Open Data and the Future of Science

Geoffrey Boulton

Munin Conference on Scholarly Publishing

Universitetet i Tromso
November 2014
A taxonomy of openness

Open science

Doing science openly

Collecting the data
Doing research

Input

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)

Output

Open access

Research publications (i.e. papers in journals)

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

Science as a public enterprise & the future of the open society
A realiseable aspiration: all scientific literature open & online, all data open & online, and for them to interoperate… and to be accessible to all?
Open communication of data: the source of a scientific revolution and the basis of scientific progress
Scientific self correction

The progress of science is strewn, like an ancient desert trail, with the bleached skeleton of discarded theories which once seemed to possess eternal life.

(Arthur Koestler)

False facts are highly injurious to the progress of science, for they often long endure; but false views, if supported by some evidence, do little harm, as everyone takes a salutary pleasure in proving their falseness; and when this is done, one path towards error is closed and the road to truth is often at the same time opened.

(Charles Darwin)
Problems & opportunities in the data deluge

90% of all the data in the world has been generated over the last 2 years.

Scientific data output increases at an annual rate of 30%.

10^20 bytes

DATA GROWTH

IT BUDGET SHORTFALL

Available storage

IT BUDGETS (INCREASE)

COST OF STORAGE/GB (DECREASE)

2011  2012  2013  2014  2015
The Challenge: the “Data Storm” is undermining “self correction”
**A crisis of replicability and credibility?**

---

**Reproducibility of Research Findings**

Preclinical research generates many secondary publications, even when results cannot be reproduced.

<table>
<thead>
<tr>
<th>Journal impact factor</th>
<th>Number of articles</th>
<th>Mean number of citations of non-reproduced articles*</th>
<th>Mean number of citations of reproduced articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;20</td>
<td>21</td>
<td>248 (range 3–800)</td>
<td>231 (range 82–519)</td>
</tr>
<tr>
<td>5–19</td>
<td>32</td>
<td>169 (range 6–1,909)</td>
<td>13 (range 3–24)</td>
</tr>
</tbody>
</table>

Results from ten-year retrospective analysis of experiments performed prospectively. The term ‘non-reproduced’ was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.


---

**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 as scientific **MALPRACTICE.**
“Scientists like to think of science as self-correcting. To an alarming degree, it is not.”
Seizing the opportunities

The opportunity: 1. identifying hitherto unresolvable patterns in phenomena

Enabling agreements/solutions needed:
- low access thresholds
- metadata
- integration
- provenance
- persistent identifiers
- standards
- data citation formats
- algorithm integration
- file-format translation
- inter-operability
- software-archiving
- automated data reading
- metadata generation
- timing of data release
- certification
- “fair data”

The semantic web?

The opportunity: 2. data-modelling: iterative integration

Satellite observation
Initial conditions
Model forecast
Model-data iteration - forecast correction

The opportunity: 3. deepening data integration

Scientific opportunity

4500 Variables: e.g.
- Annual Precipitation
- Annual Temperature
- Anthropogenic impacts on Marine Ecosystems
  - Nutrient Pollution (Fertilizer)
  - Aquaculture Production - Inland Water
  - Aquaculture Production - Marine
  - Aquaculture Production - Total
  - Arable Land
  - Arable and Permanent Crops
  - Arsenic in Groundwater - Probability of

Commercial opportunity

Monsanto

In order to:
Predict agricultural yields to ascend to "the next level of agricultural evaluation"

Purchases
For $930 million

Historic rainfall & infiltration data
Soil properties & quality

The opportunity: 4. linked sensors & machine learning

The “Internet of Things”

December 2005
1st
2nd
3rd
4th

Air conc. (NH₃)
Wet dep.
Dry dep.

acquisition – integration analysis - feedback
But seizing these opportunities depends on an ethos of data-sharing

Example:
ELIXIR Hub (European Bioinformatics Institute) and ELIXIR Nodes provide infrastructure for data, computing, tools, standards and training.
EXAMPLES OF WHERE AN OPEN DATA ETHOS OPERATES OR IS DEVELOPING

Operating
• Crystallography
• Genomics/Bioinformatics

Developing
• Geosciences
• Chemistry
• Ecology
• Longitudinal studies in social statistics
Data sharing for emergencies & global challenges

**e.g. Response to Gastro-intestinal infection in Hamburg**
- E-coli outbreak spread through several countries affecting 4000 people
- Strain analysed and genome released under an open data license.
- Two dozen reports in a week with interest from 4 continents
- Crucial information about strain’s virulence and resistance

**e.g. Global challenges – e.g rise of antibiotic resistance**
- A global challenge that inevitably needs a global response based on data sharing

*MRSA = methicillin-resistant Staphylococcus aureus; VRE = Vancomycin-resistant enterococci*  
*FQRP = Fluoroquinolone-resistant Pseudomonas aeruginosa*
But it is also vital that we apply appropriate statistical approaches and techniques to our data

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!”

