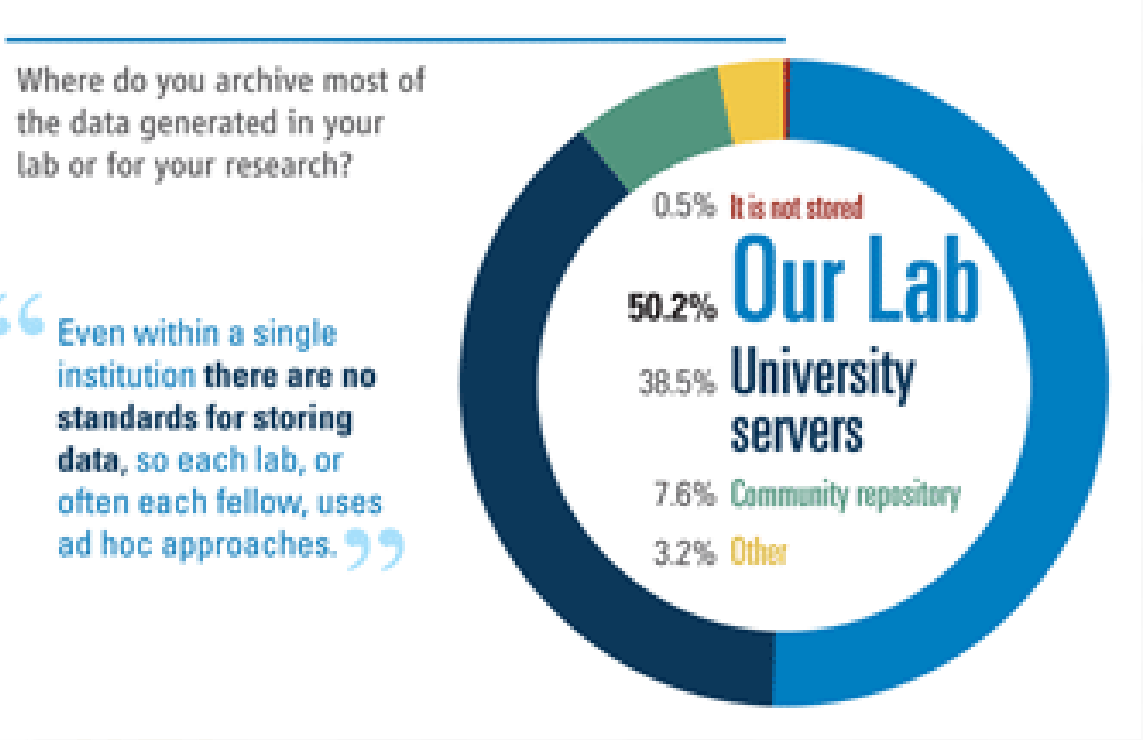
Where do researchers store/archive data?

PARSE.Insight survey 2009: 1202 respondents from different research domains and countries



Where do researchers store/archive data?

- "Science" journal 2011 survey of peer reviewers: 1700 responses, international and multi-disciplinary
- *"Where do you archive most of the data generated in your lab or for your research?"*



50.2% in our lab
38.5% university server
7.6% community repository
3.2% "other"
0.5% not stored

