

Open data infrastructure for global challenges

Geoffrey Boulton

ICRI, Athens

April 2014

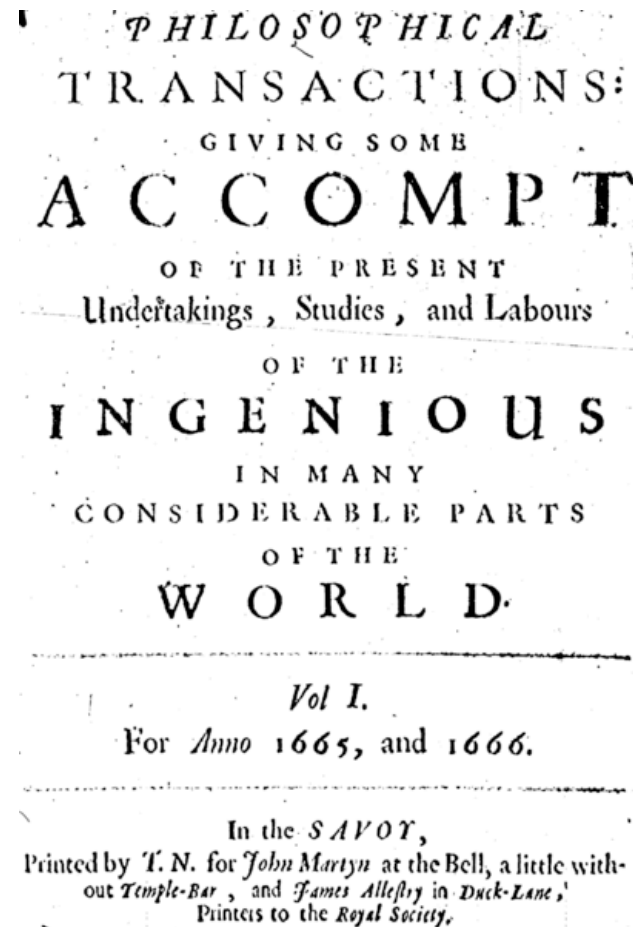
THE
ROYAL
SOCIETY



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



Henry Oldenburg

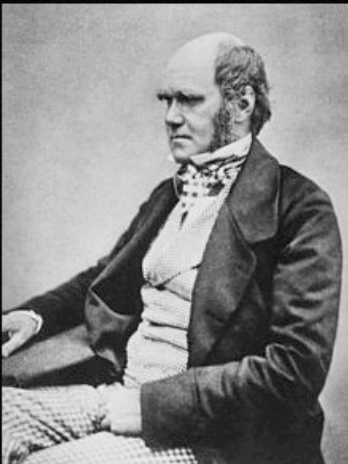


Open data & the principle of self-correction in science



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)

Protein

P4578

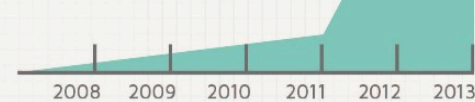
Gene

12'245'900

ATGCGGCAA

90%

of all the data in the world has been generated over the last 2 years¹



30%

JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

Scientific data output increases at an annual rate of 30%²

10²⁰ bytes

DATA GROWTH

IT BUDGET SHORTFALL

Available storage

IT BUDGETS (INCREASE)

COST OF STORAGE/GB (DECREASE)

2011

2012

2013

2014

2015

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.

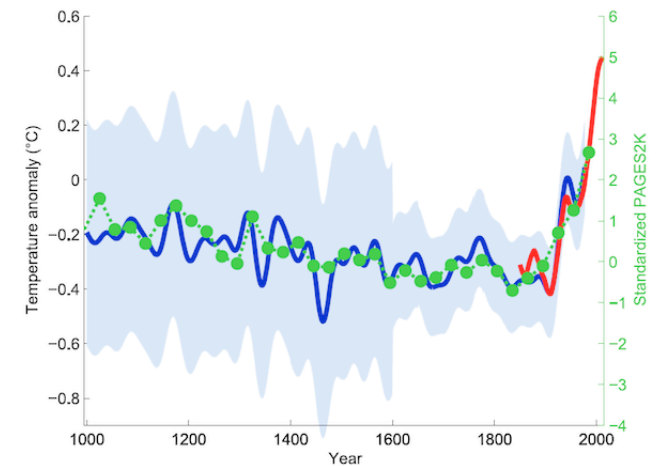
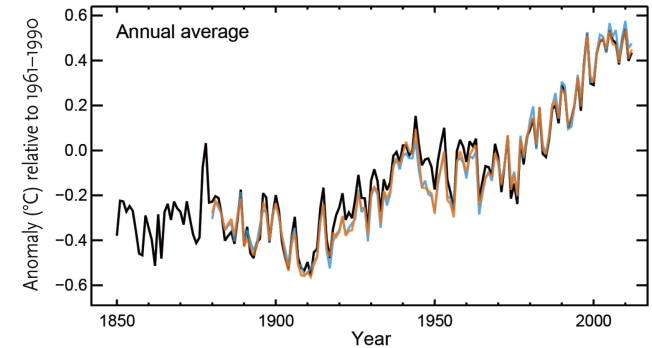
A fundamental principle: the data providing the evidence for a published concept MUST be concurrently published, together with the metadata

