ENHANCING ACCESS TO DATA FOR SCIENCE, TECHNOLOGY AND INNOVATION

GOING DIGITAL

Alan Paic
OECD Directorate for Science, Technology and Innovation

OVERVIEW

OECD instruments

- 1980/2013 Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data

- **2006 Recommendation concerning Access to Research Data from Public Funding**

- 2008 Recommendation for Enhanced Access and More Effective Use of Public Sector Information

- 2009 Recommendation of the Council on Human Biobanks and Genetic Research Databases

- 2014 Recommendation on Digital Government Strategies

- 2016 Recommendation of the Council on Health Data Governance

Other international initiatives

- European Commission's 2012 Recommendation on access to and preservation of scientific information

- FAIR principles (**F**indability, **A**ccessibility, **I**nteroperability, **R**euse)

- International Open Data Charter

- Science clouds in Europe, Australia, U.S., Africa

- Research Data Alliance recommendations

- TOP (http://cos.io/top/) policy framework

- Amsterdam call for Open Science

- Open Innovation, Open Science, Open to the World(European Commission)

- Association of American Universities (AAU) and Association of Public and Land-grant Universities (APLU) guidelines

- Overarching recommendation for open access to data, together with CDEP / PGC?

- Revision of 2006 Recommendation concerning Access to Research Data from Public Funding ?

- Policy report

TIMELINE

| Oct 2016 | 2017 | 2018 |

CSTP discussed and approved a proposed joint development of a possible new overarching recommendation on enhanced access to data, together with the Committee for Digital Economy Policy (CDEP) and the Public Governance Committee (PGC) [COM/DSTI/CDEP/STP/GOV/PGC(2016)1].

A survey was conducted to assess the current use of the Recommendation [DSTI/STP(2017)25],

13 March: Workshop: Towards New Principles For Enhanced Access To Public Data For Science, Technology And Innovation held with 28 panellists.

Next steps:
- Case studies
- Decision on Recommendation (October 2018)

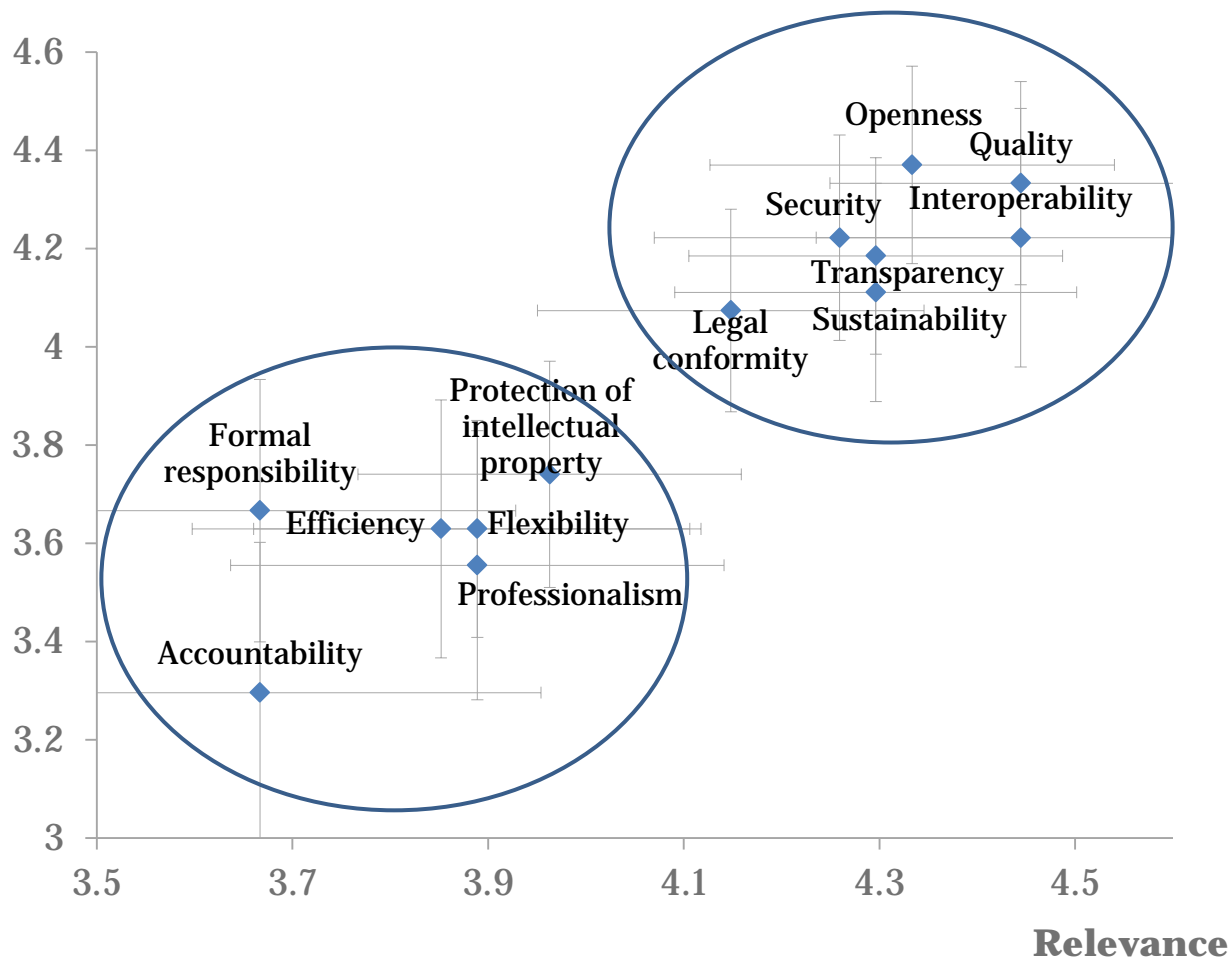# 2006 Recommendation concerning Access to Research Data from Public Funding

