Global Partnerships in Digital Scholarship

Geoffrey Boulton
The Royal Society & Edinburgh University

ARL Forum
October 2015
Washington DC
1. The Digital Revolution: challenges and opportunities
2. Responding; nationally and internationally
3. Open knowledge for the common good?
A common pursuit
A systematic organisation of knowledge that can be rationally explained and reliably applied.

Cognitive strategies:

• **Empirical claims** – how the world is (aimed at causal explanation)
• **Normative claims** – the world shaped to fit the norm (David Hume – we cannot derive “ought” from “is” - justice, freedom)
• **Formal analysis** – mathematical and analogous forms of formal logic
• **Interpretation** – e.g. of the significance to humans of objects, rituals etc – making truth-claims about objects. (aimed at understanding, meaning and significance)

Conclusion:
• All disciplines use all these approaches, though in different proportions.
• Empirical claims are based on “data”
• Data issues are the same for all of us in essence, but vary in detail
edifices, even though in science at least, electronic access anywhere, any time, is the norm, dispersed support from appropriately trained e-librarians is the need, and few scientists now darken the door of a conventional library.

The data explosion and our capacity to combine, integrate and analyse data offer powerful new ways of unravelling complexity, improving forecasts of system behaviour and detecting patterns in phenomena that have hitherto been beyond our capacity to resolve. They offer the opportunity to reuse, to combine and to recombine data in ways that deepen these capacities. Exploiting these opportunities will depend upon access to and linking between many data sets, requiring that research data should be made routinely open and readily accessible. It will depend upon developing an ethos of data sharing and facilitating new modes of collaboration that increase the creativity of the scientific enterprise through interaction of many brains and many communities unbounded by institutional walls. These changes would also enable scientific concepts and the evidence that underlies them to be more effectively disseminated through society and in education, in ways that could change the social dynamics of science, contributing towards the evolution of science as a public enterprise rather than one conducted behind closed laboratory doors.

There is, however, a downside to the 'data explosion', of which we have only recently become aware. Such are the magnitudes of much of the data that provide the evidence for scientific concepts, that traditional habits of rigorous inclusion of data, and the metadata that describes their genesis, in conventionally published work have fallen away in recent decades. As a consequence, science may have been sleepwalking into a crisis of credibility. This was exemplified two years ago by a paper in which the authors reported attempts to replicate the results of 50 benchmark papers in pre-clinical oncology. They succeeded in doing so in only 11% of cases. The failure in 89% of cases reflected in part failures of scientific logic, but in many it reflected the failure to include adequate data or metadata, such that even if the conclusions had been logically...

Figure 1. Henry Oldenburg, first secretary of the Royal Society, who launched the first and most enduring scientific journal, the Philosophical Transactions of the Royal Society, and the title page of its first volume. Oldenburg also invented 'peer review' by asking two Fellows of the Society to review submitted work and give him advice on whether it should be published.

Open communication of data: the source of a scientific revolution and the basis of scientific progress

Henry Oldenburg
“False facts are highly injurious to the progress of science, for they often long endure; but false views ... do little harm, as everyone takes a salutary pleasure in proving their Falseness.”

“No amount of experimentation can prove me right; a single experiment can prove me wrong.”

“The progress of science is strewn, like an ancient desert trail, with the bleached skeletons of discarded theories that once seemed to possess eternal life.”
The Data Deluge
problems & opportunities

Available storage
IT BUDGETS (INCREASE)
COST OF STORAGE/GB (DECREASE)

2011 2012 2013 2014 2015
The Challenge: the “Data Deluge” is undermining “self correction”

THEN AND NOW
A crisis of credibility?
100 studies of social psychology in top ranking journals in 2008

75% not reproducible
A crisis of replicability and credibility?

Pre-clinical oncology – 89% not reproducible

A fundamental principle: the data providing the evidence for a published concept MUST be concurrently published, together with the metadata.

To do otherwise should come to be regarded by all, including journals, as scientific MALPRACTICE.
“Scientists like to think of science as self-correcting. To an alarming degree, it is not.”
The issue

The possibility/probability that most research conclusions are wrong

Why?

• Absent or inadequate data and/or metadata
• Invalid reasoning
• Fraudulent behaviour

If we correct these matters, will that solve the problem?
• Not entirely - we don’t want to inhibit boldness and imagination
• Science has proven to be the best way of gaining reliable knowledge - but it will remain uncertain & provisional
The opportunity

Open knowledge

“If you have an apple and I have an apple and we exchange these apples, then you and I will still each have one apple. But if you have an idea and I have an idea and we exchange these ideas, then each of us will have two ideas.”

With the added ingredient of digital technology, Shaw’s vision for doubling, trebling and quadrupling ideas is coming to pass. ‘Open knowledge’ is the power to promote fast, creative innovation by allowing an idea to leave your hands. Essentially, something is ‘open’ if anyone is free to use, reuse and redistribute it — subject only, at most, to the requirement to attribute the author and/or a share-alike license.
New modes of technology-enabled creativity: 

e.g. Crowd-sourcing

An unsolved problem posed on his blog.

32 days - 27 people - 800 substantive contributions

Emerging contributions rapidly developed or discarded

Problem solved!

“Its like driving a car whilst normal research is like pushing it”

What inhibits such processes?
- The criteria for credit and promotion

- ALTMETRICS THE ANSWER?
... seizing the data opportunities depends on an ethos of data-sharing e.g. a growing number of disciplinary communities

EMBL-EBI services

...provide tools to help researchers use it

Classify it

Labs around the world send us their data and we...