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.”



Openness in global challenges and openness to challenge



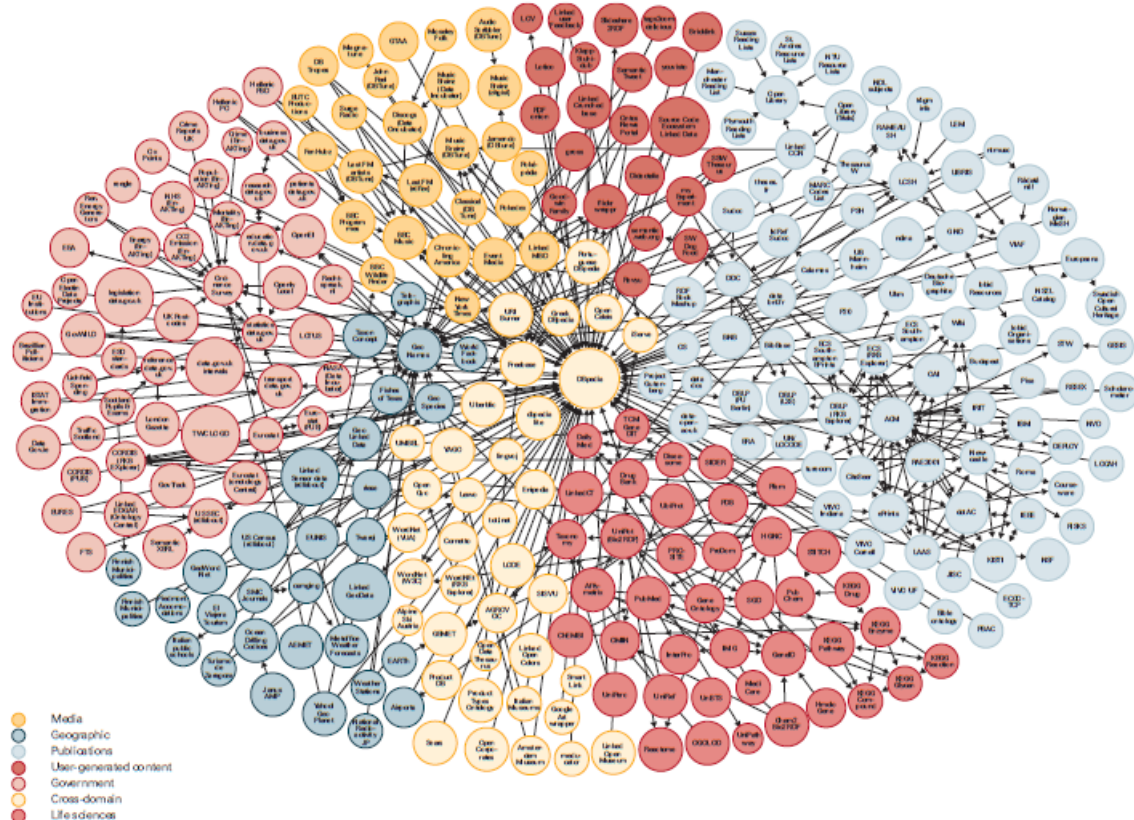
The opportunity: identifying novel patterns in data

Exploiting the potential of linked data requires:

- data integration
- dynamic data

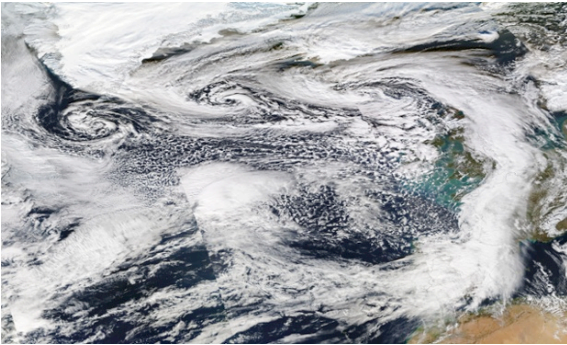
Solutions/agreements are needed for:

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

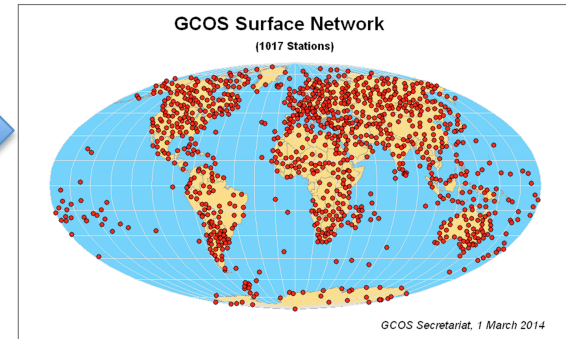


Data-modelling: iterative integration

Satellite observation

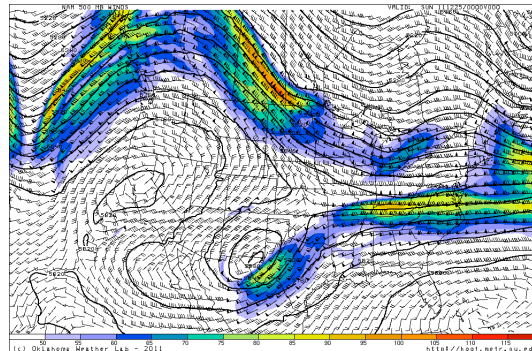


Surface monitoring

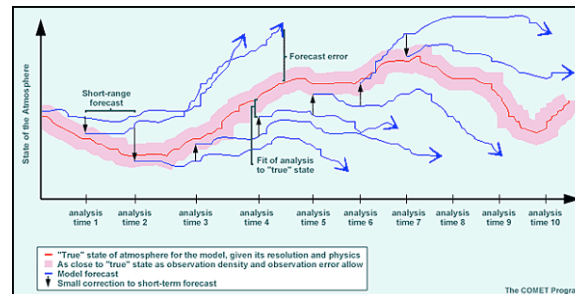


Initial conditions

Model forecast



Model-data iteration - forecast correction



Deepening data integration

Scientific opportunity



4500 Variables: e.g.

Annual Precipitation

Annual Temperature

Anthropogenic impacts on

Marine Ecosystems

- Nutrient Pollution (Fertilizer)

Aquaculture Production - Inland Waters

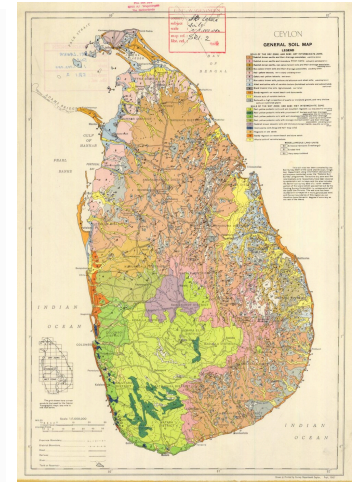
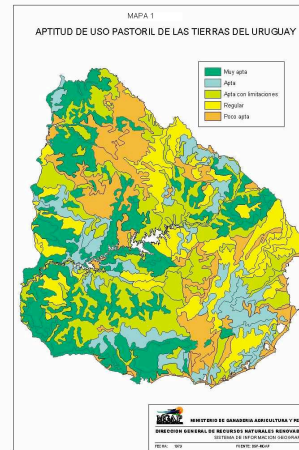
Aquaculture Production - Marine

