Science as an Open Enterprise—looking at the wider context for Research Data Management as well as Open Science [or O-Research]

Prof Colin Wright
Advisor Cyberinfrastructure: DST
cjwright@csir.co.za
20151210

Agenda

Buzzword Compliance:

- ✓ Commons
- ✓ Cloud
- ✓ Open
- ✓ Data
- ✓ Science

Pick any two of these in the title ….

Open Science

- History of Open Science
- What is it?
- Do we need it? If so, why? Rationale?
- Features
- Are there inhibitors to Open Science?

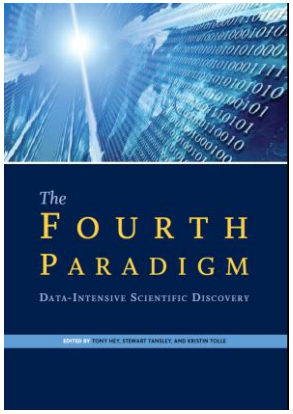
What Foundational Building Blocks have to be in place to promote Open Science? Or, what interventions promote Open Science?

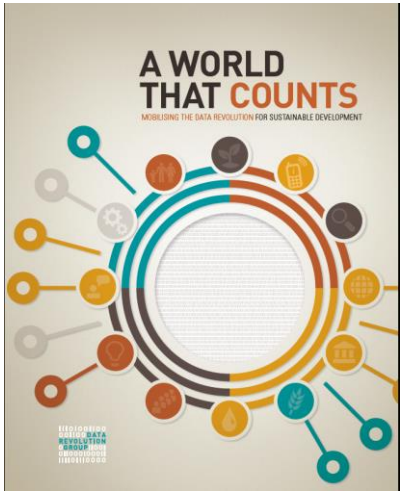
Open Data
Open Cyberinfrastructure / e-Infrastructure

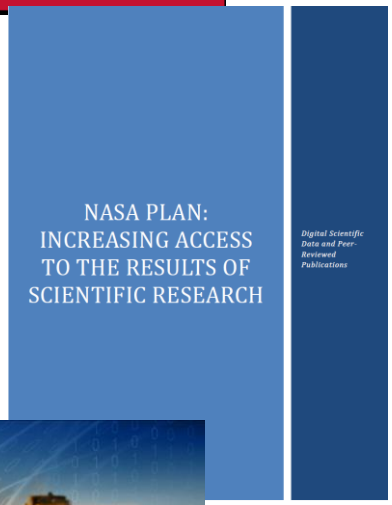
Exemplars of demonstrated consequences of practicing Open Science

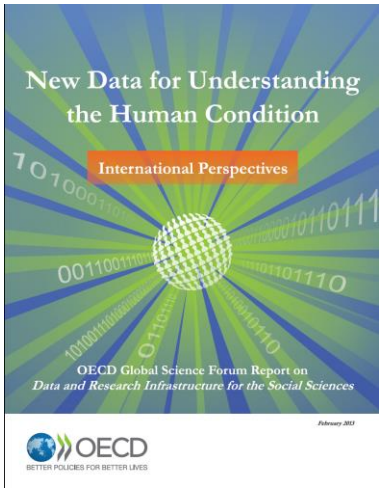
….. → Open Innovation

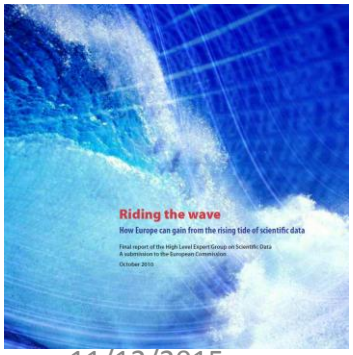
Some Useful Publications











History

Before journals (Wikipedia)

- *Before scientific journals, scientists had little to gain and much to lose by publicizing scientific discoveries.* Many, eg Galileo, Kepler, Isaac Newton, Christiaan Huygens, and Robert Hooke, claimed discoveries by describing them in papers coded in anagrams or cyphers and then distributing coded text.
- *Not publicizing discoveries quickly caused problems:* sometimes difficult to prove priority. Newton and Gottfried Leibniz both claimed priority in discovering calculus. Newton said that he wrote about calculus in the 1660s and 1670s, but did not publish until 1693. Leibniz published a treatise on calculus in 1684.
- *Aristocratic Patronage*: scientists received funding to develop either immediately useful things or to entertain, hence funding of science gave prestige to the patron → scientists under pressure to satisfy the desires of patrons, and discouraged from being open with research.
- Debates over priority inherent in systems where science not published openly—problematic for scientists who wanted to benefit from priority.
- *Philosophical Transactions of the Royal Society* (created 17th C). Publishing academic inquiry was controversial, and widely ridiculed.
- *The Royal Society was steadfast in its not yet popular belief that science could only move forward through a transparent and open exchange of ideas backed by experimental evidence.* Open Science?

