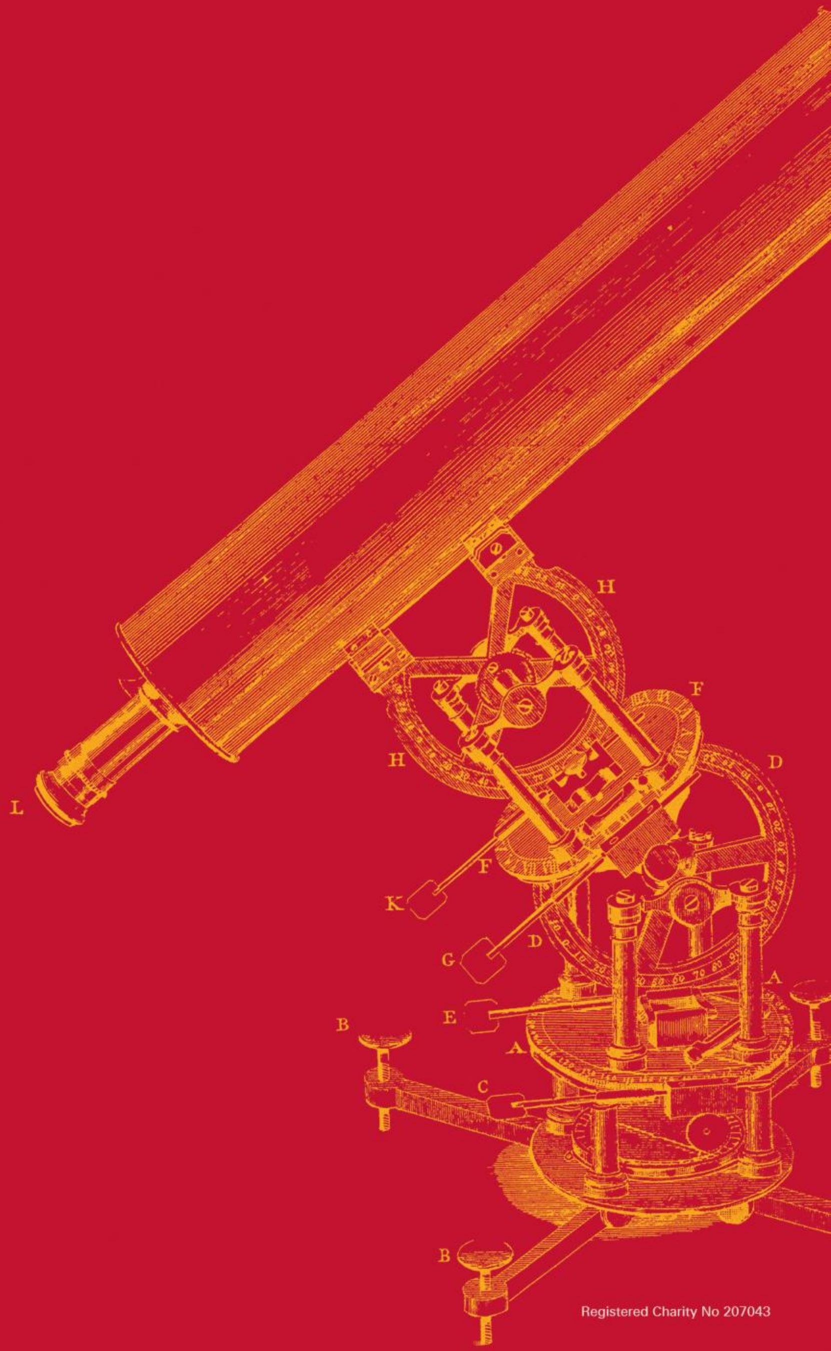
Science as an open enterprise

Geoffrey Boulton

Open Access Week

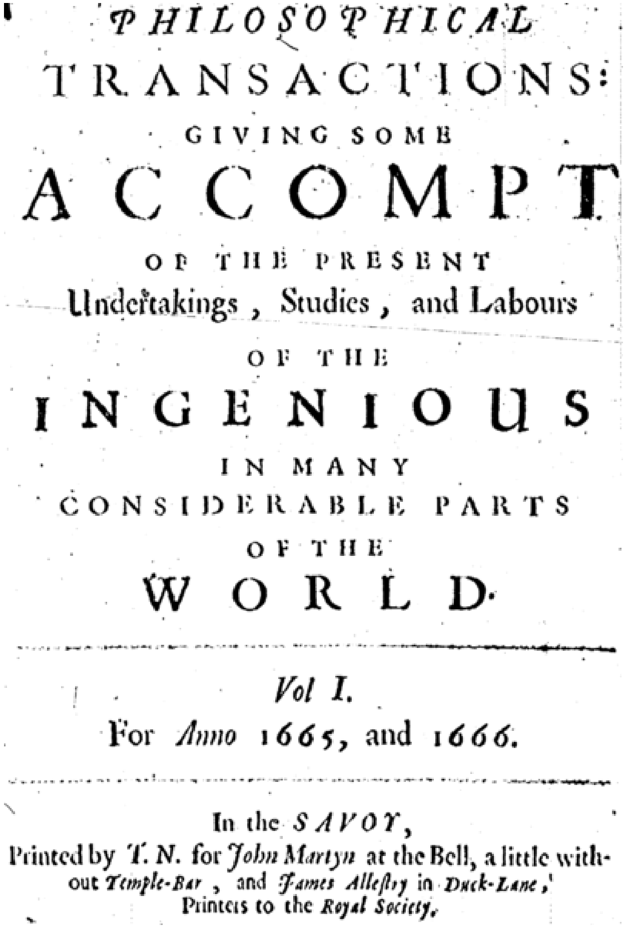
Edinburgh
October 2013



THE
ROYAL
SOCIETY

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