Principles

A. Openness - access on equal terms for the international research community at the lowest possible cost

B. Flexibility - adapt to changes in IT, diversity of research systems, legal systems and cultures

C. Transparency - information on research data to be available in a transparent way, ideally on Internet.

D. Legal conformity - respect legal rights & legitimate interests of all stakeholders in public research.

E. Protection of intellectual property - consider the applicability of copyright or of other IP laws.

F. Formal responsibility - rules and regulations regarding the responsibilities of the various parties.

G. Professionalism - establish and maintain codes of conduct to foster trust

H. Interoperability - member countries and research institutions should co-operate with international organisations charged with developing new standards., both semantic and technological

I. Quality - good practices in collection, dissemination and archiving to enable peer review, development of metadata, links to original datasets, data citation practice.

J. Security – protect against loss, destruction, modification and unauthorised access

K. Efficiency – data management practices to promote cost effectiveness; cost-benefit analysis to decide on maintaining datasets online; documentation of data to avoid duplication; incentives for researchers

L. Accountability - periodic evaluation by user groups, responsible institutions and funding agencies

M. Sustainability - administrative responsibility for the measures to guarantee permanent access to data that have been determined to require long-term retention.

Assessment of the principles concerning access to research data from public funding

Need to include (y-axis)

Relevance (x-axis)

Scatter plot points: Openness, Quality, Security, Interoperability, Transparency, Sustainability, Legal conformity, Formal responsibility, Protection of intellectual property, Efficiency, Flexibility, Professionalism, Accountability

Additional principles:

- Discoverability/findability
- Machine-readability
- Regulation of financial support of open access to data
- Definition of responsibility and ownership
- Recognition and reward system for data authors
- Possibility of licencing
- Definition of an embargo period

CSTP-GSF Workshop: Towards new principles for enhanced access to public data for STI

Main challenges discussed

- Data governance and trust: Balancing between the benefits of sharing against privacy and confidentiality

- Data standards, interoperability and re-use

- Definition of responsibility and ownership

- Recognition and reward system for data providers and stewards

- Business models for open data provision

- Building human capital and institutional capabilities at public agencies, to manage, create, curate and reuse data.

THE WORKSHOP GATHERED 28 SPEAKERS FROM ACADEMIA, GOVERNMENT, PRIVATE SECTOR AND NON-GOVERNMENTAL ORGANISATIONS

DATA GOVERNANCE AND TRUST: BALANCING BETWEEN THE BENEFITS OF SHARING AGAINST PRIVACY AND CONFIDENTIALITY

- **Issue**: Personal data provide research insights, but pose risks to individuals' **privacy**.

- Increased **scope** (including AI) and **scale** of data use (data Moore's law)

- **Trust** : rules and regulations (e.g. GDPR), consultations with stakeholders (Australia), quality control, certification, transparency, use of blockchain

- Consent cannot always be asked for. **Ethics review boards** to be guardians

- Degrees of opening: "as open as possible, as closed as necessary" / open by default (e.g. France)

- For public research data **reproducibility of scientific results** should be the driver.

