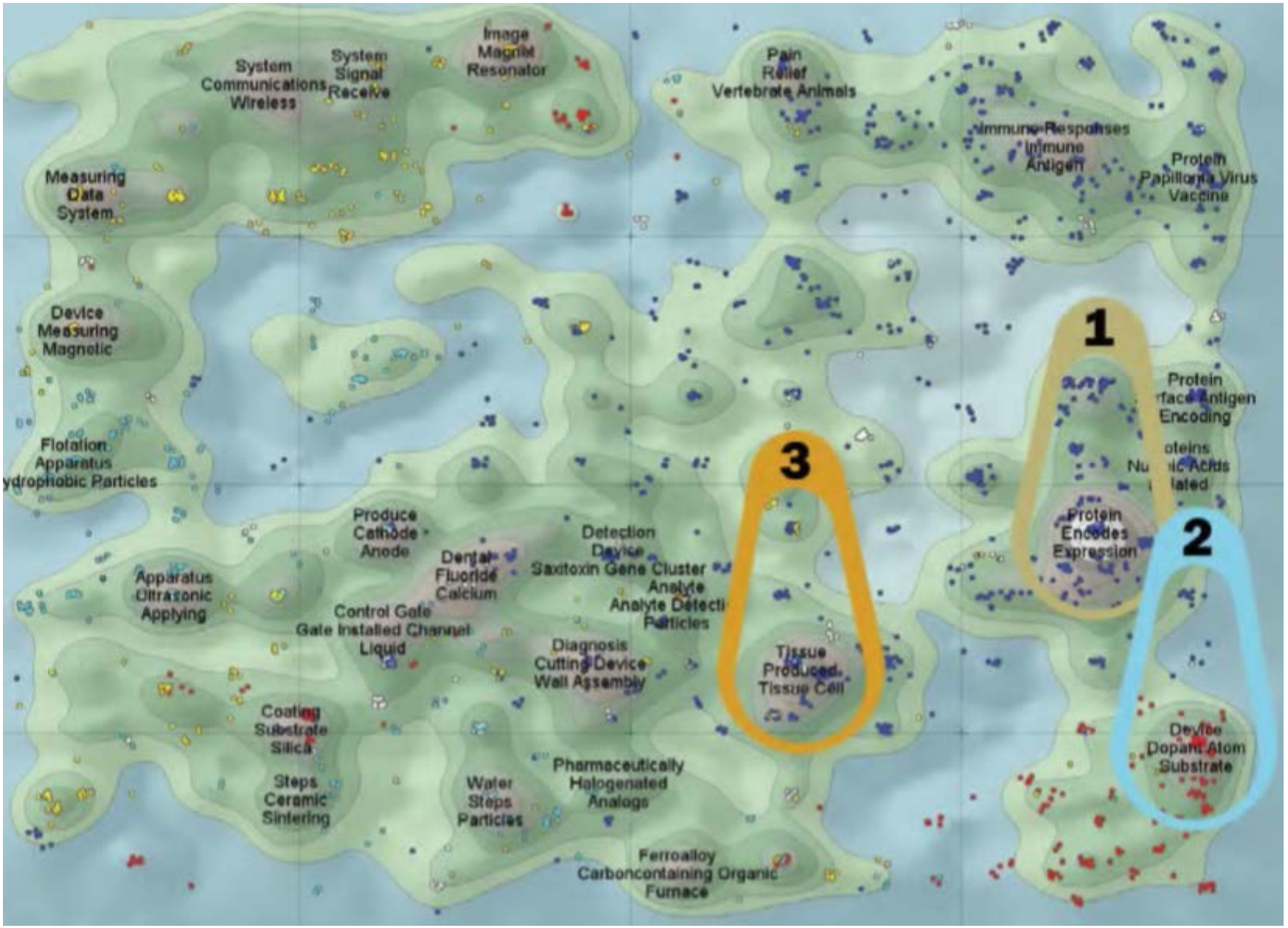
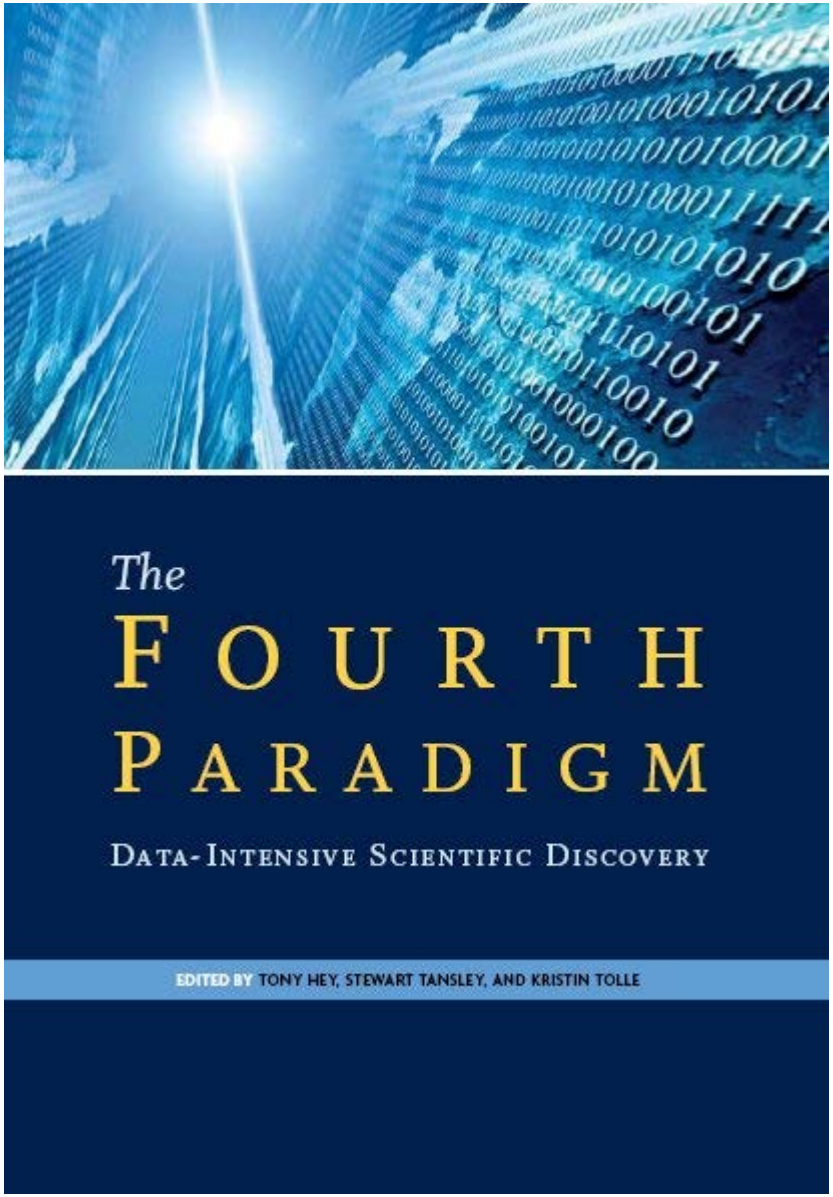
Managing Open Scientific Data: From FEAR to FREEDOM