Aquaculture Production - Total

Arable Land

Arable and Permanent Crops

Arsenic in Groundwater - Probability of



Commercial opportunity



Purchases

For \$930 million



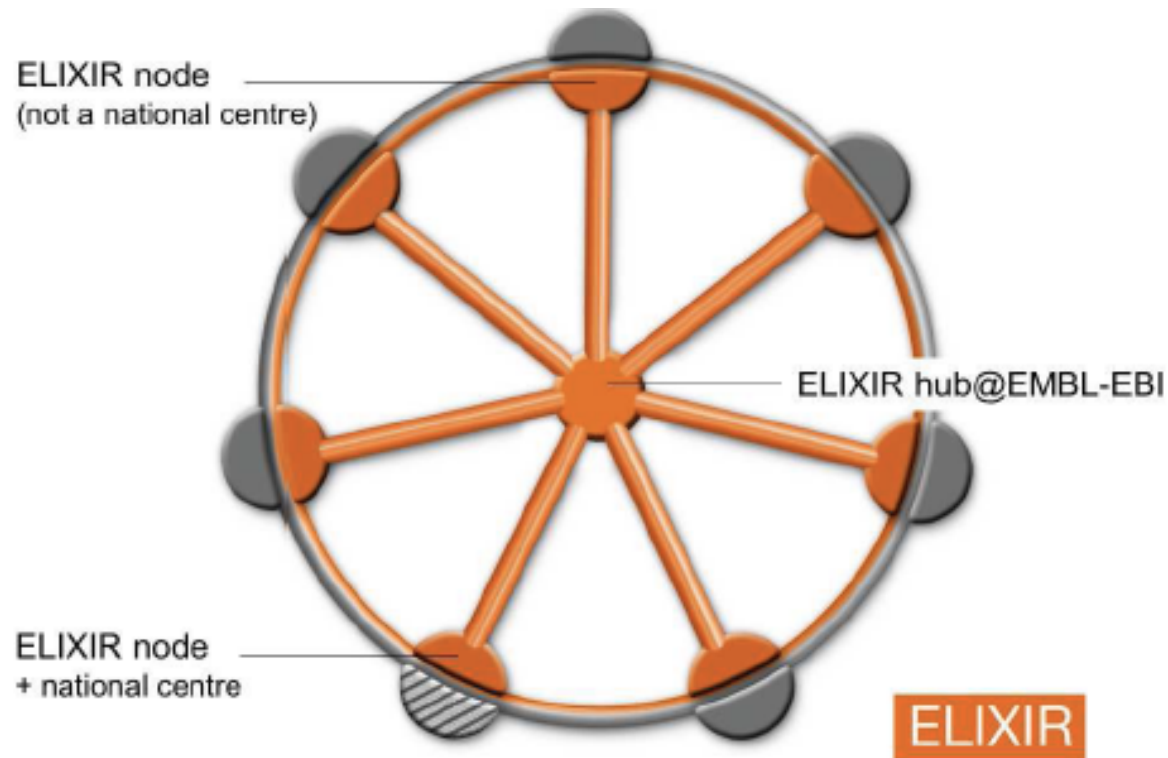
In order to:
Predict agricultural yields to ascend to
“the next level of agricultural evaluation”

Historic rainfall & infiltration data
Soil properties & quality

An ethos of data-sharing

Example:

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



Benefits of open data sharing

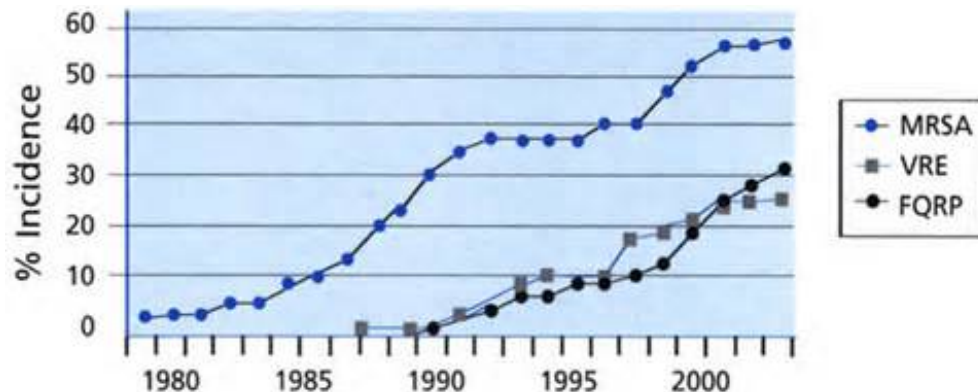
e.g. Response to Gastro-intestinal infection in Hamburg

- 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*

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:

- **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?