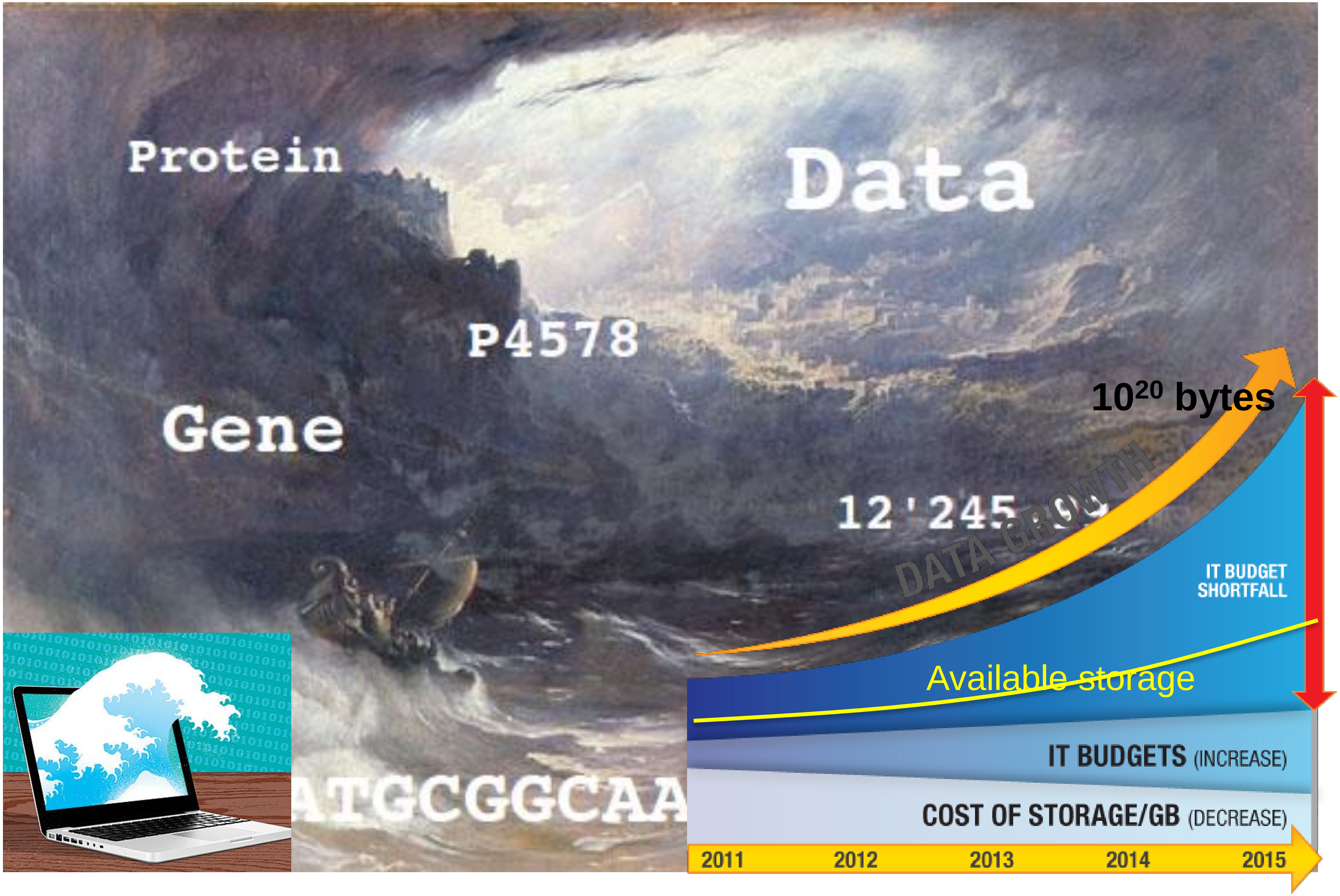
Henry Oldenburg



PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Underſtakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For *Anno* 1665, and 1666.