Open Science…? What is it?

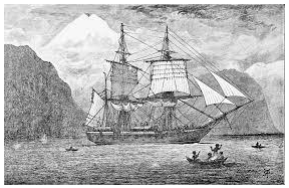
- Robert Merton (sociologist): 92% of cases of simultaneous discovery in the 17th C ended in dispute; number then dropped: 72% in 18th C, 59% by latter half of 19th C, and 33% by the first half of 20th C.
- *The decline in contested claims for priority in research discoveries can be credited to the increasing acceptance of the publication of papers in modern academic journals,* with estimates suggesting that around 50 million journal articles have been published.

MERTONIAN NORMS 1942 (Wikipedia)

CUDOS: acronym used to denote principles that should guide good scientific research. Modified definition below is most widely used.

- **Communalism:** equal access to scientific goods (IP); sense of common ownership to promote collective collaboration.
- **Universalism:** contribute to science regardless of race, nationality, culture, or gender.
- **Disinterestedness:** act for benefit of common scientific enterprise, rather than personal gain.
- **Organized Skepticism:** scientific claims exposed to critical scrutiny before being accepted.

Research become more data driven!



Data trickle

Earlier

Data Tsunami

Current

Characteristics of Modern Research

- Data
  - Heterogeneous, distributed, complex data sets
  - Flood of Social media data
  - Digitisation of historical papers and artefacts
  - Extreme data, Big data, long dark tail of data
- Social, legal Dynamics
  - Global collaborations
  - Citizen Science
  - Legal, cultural, language and proprietary rights hinder collaboration
  - Incentives to share…?
- Nature of Research & Infrastructures
  - Multi-scale; multidisciplinary
  - CERN, Large Hadron Collider—currently being upgraded
  - Distributed sensors
  - Research Infrastructures: national—SARIR and International
  - Uncoupled systems to highly coupled systems
  - Grand Challenges of 21st C transcend borders; science / research is becoming increasingly global.

Modern Research.
The scientific (research) discovery process is increasingly:

- Global
- Collaborative
- Multidisciplinary
- Integrative
- Complex
- Data requirements increasing
- And Data Driven
- … Hypothesis free Science

- But Data Skills in short supply!

Recent Press

The Economist | World politics | Business & finance | Economics | Science & technology | Culture

Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

Oct 19th 2013 | From the print edition

"I SEE a train wreck looming," warned Daniel Kahneman, an eminent psychologist, in an open letter last year. The premonition concerned research on a phenomenon known as

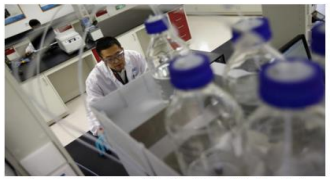
Newsweek

Scientific Case Studies

Read Customer Case Studies using QIAGEN's Bioinformatics Tools

TECH & SCIENCE

Science's Reproducibility Problem: 100 Psych Studies Were Tested and Only Half Held Up



The Economist | World politics | Business & finance | Economics | Science & technology | Cu

Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

Oct 19th 2013 | From the print edition

A SIMPLE idea underpins science: "trust, but verify". Results should always be subject to challenge from experiment. That simple but powerful idea has generated a vast body of knowledge. Since its birth in the 17th century, modern science has changed the world beyond recognition, and overwhelmingly for the better.