Note: archived ≠ curated

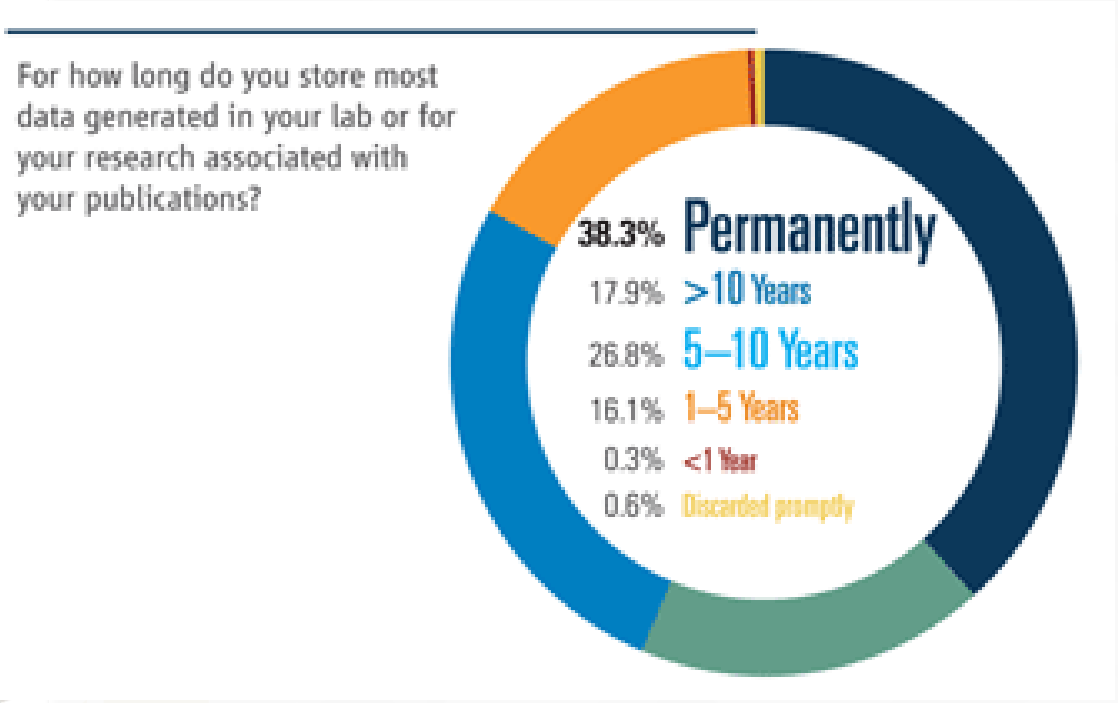
Data value & shelf life*

- Data value – perspective of individual researchers
  - understood as an asset to be exploited
  - loses value when papers are published
  - data unlikely to allow for new insights and publications
  - change of research focus, etc.
- Then the data becomes “obsolete”, remains on PCs, carrier media, servers… eventually discarded or otherwise lost
- Often not considered: potential value of the data for other, alternate, new uses, e.g. when combined with other available data

* Timeframe in which information depletes in relevance to their potential users

Stored for how long?

- "Science" (journal) 2011 survey of peer reviewers – 1700 responses, international and multi-disciplinary
- *"For how long do you store most data generated in your lab or for your research associated with your publications?"*



38.3% Permanently
17.9% > 10 years
26.8% 5-10 years
16.1% 1-5 years
0.3% > 1 year
0.6% Discarded promptly