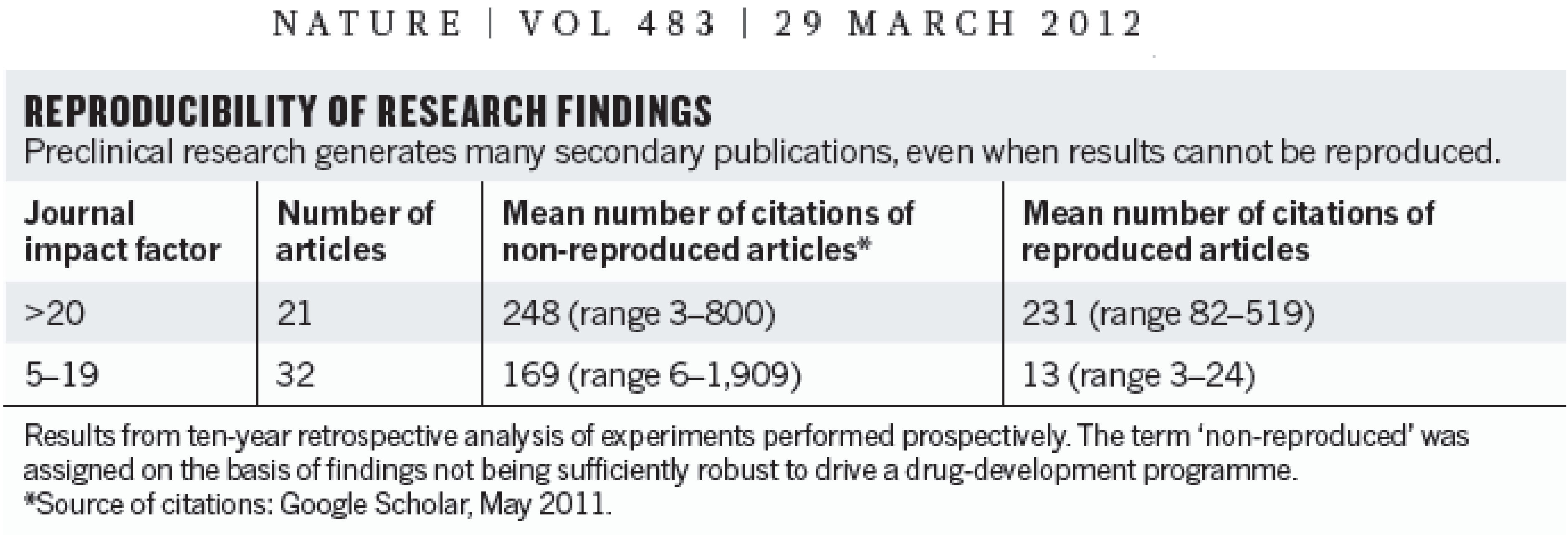
In the *SAVOY*,
Printed by *T. N.* for *John Martyn* at the Bell, a little without *Temple-Bar*, and *James Alleſtry* in *Duck-Lane*,
Printers to the *Royal Society*.


Protein
Data
P4578
Gene
12'245.99
ATGCGGCAA
DATA GROWTH
10^{20} bytes
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”

THEN AND NOW

A crisis of replicability and credibility?



NATURE | VOL 483 | 29 MARCH 2012

REPRODUCIBILITY OF RESEARCH FINDINGS

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

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

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.

*Source of citations: Google Scholar, May 2011.

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

But what about the vast data volumes that are not used to support publication?

