Science as an Open Enterprise

Geoffrey Boulton
(Royal Society, University of Edinburgh)

Open Aire – Feb 2013

Report: www.royalsociety.org
Open communication of data: the source of a scientific revolution and of scientific progress

Henry Oldenburg

THE ROYAL SOCIETY
"It is therefore thought fit to employ the [printing] press, as the most proper way to gratify those [who] . . . delight in the advancement of Learning and profitable Discoveries [and who are] invited and encouraged to search, try, and find out new things, impart their knowledge to one another, and contribute what they can to the Grand Design of improving Natural Knowledge . . . for the Glory of God . . . and the Universal Good of Mankind."

.... how do we achieve these ends in the post-Gutenberg era, when massive digital acquisition and cyber space have replaced the printing press?
Problems & opportunities in the data deluge

Available storage

IT BUDGETS (INCREASE)

COST OF STORAGE/GB (DECREASE)

2011 2012 2013 2014 2015

10^{20} bytes
The challenges & opportunities?

- Closing the concept-data gap – maintaining scientific self-correction & credibility
- Exploiting the data deluge & computational potential
- Combating fraud
- Addressing planetary challenges
- Supporting citizen science
- Responding to citizens’ demands for evidence
- Restraining the “Database State”
A crisis of replicability …… and of the credibility of science?

![Table showing reproducibility of research findings]

The data providing the evidence for a published concept MUST be concurrently published, together with the metadata.
Challenges & opportunities?

- Closing the concept-data gap – maintaining scientific self-correction & credibility
- **Exploiting the data deluge & computational potential – data sharing**
- Combating fraud
- Addressing planetary challenges
- Supporting citizen science
- Responding to citizens’ demands for evidence
- Restraining the “Database State”
Proven benefit so that data sharing becomes embedded in ethos & practice – bio-informatics

ELIXIR Hub (European Bioinformatic Institute) and ELIXIR Nodes provide infrastructure for data, computing, tools, standards and training.
New scientific knowledge from data

E.g. the potential of linked data

• data integration
• dynamic data
and the economic implications

**Big data—a growing torrent**

- $600,000 to buy a disk drive that can store all of the world's music
- **5 billion** mobile phones in use in 2010
- **30 billion** pieces of content shared on Facebook every month
- **40%** projected growth in global data generated per year vs. **5%** growth in global IT spending
- **235** terabytes data collected by the US Library of Congress by April 2011
- **15 out of 17** sectors in the United States have more data stored per company than the US Library of Congress

**Big data—capturing its value**

- **$300 billion** potential annual value to US health care—more than double the total annual health care spending in Spain
- **€250 billion** potential annual value to Europe's public sector administration—more than GDP of Greece
- **$600 billion** potential annual consumer surplus from using personal location data globally
- **60%** potential increase in retailers' operating margins possible with big data
- **140,000–190,000** more deep analytical talent positions, and
- **1.5 million** more data-savvy managers needed to take full advantage of big data in the United States
Its not just curation, retrieving and integrating data – it's also what we do with it!

Jim Gray - “When you go and look at what scientists are doing, day in and day out, in terms of data analysis, it is truly dreadful. We are embarrassed by our data!”

• Looking for inherent patterns – not just the expected/hoped for
• Partial reporting of data (cherry-picking) is scientific malpractice
• The role of Bayesian logic
Challenges & opportunities?

• Closing the concept-data gap – maintaining scientific self-correction & credibility

• Exploiting the data deluge & computational potential

• **Combating fraud**

• Addressing planetary challenges

• Supporting citizen science

• Responding to citizens’ demands for evidence

• Restraining the “Database State”
“Scientific fraud is rife: it's time to stand up for good science”

“Science is broken”

**Examples:**
- psychology [academics making up data](#),
- anaesthesiologist Yoshitaka Fujii with 172 faked articles
- *Nature* - rise in biomedical retraction rates overtakes rise in published papers

**Cause:**
Rewards and pressures promote extreme behaviours, and normalise malpractice (e.g. selective publication of positive novel findings)

**Cures:**
Open data for replication
Transparent peer review
Not just personal integrity – but system integrity
Challenges & opportunities?