Analyse, add value and integrate it

Share it with other data providers

A collaborative enterprise
The opportunity: from “simple” science to complexity, from uncoupled to highly coupled systems
Complex systems

Simulating a complex system

Emergent behaviour of a specific 6-component coupled system

Characterising a complex system

Image of brain cells in a rat
The opportunity: data-modelling: iterative integration

Satellite observation

Initial conditions

Model forecast

Surface monitoring

Model-data iteration - forecast correction
System descriptions: from simple to complex

Simple relationships
Classical statistics
- Linear regression
- Cluster analysis

Complex systems
No mathematical pipeline
Dynamic/complex behaviour
- Topological Analysis
A barrier to openness? - Analytic overload. E.g. - Global Earth Observation System of Systems

A disconnect between machine analysis & human cognition?

- What is the human role?
- Can we analyse & scrutinise what is in the black box? - & who owns the box?
- What does it mean to be a researcher in a data intensive age?
Responsibilities & actions

• Researchers:  
  - changing the mindset
  - developing the skills

• Learned Societies:  
  - influencing their communities

• Universities/Insts:  
  - support processes
  - incentives & promotion criteria
  - proactive, not just compliant
  - strategies (e.g. the library)
  - management processes

• Funders of research:  
  - mandate intelligent openness
  - accept diverse outputs
  - cost of open data is a cost of science
  - strategic funding for technical solutions
    (a priority for international collaboration)

• Publishers & editors:  
  - mandate concurrent open deposition
A national data-intensive system
UK response to the challenge: Research Data Forum
Universities/Institutes; Funders; Publishers; Learned Societies; Technical Bodies
(UUK, Russell Group, RCUK, HEFCE, British Library, JISC, RIN, RSC, W3C, PLOS, Nature, Wellcome Trust, Dryad, CODATA, W3C etc)

Purpose
• Articulate rationale, principles, processes and priorities
• coherent approach across the research process – all the disciplines
• consistent with and influencing international developments (CODATA/RDA)
• practical steps to implement an open data regime & remove barriers
• advise Govt on its proper role (thro’ RSTB)

First targets
• RC/FC/Univs/Insts concordat (similar to that on research integrity)
• Data citation using Datacite
• Adoption of “intelligent openness” criteria by RCs
• Database registers
• International conversations

Dangers on the flank
• Publishers inhibition of text and data mining
• EU confidentiality regulation
An international voice for science

The consortium

- International Council for Science (ICSU)
- International Social Science Council
- The World Academy of Science
- Inter-Academy Panel
- Inter-Academy Medical Panel
- Inter-Academy Partnership

First meeting of Science International
South Africa – 7-9 December 2015
Agenda - Big Data/Open Data
- Principles for Open Data
- Capacity Mobilisation in the Global South
i. Responsibility of scientists and their institutions
Publicly funded scientists and scientific institutions have a responsibility to contribute to the public good through the creation and communication of new knowledge, of which associated data are intrinsic parts. They have a responsibility to make such data openly available to others in ways that permit them to be re-used and re-purposed.

iii. Responsibility of publishers of scientific results
Publishers of research papers that present scientific concepts should require the evidential data to be concurrently made intelligently open, if possible in a reliable data repository.

iv. Boundaries
Open data should be the default position for publicly funded research, although there should be proportional exceptions for cases of legitimate commercial exploitation, privacy, confidentiality, safety and security.

viii. Text and data mining
The historical record of scientific discovery and analysis published in scientific journals should be accessible to text and data mining (TDM) at no additional cost by scientists from journals to which their institution already subscribes. (ENGLISH!)
International Research Data Collaboration

CODATA
- Policies & practice
- Frontiers of data science
- Capacity Building

WDS
- Data stewardship
- Data standards

RDA
- Interoperability
Regional Platforms for Open Science

Shared investment in infrastructure; harvesting and circulating good ideas; spreading and supporting good practice; capacity building; promoting applications; linking to international programmes and standards.
Open science

Inputs

- Collecting the data
- Doing research

Doing science openly

Open data

- Administrative data (held by public authorities e.g. prescription data)
- Public Sector Research data (e.g. Met Office weather data)
- Research Data (e.g. CERN, generated in universities)
- Research publications (i.e. papers in journals)

Outputs

Researchers - Govt & Public sector - Businesses - Citizens - Citizen scientists

Open access

Science as a public enterprise & the future of the open society
Open Knowledge

Open Science

Data / Publications

Monomodal/Multimodal
Interdisciplinary

Stakeholders

Rigour

Innovation

Policy

Solutions

Researchers
DIGITAL REVOLUTION
Committee on Data for Science and Technology

**Strategic objectives**
- Frontiers of Data Science (ie. Valid reasoning)
- Practice and Policies of data system management
- Capacity Mobilisation in the Global South

**Examples of CODATA Task Groups**
- Non-Replicability
- Data at Risk
- Management of Physical Objects in the Digital Era
- Linked Open Data for Global Disaster Risk Reduction
- Preservation and Access to Data for Developing Countries
- Policy, data supply, information creation, citizen science, action
- Interoperable Data Publications
- Data Citation Standards and Practices
- Legal interoperability
Changing technology & the historic role of the library
to collect, to organize, to preserve knowledge, and to make it accessible

What does this mean in a post-Gutenberg world?
• vast data volumes
• vast computational capacity
• instantaneous communication
• access anywhere, anytime
A realiseable aspiration: all scientific literature open & online, all data open & online, and for them to interoperate

... but, this is a process, not an event!