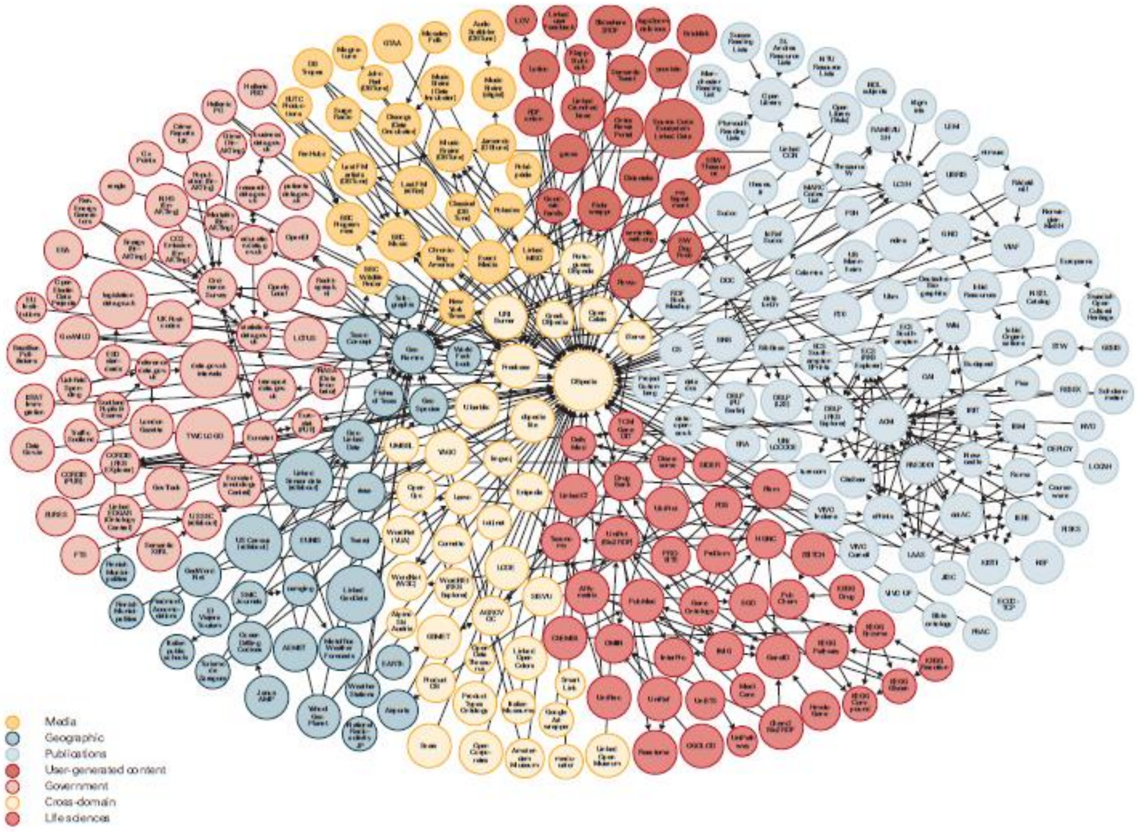
The opportunity: new scientific knowledge from data

Exploiting the potential of linked data requires:

- data integration
- dynamic data

Solutions/agreements are needed for:

- provenance
- persistent identifiers
- standards
- data citation formats
- algorithm integration
- file-format translation
- software-archiving
- automated data reading
- metadata generation
- timing of data release



Its not just accumulating and linking data– its 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!"

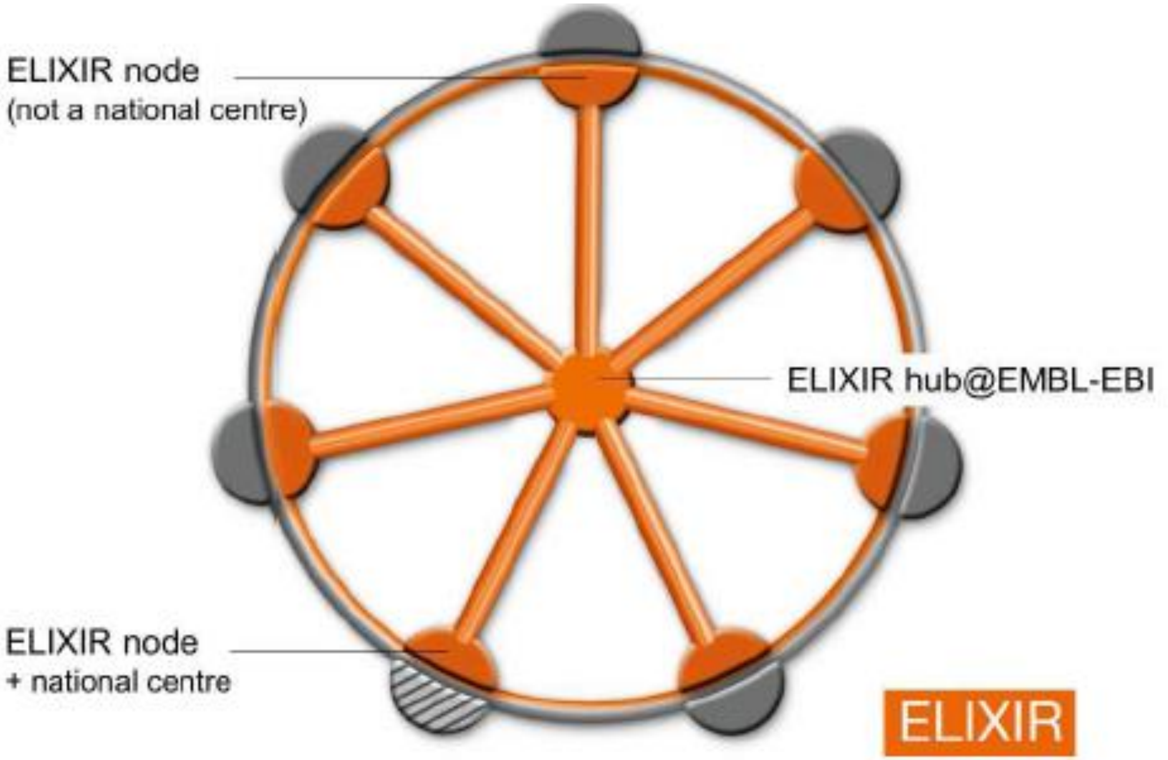
So what are the priorities?