- Closing the concept-data gap – maintaining scientific self-correction & credibility
- Maintaining the credibility of science
- Exploiting the data deluge & computational potential
- Combating fraud
- **Addressing planetary challenges**
- Responding to citizens’ demands for evidence
- Supporting citizen science
- Restraining the “Database State”
Why is open data an urgent issue?

• Closing the concept-data gap

• Maintaining the credibility of science

• Exploiting the data deluge & computational potential

• Combating fraud

• Addressing planetary challenges

• **Responding to citizens’ demands for evidence**

• Supporting citizen science

• Restraining the “Database State”
Why is open data an urgent issue?

- Closing the concept-data gap
- Maintaining the credibility of science
- Exploiting the data deluge & computational potential
- Combating fraud
- Addressing planetary challenges
- Responding to citizens’ demands for evidence
- **Supporting citizen science – the 2030 question**
- Restraining the “Database State”
Opening-up science: 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!

“It’s like driving a car whilst normal research is like pushing it”

What inhibits such processes?
- The criteria for credit and promotion.
Why is open data an urgent issue?

• Closing the concept-data gap
• Maintaining the credibility of science
• Exploiting the data deluge & computational potential
• Combating fraud
• Addressing planetary challenges
• Supporting citizen science
• Responding to citizens’ demands for evidence
• Restraining the “Database State”
Openness of data \textit{per se} has no value. Open science is more than disclosure.

For effective communication, replication and re-purposing we need \textit{intelligent openness}. Data and meta-data must be:

- Accessible
- Intelligible
- Assessable
- Re-usable

Only when these four criteria are fulfilled are data properly open.

Metadata must be audience-sensitive.

Scientific data rarely fits neatly into an EXCEL spreadsheet!
Which publicly funded data for what purpose?

Data supporting the argument of a published paper?
- simultaneous deposition of citable data

Why should other data be open?
- greater benefit to science
- it’s not “our” data

Who should it be intelligently open to?
- other scientists
- citizen scientists
- the wider public

The dilemma of choice

Contradictory injunctions
Pressure to:
- commercialise, or
- share, collaborate, disseminate
Boundaries of openness?

Openness should be the default position, with proportional exceptions for:

• **Legitimate commercial interests** *(sectoral variation)*

• **Privacy** *(completely anonymised data is impossible)*

• **Safety & security** *(impacts contentious)*

All these boundaries are fuzzy
Commercial interests: potential by sector

A heat map shows the relative ease of capturing the value potential across sectors.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Sectors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goods</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Manufacturing</td>
</tr>
<tr>
<td></td>
<td>Construction</td>
</tr>
<tr>
<td></td>
<td>Natural resources</td>
</tr>
<tr>
<td></td>
<td>Computer and electronic products</td>
</tr>
<tr>
<td></td>
<td>Real estate, rental, and leasing</td>
</tr>
<tr>
<td></td>
<td>Wholesale trade</td>
</tr>
<tr>
<td></td>
<td>Information</td>
</tr>
<tr>
<td><strong>Services</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Transportation and warehousing</td>
</tr>
<tr>
<td></td>
<td>Retail trade</td>
</tr>
<tr>
<td></td>
<td>Administrative, support, waste management, and remediation services</td>
</tr>
<tr>
<td></td>
<td>Accommodation and food services</td>
</tr>
<tr>
<td></td>
<td>Other services (except public administration)</td>
</tr>
<tr>
<td></td>
<td>Arts, entertainment, and recreation</td>
</tr>
<tr>
<td></td>
<td>Finance and Insurance</td>
</tr>
<tr>
<td></td>
<td>Professional, scientific, and technical services</td>
</tr>
<tr>
<td></td>
<td>Management of companies and enterprises</td>
</tr>
<tr>
<td><strong>Regulated and public</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Government</td>
</tr>
<tr>
<td></td>
<td>Educational services</td>
</tr>
<tr>
<td></td>
<td>Health care and social assistance</td>
</tr>
<tr>
<td></td>
<td>Utilities</td>
</tr>
</tbody>
</table>