- **Software** to be addressed by open policies at equal footing

- **Definitions**: distinction between legislation applicable to public sector information and public research data is not clear.

- Open data may actually aggravate the **digital divide** and social exclusion, as well as geographical disparity. Therefore socio-economic impact assessments should be considered.

Data standards, interoperability and re-use

- **F**indability: Data catalogues, or search engines to make data more findable. Science clouds being established.

- **A**ccessibility : As open as possible, as closed as necessary. Not only for academia, also for business

- **I**nteroperability : semantic (scientific vocabulary), legal (rights) and technical (machine readability).
  - Interoperability **across disciplines** is a specific challenge.

- **R**e-use : machine readable meta-data, for the long term.
  - The Open Archival Information System (OAIS) guidelines are a useful reference.

- **Pace**: Standard setting is slow & iterative process (vs. technology evolution)
  - Private sector may impose their own standards.
  - Certification alone cannot ensure cybersecurity.

- Use the **hourglass model** with a small number of key standards, and freedom to implement.

- Governments should commit to supporting **data infrastructure in the long term** for sustainability

- The Research Data Alliance (RDA) was initiated in 2013 with the goal of building the social and technical infrastructure to enable open sharing of data. Addresses interoperability of data, data citation, data catalogues, various standards and research data publishing.

DEFINITION OF RESPONSIBILITY AND OWNERSHIP

- Legislation should be **harmonised** across data custodians and funding agencies (U.S.)

- **Exceptions** from copyright law (e.g. for TDM) sometimes granted for non-commercial purposes – what about value creation?

- Private-law mechanisms, such as **licenses** and **agreements**, have superseded public-law instruments, such as copyright and other IP statutes.

- Good Research Data Management (RDM) principles are needed and researchers should set up coherent data management plans

- Issues of data ownership for data created through **public-private partnerships** (PPPs), hidden but increasing risk of 'privatisation' of data arising from publicly-funded research

- **Text and data mining** in PPP is not problematic if there is the right legal context; **very** problematic on EU level at the moment.

- **Licensing information** should be included as metadata.

- If publicly funded, research data have public good (economic), public domain (legal) and public interest (political or policy) characteristics.

RECOGNITION AND REWARD SYSTEM FOR DATA PROVIDERS AND STEWARDS

- Cultural change is a long process (researchers tend to be possessive about data)

- Perceived **barriers** and **risks** of open data to be mitigated by appropriate **acknowledgement** and **reward** systems.

- **Reproducibility**: make the data that support research results available to support independent verification.

- **Data citation** practices and protocols being developed but not yet widely implemented
  - missing: standard formats, citation metrics.

- Need to **reward** data release, include in **evaluation** and **recruitment** criteria

- **Incentives** to publish results as publications, but incentives to release data need to be developed:
  - allow embargo period for data sharing
  - require use of Persistent Identifiers (PIDs) to support data citation,
  - tracking data use for better metrics.
  - recognise data management skills and data products at a par with publications (NSF)

BUSINESS MODELS FOR OPEN DATA PROVISION

- **Costs** are **borne by the providers**, but the **benefits accrue mostly to users**.

- Public financing through **short term** project funding, responsibilities are **long term**.

- Access does not always have to be free, but in many cases should be free **at point of use.**

- Rapid growth in data production requires continuous **investment** in storage

- Centralisation allows to take advantage of **economies of scale**, but networked repositories offer better **ownership** and flexibility.

- **Cost/benefit analysis** is crucial, since some data may be too costly to curate and provide on open data platforms.

    - However, there is no standardised method of calculating the cost-benefit of curation of a dataset

- In particular, **benefits** are difficult to estimate – estimates provided by the Australian Productivity Commission show that estimates can vary up to a factor of 100!

- The automation of certain data curation functions needs to be evaluated and implemented where appropriate.

- **Researchers** lack data management planning and curation skills
- **Users** do not always have skills for correct interpretation and analysis.
- **Technical staff** in data repositories need training on data standards, and good policy and practice.
- **Lack of skills** breeds **lack of trust**
- More **higher education curricula** and **lifelong training** in data management are needed to maximize the value of preserved research data.
- **Data skills** are **not just for data stewards** – training needs to be self-evident in any curriculum. Such courses should be awarded ECTS credits.
- Need to care about **data scientists' career paths**:
  - research teams at NYU, UCB integrated them within the research teams with adequate career prospects and reward structure

Contact

Alan.Paic@oecd.org