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 we need a new breed of informatics-trained data scientist as the new librarians of the post-Gutenberg world

A new ethos of data-sharing?

Example:
ELIXIR Hub (European Bioinformatic Institute) and ELIXIR Nodes provide infrastructure for data, computing, tools, standards and training.



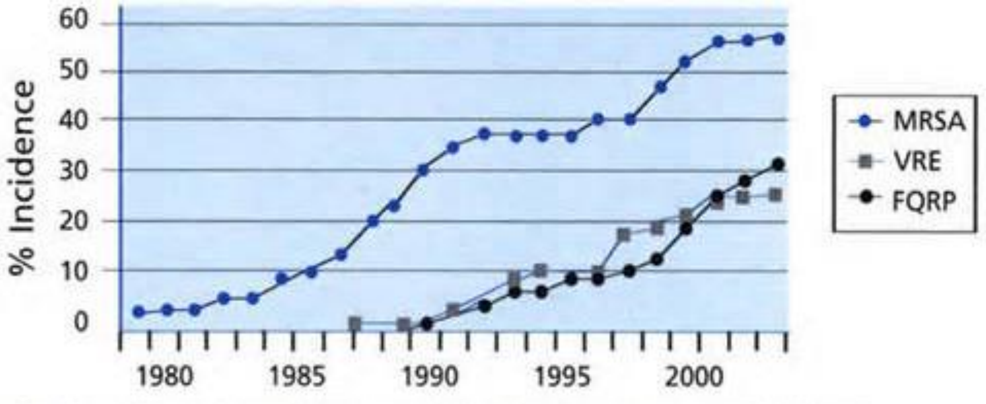
Benefits of open science:

1. 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

2. Global challenges – e.g rise of antibiotic resistance

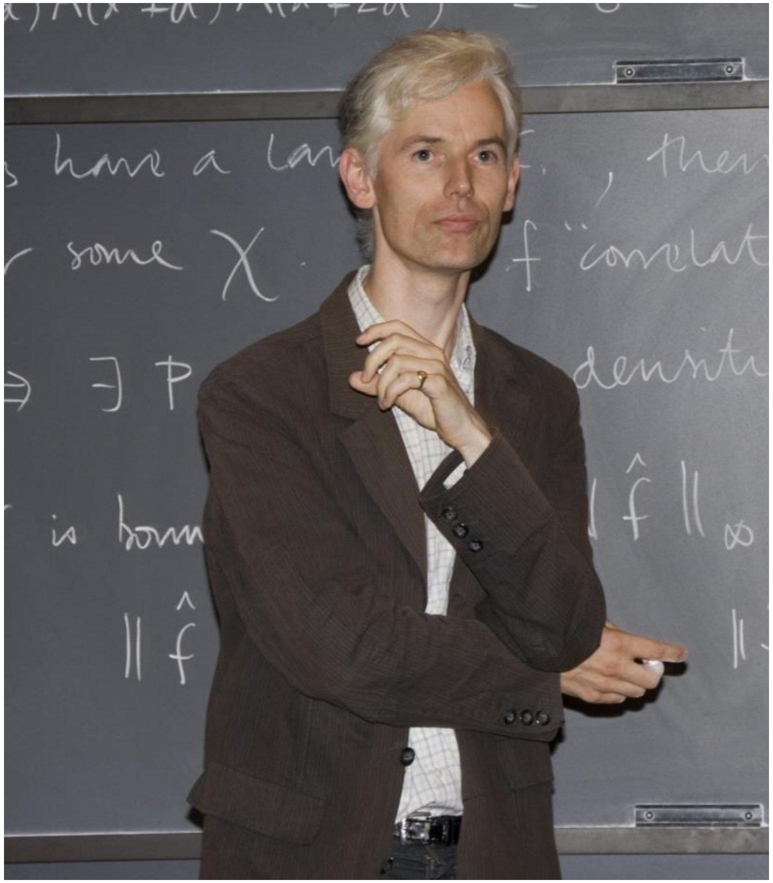
- A global challenge that Inevitably needs a global response