Overall ease of capture index:
- **Dark blue**: Top quintile (easiest to capture)
- **Medium blue**: 2nd quintile
- **Light blue**: 3rd quintile
- **Gray**: Bottom quintile (most difficult to capture)
- **White**: No data available

1. See appendix for detailed definitions and metrics used for each of the criteria.

SOURCE: McKinsey Global Institute analysis
A data management ecology?

The role of the top-down and the bottom-up?

Sum (little science data) > Sum (big science data)?

Massive data loss

Tier 1: major international resources: eg Worldwide Proton Databank

Tier 2: national data centre: eg UK Data Archive

Tier 3: institutional repository: eg ePrints Soton

Tier 4: individual collections

Breadth of the value of data increases up the tiers: from individual to community to social value.

Each higher tier brings greater responsibility and demands for access.

And so, as infrastructure increases so must the attention given to standards, sustainability and provenance (see appendix 2).
Views of young scientists

• The generation gap: younger researchers typically produce more data; recognise data sharing as maximising value; have most potential to develop data sharing tools; and they are the future. We should listen to them!

1. a shift away from a research culture where data is viewed as a private preserve
2. the data evidence for a published argument MUST be intelligently open at the time of publication
3. data management should be embedded in the community producing and using the data
4. science data should be as easy to "remix" as music is to a DJ
5. replication is by far the best guarantee of preservation (e.g. LOCKSS)
6. give credit for useful data communication and novel ways of collaborating
7. common standards for communicating data (correct?)
8. the cost of intelligent openness is an integral part of the cost of doing science
9. Training and support
Essential enabling tools & processes:

key issues for research & implementation

- data integration
- supporting dynamic data
- providing provenance
- annotation
- metadata generation
- citation
- access to data scientists
- changing the library
## Scripts for the actors in open science

<table>
<thead>
<tr>
<th>Scientists – changing cultural assumptions</th>
<th>Publishers of research – mandating open data; open up to data mining; be careful not to be obstacles to the progress of science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employers (universities/institutes) – data responsibilities; crediting researchers; the role of libraries</td>
<td>Business – exploiting the opportunity; awareness &amp; skills</td>
</tr>
<tr>
<td>Funders of research - the cost of curation is a cost of research</td>
<td>Government – efficiency of the science base; exploiting its data</td>
</tr>
<tr>
<td>Learned societies – influencing their communities</td>
<td>Governance processes for privacy, safety, security - proportionality</td>
</tr>
</tbody>
</table>
Challenges for universities

- Will they rise to the scientific challenge, or leave things to the information business?
- Will they be responsible for the knowledge they create?
- The university library; doing the wrong things through the wrong people?
- Adapting scientific education?
- Training data scientists?
- Supporting the data manipulation needs of their researchers?
- Supporting intelligent openness
- Open data and commercial imperatives
The levels of influence

**National**
E.g. UK: Government “Transparency Boards” (Research, Business, Govt data) – chaired by Minister for Science

**European**
DGs Connect & Research

**International**
ICSU (International Scientific Unions)
CODATA
UK-US-Chinese-Indian science academies

**BUT:** the science community is the driver of creative, workable, flexible solutions – the roles of the above bodies are:

1. Remove barriers
2. Intelligent facilitation
Challenge for the Commission as a funder of science

Top-down (present understanding)

Bottom-up (new knowledge and experiment)

Optimal Flexible Solution?

Digital European Research Area
1. Developing an open, interoperable e-infrastructure
2. Organising the European data space, through an open science policy
3. Opening communities, engaging individuals

... and remember - science is international!
A realiseable aspiration: all scientific literature online, all data online, and for them to interoperate

... and don’t forget, this is a process, not an event!
Science as an open enterprise

June 2012

THE ROYAL SOCIETY