REPLICATING RESULTS

Why Biomedical Research Has A Reprodu Problem

February 26, 2014

SIDENOTES

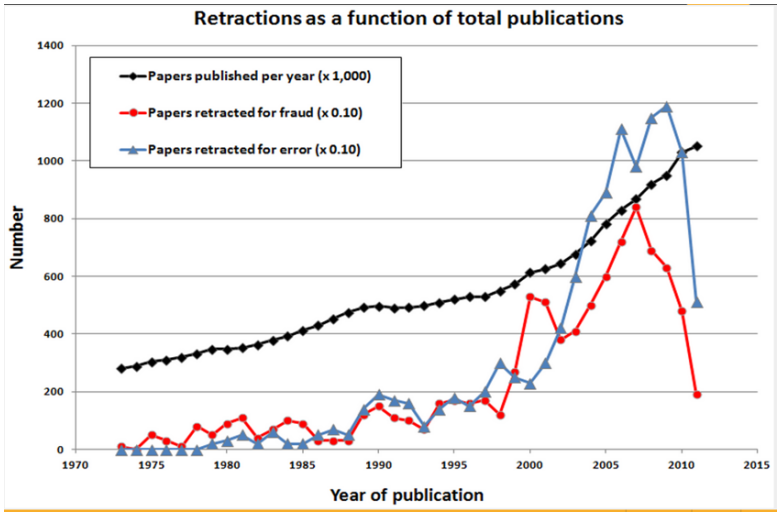
One of the fundamental principles of science is *reproducibility* – the idea that a discovery is valid only if any scientist in any lab can conduct the same experiment under the same conditions and obtain the same results. Without reproducibility, we could not distinguish scientific fact from error or chance, and scientific "laws" would vary from place to place and scientist to scientist.

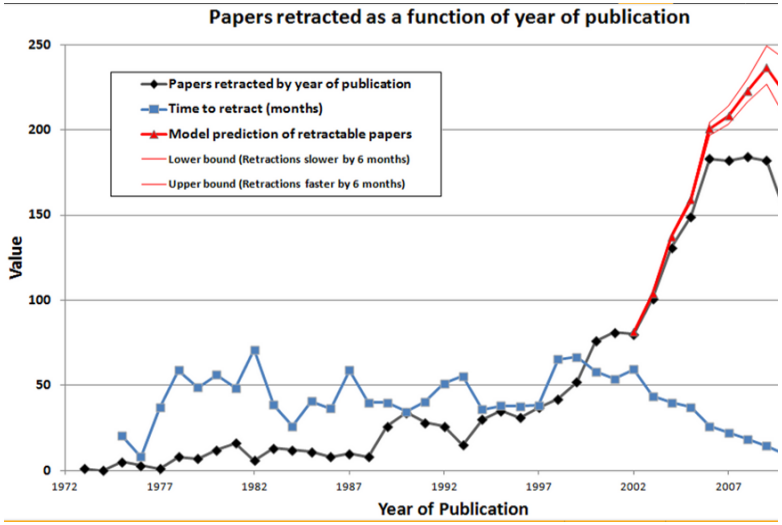
CONTRIBUTOR
Sarah Weil
Columbia University

While reproducibility is an essential principle of the scientific process, it isn't always easy to achieve. Recent[1] studies[2] in the field of biomedicine show that findings from an alarming percentage of scientific papers in even the top journals cannot be reliably reproduced by other researchers. Why does science fail to meet the basic standard of reproducibility? The current state of affairs results from a combination of the complex nature of modern scientific research, a lack of accountability for researchers, and the incentives created by a publish-or-perish culture in academia.

Reproducibility of published research?

Retractions





Reasons

- Fraudulent behaviour
- Invalid reasoning
- Absent or inadequate data or metadata

Causes

- Pressure to publish
- Pressure to make excessive claims
- Data hoarding
- Poor data science

Publication of all (not selected) data; algorithms; software…

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0068397

Open Science Project

What, exactly, is Open Science?
Posted on July 28, 2009 by Dan Gezelter
"I was recently asked to define what *Open Science* means. It would have been relatively easy to fall back on a litany of "Open Source, Open Data, Open Access, Open Notebook", but these are just shorthand for four fundamental goals":

- Transparency in experimental methodology, observation, and collection of data.
- Public availability and reusability of scientific data.
- Public accessibility and transparency of scientific communication.
- Using web-based tools to facilitate scientific collaboration

Michael Faraday's advice to junior colleague to: "Work. Finish. Publish." needs to be revised. It shouldn't be enough to publish a paper anymore. If we want open science to flourish, we should raise our expectations to: "Work. Finish. Publish. *Release.*"

NB "Data" here includes ….

ZA

Open Science (OS) essentially a set of related practices, e.g.:

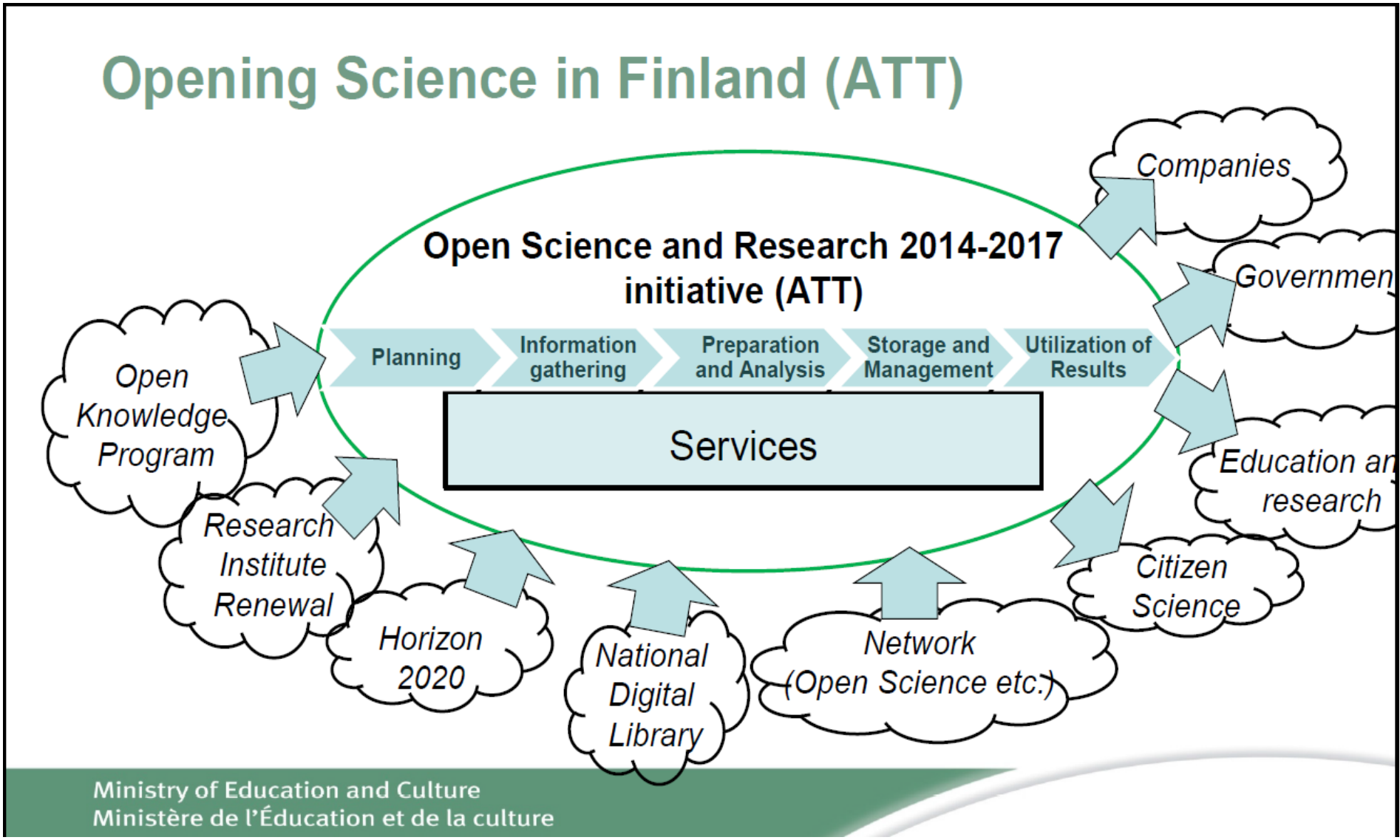
- Publications under Open Access rules (fully open, partially open);
- Community discussions and comments on preprints;
- Increased sharing of intermediary research results (open notebook research blogs), research data and software;
- Use of text and data mining techniques to study previous research results;
- Research oriented social networks;
- Virtual collaboration, virtual infrastructures;
- Seeing the public as a resource / partner, citizen science.

- OS practices are driven by scientists themselves: a re-appreciation of the role of data: Data is increasingly treated as the raw material of research, rather than purely as a means to confirm hypotheses.
- OECD: The value of data lies in their use. Full and open access to scientific data should be adopted as the international norm for the exchange of scientific data derived from publicly funded research.

Opening up the entire research process from agenda setting, data generation and data analysis, to dissemination and use with the aid of various emerging social platforms and tools.