**Its not just accumulating and linking data/
information– 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**

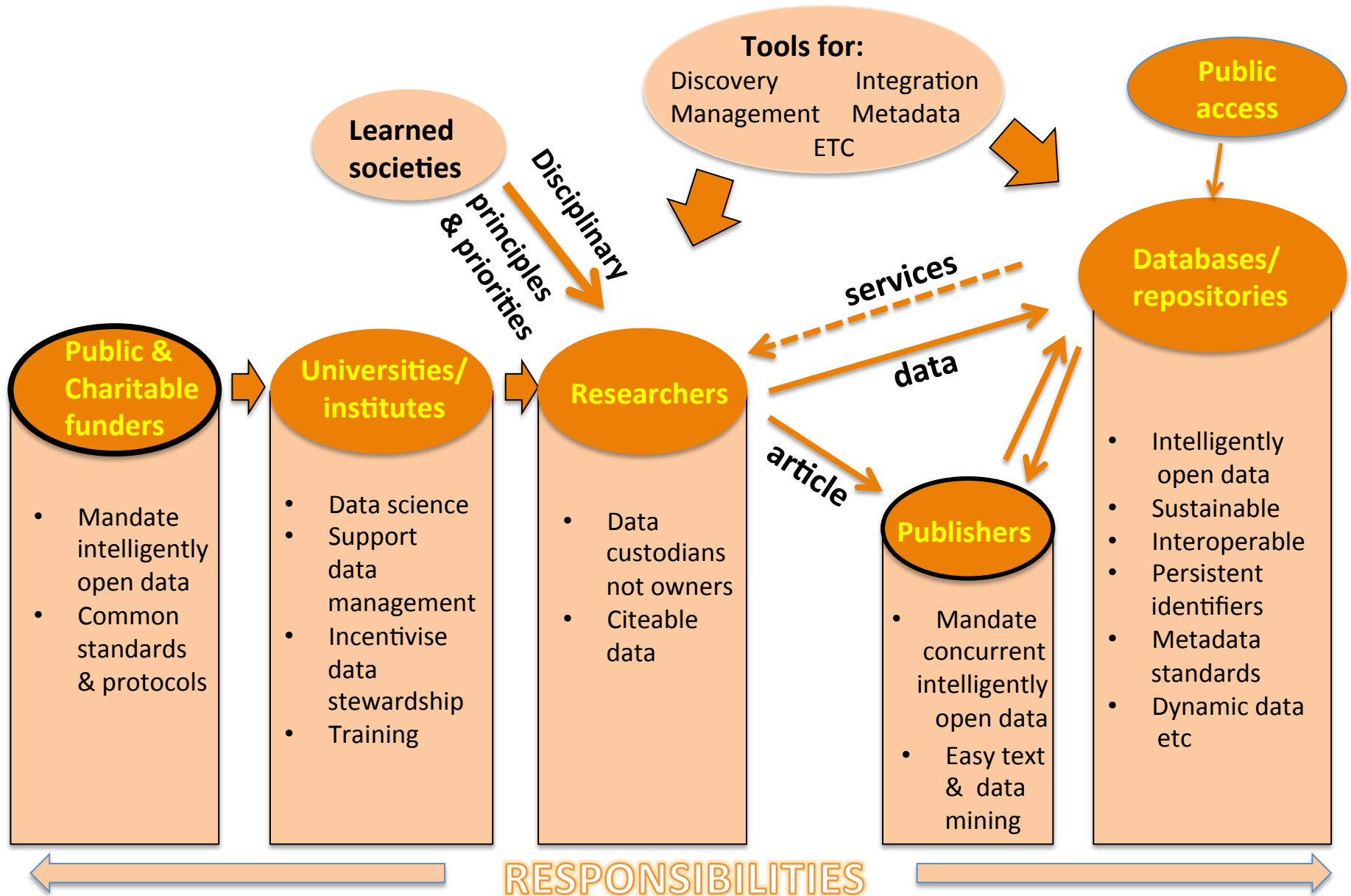
Boundaries of openness?

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

A data infrastructure ecology: drivers and self-organising components



Examples of evolving stimulation/coordination

THE
ROYAL
SOCIETY

UK OPEN RESEARCH
DATA FORUM

National

Horizon2020

European



Research Data Sharing
without barriers

Global



[illegible]