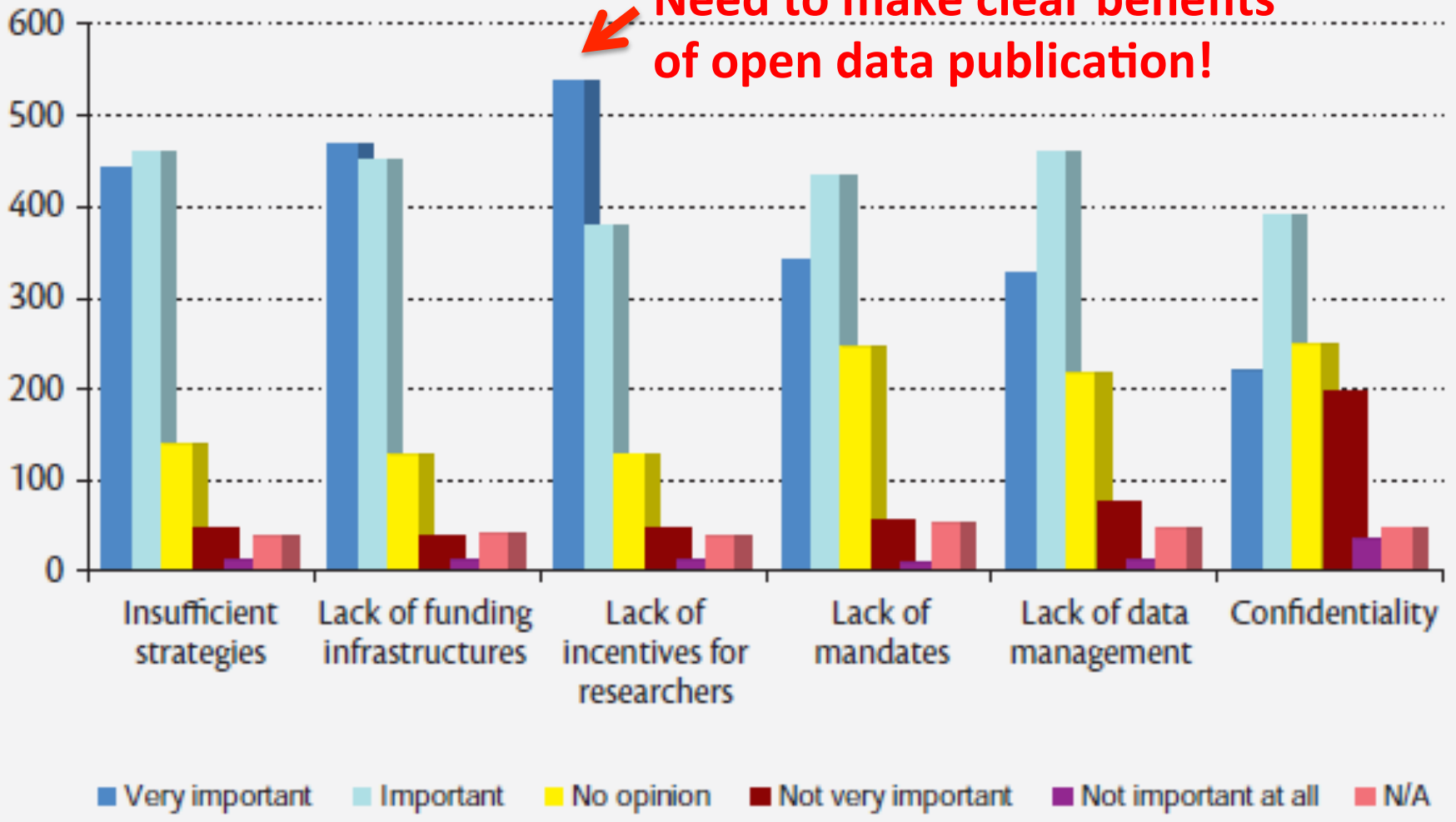
Note: stored ≠ curated

Reasons for lack of data sharing

- Many obstacles/barriers to providing open access to reusable data
  - Priority of published papers / little academic reward for development and sharing of datasets/DB
  - Existing copyrights, confidential and sensitive data
  - Concerns of researchers that data could be scooped, misused or misinterpreted
  - Potential reputational risk (e.g. data quality, errors,…)
  - Required effort to share re-usable data, incl. formatting, metadata creation, licensing etc.
  - Perceived lack of appropriate data archives (trusted, sustainable, ...)

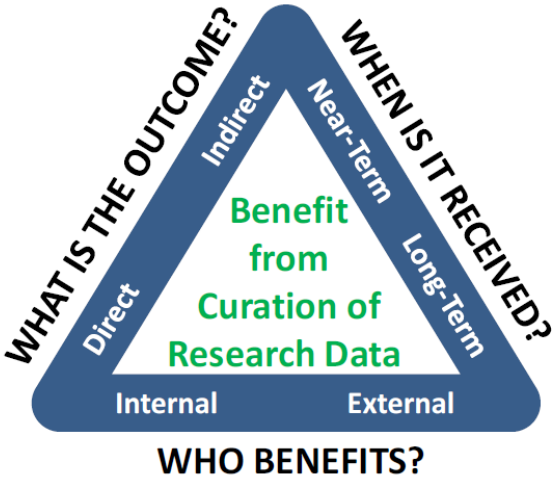
EC 2012 survey „How would you rate the importance of the following potential barriers to enhancing access to research data?"

Total survey participants: 1140. Germany: 422, France: 120, UK: 127, Italy: 95, NL: 39, Austria: 38, Belgium: 36, Greece: 27, …. (42 countries); N below =?



Examples of benefits

- Charles Beagrie Ltd: Keeping Research Data Safe (KRDS) benefits framework
- Some 30 examples of benefits for researchers, institutions, society:
  - Scholarly communication/access to data
  - Verification of research/research integrity
  - Increased visibility/citation
  - Motivating/input for new research
  - Stimulating new networks/collaborations
  - Re-use/-purposing of well curated data
  - No re-creation of data
  - No data lost from Post Doc turnover
  - …



Authors‘ benefits focus /1

- Recognition and academic reward for data providers – at least same as for other publications (maybe more)
- Core mechanism = citation of published data/set
- Confirms value of the data contributed
- Makes identification of good data easier, and promotes further re-use/-purposing
- Allows the impact of the data to be tracked and measured (citation metrics)

Authors‘ benefits focus /2

- Open data – longer shelf life
  - Data that is accessible, used and enriched by a research community gains in value
  - Consequently it will be kept on the shelf and curated for long-term access
- Authors and archives are partners – archives need to demonstrate relevance, ensure funding

How to reap the benefits? / 1

- Deposit reusable data in a community recognised and reliable repository
  - See Data Seal of Approval; Trusted Repositories Audit & Certification (TRAC) and other checklists
  - Should provide unique persistent identifiers (e.g. DOIs)
  - Require following citation standard as part of user agreement (e.g. DataCite; citation in reference list)
- Provide good metadata – “no pain, no gain”*
  - Key for data re-use without direct contact with creator
  - * Costs of preparing data and metadata for publication should be included in project funding
- Apply a license not impeding reuse (e.g. CC-BY)