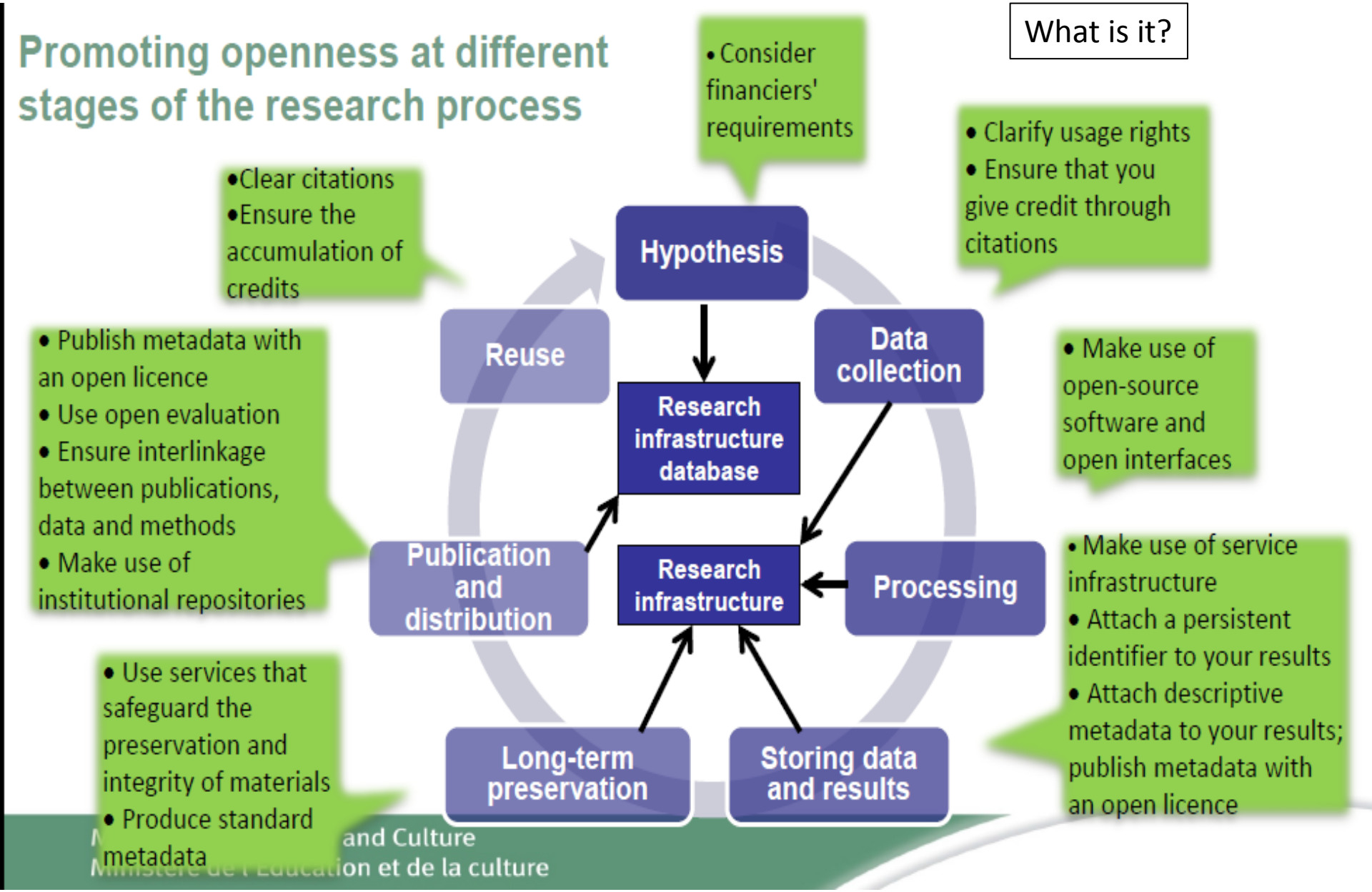
FINNISH OPEN SCIENCE AND RESEARCH INITIATIVE 2014–2017

Senior Advisor Sami Niinimäki



What is it?

Promoting openness at different stages of the research process



Consequences of openness in Science and Research:

- Improving efficiency in Research
  - reducing duplication and cost of creating, transferring and re-using data;
  - allowing more research from the same data;
  - multiplying opportunities for domestic and global participation in the research process;
  - speeding the transfer of knowledge;
  - Science more reliable, eg re-use of data
- Research validation
  - Increasing transparency and quality of the research validation process by allowing a greater degree of replication and validation of scientific results, more integrity
- Economy
  - Increasing knowledge spill overs to the economy
  - Data-intensive science a key economic driver
  - Growing evidence that Open Science has an impact on the research enterprise, business and innovation, and society more generally.
- Collaboration
  - Addressing global challenges more effectively—global challenges require co-ordinated international actions.
- Societal engagement
  - Recent analysis reveals that enhanced public access to scientific publications and research data increases the visibility of, and spill overs arising from, scientific research
  - Promoting Citizen's engagement in Science and Research.
  - Reconnect science and society
  - Science more responsive to societal challenges, also through Crowd Funding
- Sustainability: Crowd-funding an important research funding source.