MRSA = methicillin-resistant *Staphylococcus aureus*; VRE = Vancomycin-resistant *enteroccoci*
FQRP = Fluoroquinolone-resistant *Pseudomonas aeruginosa*

Benefits of open science:

3. Crowd-sourcing



Tim Gowers
- crowd-sourced mathematics

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”

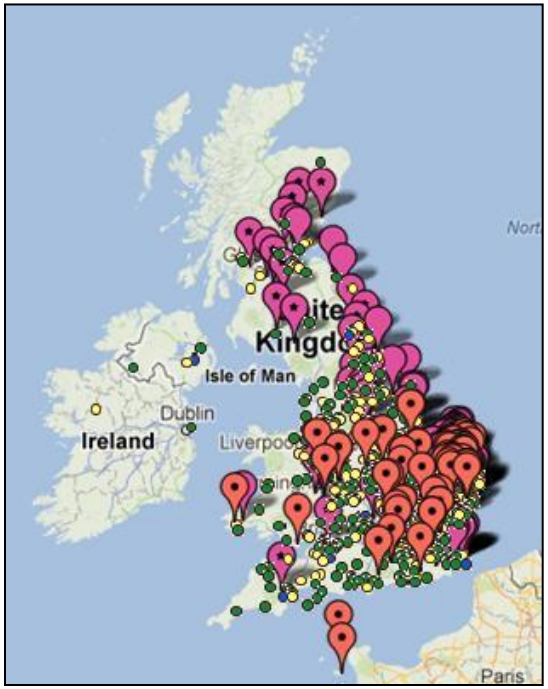
What inhibits such processes?
- The criteria for credit and promotion.

4. …… & the changing social dynamic of science

Citizen science

ASHTAG

Available on the App Store



Openness to public scrutiny

5. Fraud and malpractice

theguardian

"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

Malpractice

- **Non-publication of evidence for a published claim"**
- **"Cherry-picking" data & selective publication**
- **Partial or biased reporting – e.g. clinical trials**
- **Failure to publish refutation**

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 and meta-data must be:

- **Accessible**
- **Intelligible**
- **Assessable**
- **Re-usable**

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

But, intelligent openness must be audience sensitive.

Open data to whom and for what?

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

Responsibilities & actions

- **Scientists:** - changing the mindset
- **Learned Societies:** - influencing their communities
- **Universities/Insts:** - 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:** - mandate concurrent open deposition
- **Governments & the EU:** - do not over-engineer an ecology with emergent properties

Its mostly people & institutions – not systems, regulation & hardware

Can libraries rise to the challenges of a post-Gutenberg world?

"Libraries do the wrong things, employ the wrong people"

People

- Funders mandate novel customers – the public
- Can they attract data scientists?
- Support for researchers & students

Policies

- Reversing centralisation
- A data repository – directory - metadata – background
- Dynamic data
- Selection problem
- Compliant or proactive?

How are institutions responding?

International

- G8 statement
- OECD
- Engagement of ICSU bodies (e.g. CODATA)
- Inter-academy collaboration
- Research Data Alliance