....and Big Data compounds the problem.

So what are the priorities?

1. Ensuring valid reasoning
2. Innovative manipulation to create new information
3. Effective management of the data ecology
4. Education & training in data informatics & statistics
...and a new fundamental debate in the petabyte world

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

Thesis: Correlation is not causation.

Anti-thesis: Correlation is enough.

Question: If we know “how things are”, do we need to know “why they are?”
The nightmare: disconnect between machine analysis & human cognition

What is the human role?
Can we analyse & scrutinise what is in the black box?
What does it mean to be a researcher in a data intensive age?
Who owns the box: the tragedy of the commons in understanding?
The future of “science”? 

The present/future

Open data = Science

Closed data ≠ Science
Openness of data *per se* has little value: open science is more than disclosure

For effective communication, replication and re-purposing we need **intelligent openness**. Data, meta-data and, increasingly software/machine codes must be:

- Discoverable
- Accessible
- Intelligible
- Assessable
- Re-usable

Only when these criteria are fulfilled are data properly open.

**But, intelligent openness must be audience sensitive.**

Open data to whom and for what?

- Scientists – Citizen scientists – Citizens
Boundaries of openness?

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

- Legitimate commercial interests (sectoral variation)
- Privacy ("safe data" v open data – the anonymisation problem)
- Safety, security & dual use (impacts contentious)

All these boundaries are fuzzy
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?
a changing social dynamic in science?

Citizen science

Opening the evidence to public scrutiny
Open data & the inhibition of scientific fraud

“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
Infrastructure: e.g. 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
• interactivity
• access anywhere, anytime
Changing and adapting: whose responsibilities?

- **Scientists:** changing the mindset
- **Learned Societies:** influencing their communities
- **Universities/Insts:** incentives & promotion criteria, proactive, not just compliant, the library function, 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:** mandate concurrent open deposition
- **Governments & the EU:** do not over-engineer an ecology with emergent properties

Its mostly people & institutions – not systems, regulation & hardware
Don’t preach – Incentivise

Researchers
• Advancement & promotion
• Data citation – 2 for the price of 1

Universities/institutes
• Funding incentives for open data
• Greater potential for scientific value
International

CODATA
- Standards
- Protocols
- Tools
- Interoperable systems

Research Data Alliance
- Domain specific solutions
- Community stimulation

Data Bases
- WDS
- GEO
- Etc

Inter-Govt support
- Horizon 2020
- G8 statement
- Obama White House

Systems

National

Funding bodies
- Research Councils
- University Funding Councils
- Research charities

Research performers
- Universities
- Institutes

Learned societies
- National academies
- Disciplinary societies

Technical bodies
- British library
- JISC
- PLOS
- etc

Janus

Publishers
UK 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 the rationale, principles, processes and priorities
• coherent approach across the research process
• consistent with and influencing international developments
• 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
• Joint development of SHARE with US “Coherence committee”

Dangers on the flank
• Publishers inhibition of text and data mining
• EU confidentiality regulation
A data infrastructure ecology: drivers and self-organising components (the rationale for the UK Open Data Forum)

- **Mandate open data**
- **Intelligently open data**
- **Sustainable**
- **Interoperable**
- **Persistent identifiers**
- **Metadata standards**
- **Dynamic data etc**

---

**Public & Charitable funders**
- Mandate intelligently open data
- Common standards & protocols

**Universities/institutes**
- Promotion & reward
- Data science
- Support data management
- Incentivise data stewardship
- Training

**Researchers**
- Data custodians not owners
- Depositing citeable data

**Publishers**
- Mandate concurrent intelligently open data
- Easy text & data mining

---

**Public access**

**Tools for:**
- Discovery
- Integration
- Management
- Metadata
- ETC

---

**Databases/repositories**
- Intelligently open data
- Sustainable
- Interoperable
- Persistent identifiers
- Metadata standards
- Dynamic data etc

---

**Learned societies**

**RESOURCES**
A taxonomy of openness

Open science

Open data

Open access

Inputs

Collecting the data
Doing research

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

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