Rationale

Arguments against Open Science:

- **Too much unsorted information overwhelms scientists.**
- **Science will be used for bad things.**
- **The public will misunderstand science data.**
- **Increasing the scale of science will make verification of any discovery more difficult.**

Arguments for Open Science:

- **Open access publication of research reports and data allows for rigorous peer-review**
- **Science is publically funded so all results of the research should be publically available**
- **Open Science will make science more reproducible and transparent**
- **Open Science has more impact**

http://digitheadslabnotebook.blogspot.co.uk/2014/01/guide-to-open-science.html

EC Assessment of Science 2.0 (Open Science)

Key Drivers of Open Science

1. Availability of digital technologies and their increased capacities
2. Researchers looking for new ways of disseminating their output
3. Researchers looking for new ways of collaborating
4. Increase of the global scientific population
5. Growing criticism of the current peer review system
6. Public demand for better and more effective science
7. Public funding supporting Science 2.0
8. Growing public scrutiny of research
9. Public demands for faster solutions to societal challenges
10. Scientific publishers engaging in Science 2.0
11. Citizens acting as scientists

Barriers for Science 2.0

1. Concerns about quality assurance
2. Lack of credit-giving to Science 2.0
3. Lack of integration in the existing infrastructures
4. Limited awareness of benefits of Science 2.0 for researchers
5. Lack of financial support
6. Uncertain benefits for researchers
7. Legal constraints, eg copyright law
8. Lack of research skills fit for Science 2.0
9. Lack of incentives for junior scientists to engage with Science 2.0
10. Concerns about ethical and privacy issues

EU Assessment of Science 2.0 (Open Science)

Successful exploitation of Open Research will come from six changes:

1. Shift away from a research culture where **data is viewed as a private preserve**;
2. Expanding criteria used to evaluate research to **give credit** for useful data communication and novel ways of collaborating;
3. Development of **common standards** for communicating data;
4. **Mandating intelligent openness for data** relevant to published scientific papers;
5. **Strengthening the cohort of data scientists** needed to manage and support the use of digital data;
6. Development and **use of new software tools to automate and simplify the creation and exploitation** of datasets, ie good Data Management and exploitation.

What infrastructure is required by Open Science to be viable

A. Policies: National, Funding Agencies; Institutions
B. E-Research skilled workforce
C. Open, sustainable e-Infrastructures
D. Collaborations: national and international
E. Skilled People

Data Principles

GSO Research Data Working Group developed a White Paper on Data with 5 Principles for an Open Data Infrastructure and effective management. At a meeting on 12 June 2013 (London, UK), the G8 Science Ministers endorsed the aforementioned actions and adopted the Framework and Data White Paper

Five Principles for an Open Data Infrastructure and effective management .

- **Discoverable** – using conventional search methods to easily find data;
- **Accessible** – openly available with as few restrictions as possible;
- **Understandable** – data plus additional or supporting information and documentation needed for understandability and effective use;
- **Manageable** – data management policies and plans at project and institutional level, maintenance of available data, coordination of technologies and services; and
- **People** – high-skilled and adaptable workforce and culture to capture data and make it available
- \+ Metrics

Royal Society

- Accessible
- Intelligible
- Assessable
- Usable

OECD principles

- Openness
- Flexibility
- Transparency
- Legal conformity
- Protection of intellectual property
- Formal responsibility
- Professionalism
- Interoperability
- Quality
- Security
- Efficiency
- Accountability
- Sustainability

Impact

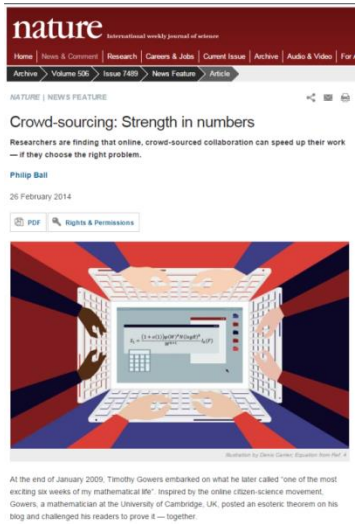
Examples: Open Science, Open Data and Open Innovation:

- Genome Project
- Australia (Houghton and Sheehan) 2014 Conservatively, we estimate that the value of data in Australia's public research to be at least $1.9 billion and possibly up to $6 billion a year at current levels of expenditure and activity. Research data curation and sharing might be worth at least $1.8 billion and possibly up to $5.5 billion a year, of which perhaps $1.4 billion to $4.9 billion annually is yet to be realized. Hence, any policy around publicly-funded research data should aim to realise as much of this unrealised value as practicable
- UK
- US NASA Satellite imagery, until 2008 sold at $600 per scene when it became free. Usage jumped from 19 000 scenes per year to 2 100 000 p.a. Much scientific benefit. Created $935m pa Environmental Management industry, direct benefit >$100mpa to US economy
- Crowdsourcing Science
  - A classic example of this approach is Tim Gowers, who **posted in his blog a mathematical question** and in a matter of days the commenters had solved it. Published proof under pseudonym DHJ Polymath
  - Distributed computing used to find Marsenne Prime Numbers (**GIMPS**)

Open Innovation
- IBM desktop computer
- Google and Android
- Elon Musk

In June 2014, Elon Musk opened up all Tesla patents. In a blog post announcing this, he wrote that patents "serve merely to stifle progress, entrench the positions of giant corporations and enrich those in the legal profession, rather than the actual inventors."

in science. For example, in 2009 the Fields medallist mathematician Tim Gowers posted an unsolved mathematical problem on his blog with an invitation to others to contribute to its solution. In just over a month and after 27 people had made more than 800 comments, the problem was solved. At the last count, ten similar projects are under way to solve other mathematical problems in the same way.


nature

Crowd-sourcing: Strength in numbers

Researchers are finding that online, crowd-sourced collaboration can speed up their work — if they choose the right problem.

Philip Ball

26 February 2014

Some SA actions:

ZA

- Berlin declaration (>=11 Univs, ASSAf, LIASA)
- CODATA (NRF)
- **NRF: Statement on Open Access to Research Publications from the National Research Foundation (NRF)-Funded Research**
- **Data Plans: various Universities**
- **ASSAf:** OPEN ACCESS: SciELO SA is an open-access (free to access and free to publish), full-text searchable database of selected, high-quality SA scholarly journals. At September 2014 there are 47 journals in the collection, with one title being added per month, as well as 585+ issues and 9 500+ articles. An average of 3 050 visits per day is currently being recorded. Foreseen that eventually, after the ASSAf Peer Review Panels evaluations are complete, more than 180 SA Scholarly journals may be published on the collection. SciELO: Scientific Electronic Library Online.
- e-Research initiatives
- Earth Observation Statement finalised Nov 2015 at Ministers' summit
- DIRISA + SAEON

The Berlin Declaration urges its signatories to:

1. Increase the support for and interoperability of OA repositories; reduce and where possible eliminating embargoes; improve ability to re-use works;
2. Support new and innovative OA publishing models that meet the highest possible scholarly standards; invest into a publication infrastructure optimised for the needs of research and scholarship; and
3. Cooperate with one another to ensure a smooth transition to a stable and functioning, truly open scholarly publishing system, including access to scholarly source and cultural heritage data, where the full text of every research work is open immediately upon publication.

SA Legal Environment

- Intellectual Property Rights from Publicly Financed Research and Development Act (Act No 51 of 2008): potentially restricts the access to research data and outputs which would conflict with the principles of data democracy.
  - NIPMO established
- Spatial Data Infrastructure Act (Act No 54 of 2003): Improves discoverability but does not guarantee access. Could limit data availability because of the legal obligations of 'custodianship', which may make institutions unwilling to publish all available data sets.
- South African Weather Service Amendment Bill: An example of a step in the wrong direction, it impedes climate change research and discriminates against local enterprise while foreign entities are free to continue with services and data in competition with the SA Weather Service.
- Promotion of Access to Information Act, 2000 (Act No. 2 of 2000):
- Protection of Information Bill: policy or regulation required to declassify and exempt the data.
- National Archives and Records Service of South Africa Act, 1996 (Act No. 43 of 1996)
- Protection of Personal Information Act (Act 912, 2013)

Main findings and policy messages

- Open science is a means and not an end
- Open science is more than open access to publications or data: it includes many aspects and stages of research processes
- Policies to promote open data are less mature than those to promote open access to scientific publications
  - Better incentive mechanisms to promote data sharing practices among researchers are needed
  - Data related skill development is essential
  - Training of and awareness-raising among researchers is important for the development of an open science culture
- ❖ Repositories and online platforms will not have impact if the information they contain is not of good quality
- ❖ The long term preservation costs of openly available research output need to be considered