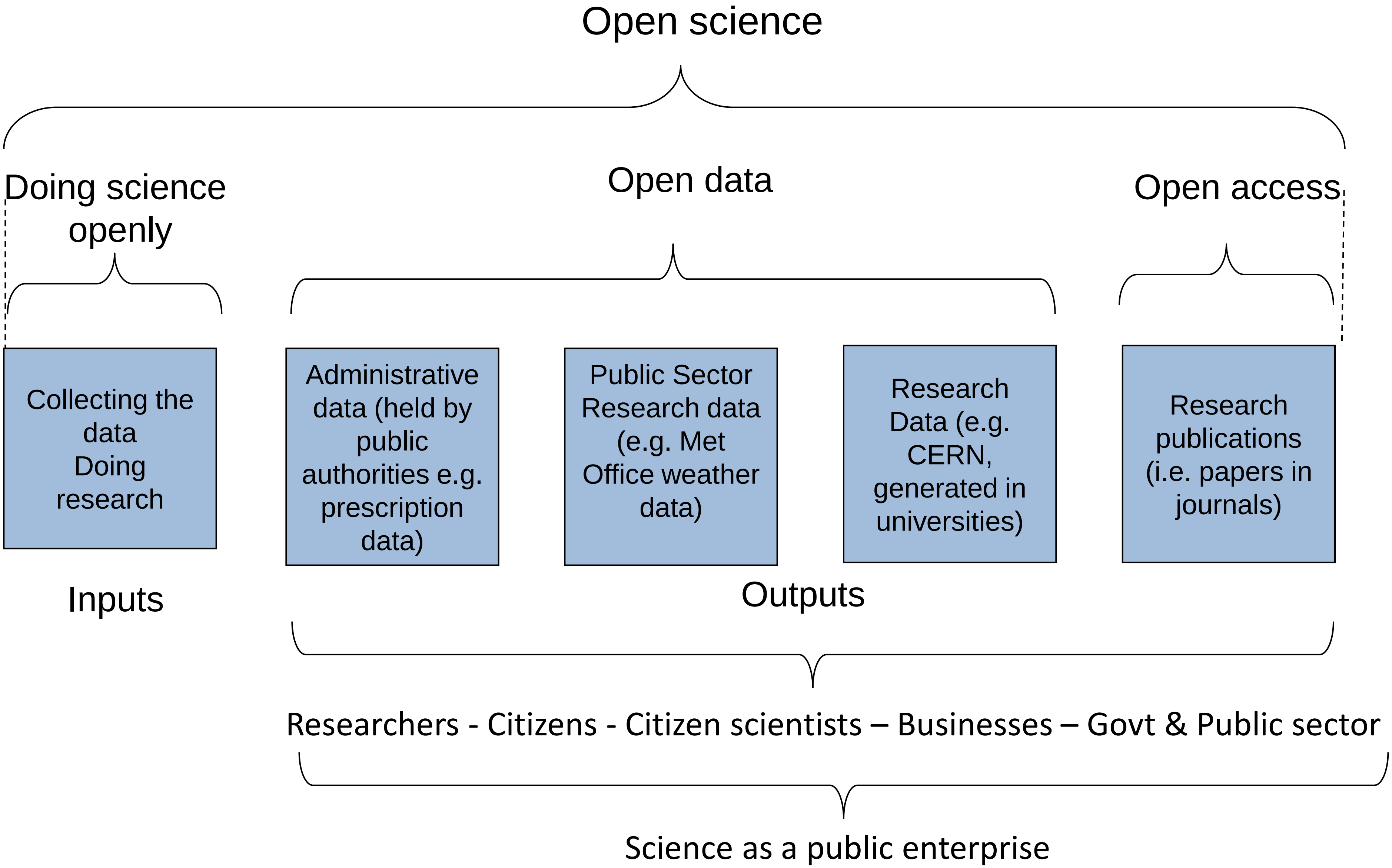
European

- A principle of Horizon 2020 (trial runs shortly)
- Engagement by EUA, LERU, LIBER
- EC initiatives (e.g. Medoanet)

UK

- Research Councils
- Government Research Data Transparency Board
- UK Science Data Forum

A taxonomy of openness


Open science

Doing science openly

Open data

Open access

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)