How to reap the benefits? /2

- The above when
  - publishing data/datasets (stand-alone)
  - publishing papers: to make available the data that underpins your research results (e.g. supplemental material)
  - publishing a "data paper"
- Demand proper citation by others who re-use your data/sets
- Promote/cite your data when appropriate
- Look for options to co-author papers with data re-users

Key takeaway points

- Researchers as open data publishers and consumers
  - Publish open data to reap benefits – individually and as research community
  - Recognise colleagues who share data, cite their datasets properly
- Research institutions
  - Reward researchers who publish data/sets
  - Change mind-sets by doing (not teethless mandates)
- Archives/repositories
  - Need sustained funding – importance of demonstrating usage/impact

References and additional material

- ADS – Archaeology Data Service, http://archaeologydataservice.ac.uk
- Charles Beagrie Ltd.: Keeping Research Data Safe (KRDS) Benefits Framework, http://beagrie.com/krds-i2s2.php
- Borgman, C.L: Research Data: Who will share what, with whom, when, and why? Fifth China – North America Library Conference 2010, Beijing, 8-12 September 2010, http://works.bepress.com/cgi/viewcontent.cgi?article=1237&context=borgman
- Data Seal of Approval, http://www.datasealofapproval.org
- DataCite, http://www.datacite.org
- DataCite Metadata Schema for the Publication and Citation of Research Data, V3.0, July 2013, http://schema.datacite.org/meta/kernel-3/index.html
- Digital Object Identifier (DOI), http://www.doi.info
- DRYAD, http://datadryad.org
- EC – European Commission: Online survey on scientific information in the digital age, Brussels, 2012, http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-scientific-information-digital-age_en.pdf
- EC Communication: Open data. An engine for innovation, growth and transparent governance (12.12.2011), http://ec.europa.eu/information_society/policy/psi/docs/pdfs/opendata2012/open_data_communication/en.pdf

References and additional material

- EC Communication: Towards better access to scientific information: Boosting the benefits of public investments in research (17.7.2012), http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf
- EDNA – e-depot Nederlandse archeologie, http://www.edna.nl
- European High-level Expert Group on Scientific Data (2010): Riding the wave. How Europe can gain from the rising tide of scientific data. A submission to the European Commission, October 2010, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf
- Heidorn, P.B: Shedding Light on the Dark Data in the Long Tail of Science. Library Trends 57(2), 2008, http://hdl.handle.net/2142/9127
- Internet Archaeology: Data Papers, http://intarch.ac.uk/authors/data-papers.html
- Journal of Open Archaeology Data, http://openarchaeologydata.metajnl.com
- MAPPA Open Data, http://mappaproject.arch.unipi.it/?lang=en
- OECD: Declaration on Access to Research Data from Public Funding (30.01.2004), http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157&Lang=en&Book=False
- OECD: Principles and Guidelines for Access to Research Data from Public Funding (2007), http://www.oecd.org/science/sci-tech/38500813.pdf

References and additional material

- Opportunities for Data Exchange (ODE) project / Kotarski R. et al. (2012). Report on best practices for citability of data and on evolving roles in scholarly communication, http://www.alliancepermanentaccess.org/index.php/community/current-projects/ode/outputs/
- PARSE.Insight: Insight into digital preservation of research output in Europe. Project deliverable D3.4: Survey Report, 9 December 2009, http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf
- Research Data Alliance, https://www.rd-alliance.org
- Science magazine: Science Staff introduction to the Special Issue "Dealing with Data", Science, Vol. 331 no. 6018, 11 February 2011, pp. 692-693, http://www.sciencemag.org/content/331/6018/692.short
- tDAR - The Digital Archaeological Record, http://www.tdar.org
- The Royal Society: Science as an Open Enterprise, June 2012, http://royalsociety.org/policy/projects/science-public-enterprise/report/
- Thessen, A.E & Patterson, D.J (2011) Data issues in the life sciences. In: ZooKeys 150: 15–51, http://www.pensoft.net/journals/zookeys/article/1766/data-issues-in-the-life-sciences
- zenodo, http://www.zenodo.org (CERN, related to OpenAIRE)

Disclaimer

ARIADNE is a project funded by the European Commission under the Community's Seventh Framework Programme, contract no. FP7-INFRASTRUCTURES-2012-1-313193.

The views and opinions expressed in this presentation are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.