Main findings … 2

- Open science policies should be principle based but adapted to local realities
- Clear legal frameworks for the sharing of publications and re-use of data sets are needed at the national and international levels
- Policy makers need to promote openness in science while at the same time preserving competition
- Consultative approaches that involve all relevant actors for open science are a key component of successful open science strategies
- International collaboration in the area of open science is necessary to address global challenges

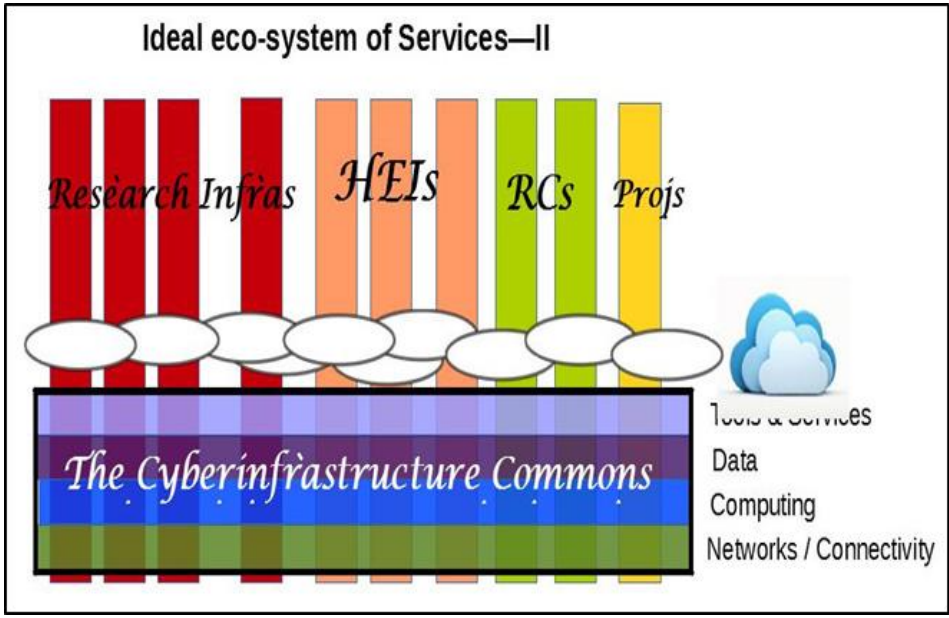
EAG

H2020 RI and e-I EAG recommendation to EC:

1. A key requirement for the future development of Open Science is the provision of sustainable and accessible research infrastructures where the science community can have their research data stored, enriched, curated and made available for re-use.
2. Open Science infrastructures could offer [European] researchers an environment with open and seamless services for data storage, management and analysis, notably by federating existing and newly developed e-infrastructures under appropriate governance schemes and by supporting the development of services for data analysis and exploitation.
3. The proposed European Open Science Cloud must be built upon existing and emerging Research Infrastructures that demonstrate excellence, are well-rooted in the research community and the research organizations, and that are openly accessible and have a data management framework that enables open access and data exploitation.
4. It is essential that the governance model for the proposed European Open Science Cloud be structured so as to achieve this. It must be designed in consultation with key scientific communities, and requires the continued involvement of stakeholders for its effective operation and sustainability.

Pillars supporting Open Science / Research

- Good, consistent policies and legal framework
- Sustainability
- Coherent Cyberinfrastructure: International, National & Local
- Open Cyberinfrastructure Commons and Advanced Services
- Data Standards which cohere with international Standards
- Robust Data Stewardship Practices
- Skilled People



1. Making Open Science a reality, OECD, 2015
2. Rome Declaration: Open Science for the 21st century, 11-12 April 2012
3. Science as an open enterprise, The Royal Society, June 2012
4. http://digitheadslabnotebook.blogspot.co.uk/2014/01/guide-to-open-science.html
5. Data-Driven Innovation: Big Data for growth and well-being, OECD, 2015
6. The Fourth Paradigm: Data Intensive Scientific Discovery, Editors Tony Hey, Stewart Tansley and Kristin Tolle, Microsoft Research, 2009
7. A world that Counts, Report by Independent Expert Advisory Group to UN Secretary General, 2014
8. Data Driven Innovation: Big Data for Growth and Well-Being, OECD Report, 2015
9. Data Intensive e-Science Frontier
10. Making Open Science a Reality, OECD Report, 2015
11. NASA Plan: Increasing Access to the results of Scientific Research, 2014
12. New Data for Understanding the Human Condition: Data and Research Infrastructure for the Social Sciences, OECD Report, 2013
13. Riding the Wave, How Europe can gain from the rising tide of scientific data, Report HLEG, 2010
14. The Data Harvest, RDA 2014