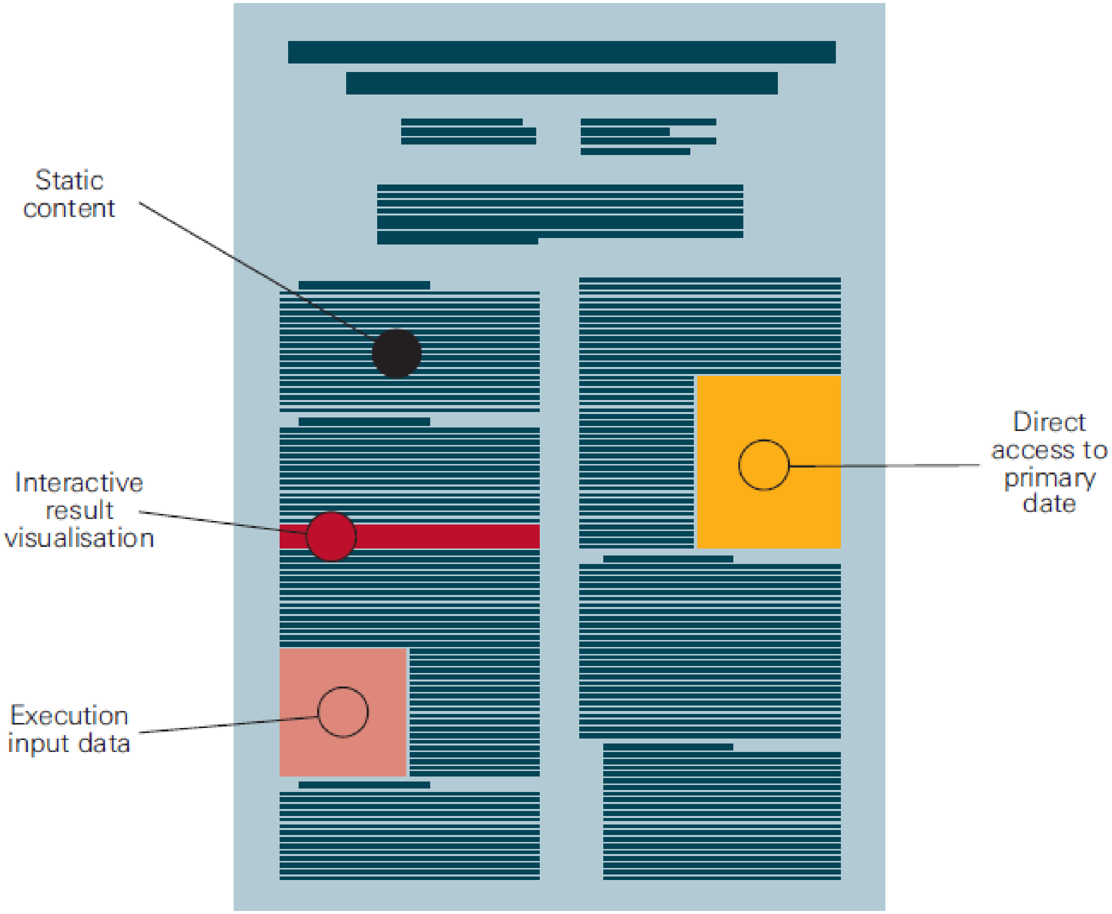
Inputs

Outputs

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

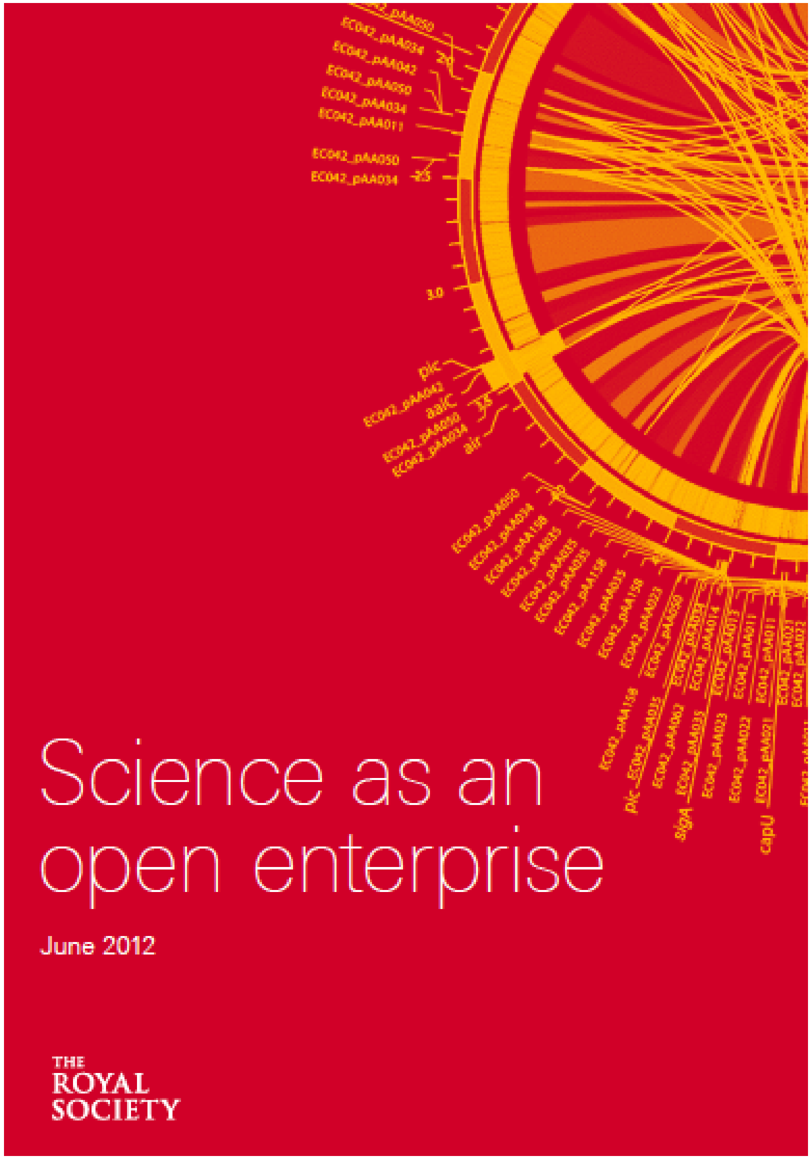
Science as a public enterprise

A realiseable aspiration: all scientific literature open & online, all data open & online, and for them to interoperate


Static content
Interactive result visualisation
Execution input data
Direct access to primary date

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

www.royalsociety.org



Science as an open enterprise

June 2012

THE
ROYAL
SOCIETY