System Communications Wireless
System Signal Receive
Image Magnit Resonator
Pain Relief Vertebrate Animals
Immune Responses Immune Antigen
Protein Papilloma Virus Vaccine
Measuring Data System
Device Measuring Magnetic
Flotation Apparatus Hydrophobic Particles
Produce Cathode Anode
Dental Fluoride Calcium
Detection Device Saxitoxin Gene Cluster Analyte
Analyte Detecti Particles
Apparatus Ultrasonic Applying
Control Gate Gate Installed Channel Liquid
Diagnosis Cutting Device Wall Assembly
Tissue Produced Tissue Cell
Protein Encodes Expression
Protein Encoding
Coating Substrate Silica
Steps Ceramic Sintering
Water Steps Particles
Pharmaceutically Halogenated Analogs
Ferroalloy Carboncontaining Organic Furnace
Dopant Atom Substrate
1
2
3

Data, computers and collaboration extending the frontiers of science



FOSTER

Key differences between open publications and open data

Publications

- Available in required formats at the time of publication or earlier, no additonal effort required
- Incentives to publish and increase impact.
- Libranians play a key role.
- One size fits all.

Data

Require curation and substantial inputs from researchers.

No incentives to curate and share data.

Research managers and data scientists to play a key role.

Always contextual

FOSTER

Why OPEN DATA?

"Ten or 20 years ago we might have been able to repeat an experiment. They were simpler, cheaper and on a smaller scale. Today that is not the case.
So if we need to re-evaluate the data we collect to test a new theory, or adjust it to a new development, we are going to have to be able reuse it. That means we are going to need to save it as open data."

Heuer, Former CERN Director

FOSTER

Research Data Management

The process of organising, manipulating, storing, curating and using research data to enhance its preservation and access into the future.

Credit: Michael Gibbs

FOSTER

THE PROMISE of RDM

Digital curation of data for public release, that is making scientific data available and useful to unknown audiences, and for unanticipated purposes.

RDM REQUIRES COLLABORATION AND TRUST.

FOSTER

FEAR of CHANGE

"Individuals, work teams, departmental units, leadership groups and even whole organisations are naturally averse to change as they unknowingly protect themselves from making the very changes they most desire."

Robert Kegan and Lisa Laskow Lahey, *Immunity to Change*

FOSTER

Managing data? What DATA?

DATA is 'everything', means different things to different people – from BIG data to LITTLE data, RAW to PROCESSED data.

Open data is free for anyone to use, reuse and redistribute it – subject only and at most, to the requirement to attribute or share-alike.

Emergent agreement: Research data are the evidence used to support the findings and arguments in science, research or scholarship.

FOSTER

What is 'the evidence'?

- *data captured from instruments*
- *derived data*
- *Documents*
- *spreadsheets & databases*
- *lab notebooks ???*
- *visualisations*
- *models,*
- *software*
- *images, measurements and numbers.*

FOSTER

What is 'data'?

- **DATA is always contextual** – the definition varies not only across scientific disciplines but also across different projects and different people.
- **MINIMUM STANDARD:**
  - 'Data'
  - 'Metadata'
  - 'Software'???? – IP clearance and version control are necessary. If software cannot be shared, the version used should be listed.

FOSTER

What is 'data'?

- **DATA is always contextual** – the definition varies not only across scientific disciplines but also across different projects and different people.
- **MINIMUM STANDARD:**
  - 'Data'
  - 'Metadata'
  - 'Software'???? – IP clearance and version control are necessary. If software cannot be shared, the version used should be listed.

FOSTER

So what is 'data'?

- **DATA is always contextual** – the definition varies not only across scientific disciplines but also across different projects and different people.
- **MINIMUM STANDARD – OPEN DATA:**
  - 'Data'
  - 'Metadata'
  - 'Software'???? – IP clearance and version control are necessary. If software cannot be shared, the version used should be listed.

FOSTER

METADATA

Not just 'data about data'.

METADATA explains, describes, locates or otherwise makes it easier to retrieve, use, or manage the primary data.

National Information Standards Organisation

MAKES DATA SENSIBLE and USABLE.

ISSUE: No established protocols for recording metadata.

FOSTER

Open 'by default' or 'controlled access'

DEFAULT

- Data underpinning scientific publications
- Most geospatial data to facilitate access to public infrastructure and manage public health issues
- Genomic data

CONTROLLED

Most research data where data interpretation, assumptions and detailed analyses are required, e. g. particle physics, clinical trials.

FOSTER

LEVELS OF CONTROL (CERN)

Level 1: the content of scientific publications and data, such as figures embedded in these publications. OPEN

Level 2: simplified data formats for immediate re-use, for example, in education outreach activities. OPEN

Level 3: simulation data, along with software, workflow analyses and other documentation needed to reproduce published results.CLOSED

Level 4: raw data and software that enable full reconstruction of the CMS experiments. CLOSED

FOSTER

Barriers to sharing

FOSTER

Too many stakeholders, too many interests

- Data collectors/authors
- Data collaborators and colleagues
- Software developers
- Data curators
- Data managers
- Research team
- Principal Investigator
- Research subjects
- Citizen scientists

Research funders

Policy makers

Management of research organisations

Librarians

Data custodians

Publishers

FOSTER

FEARS of researchers

No time to curate data.

Validity of secondary data analyses

The lack of incentives to curate and share data

The lack of incentives to (re)use somebody else's data.

FOSTER

FEARS of policy makers

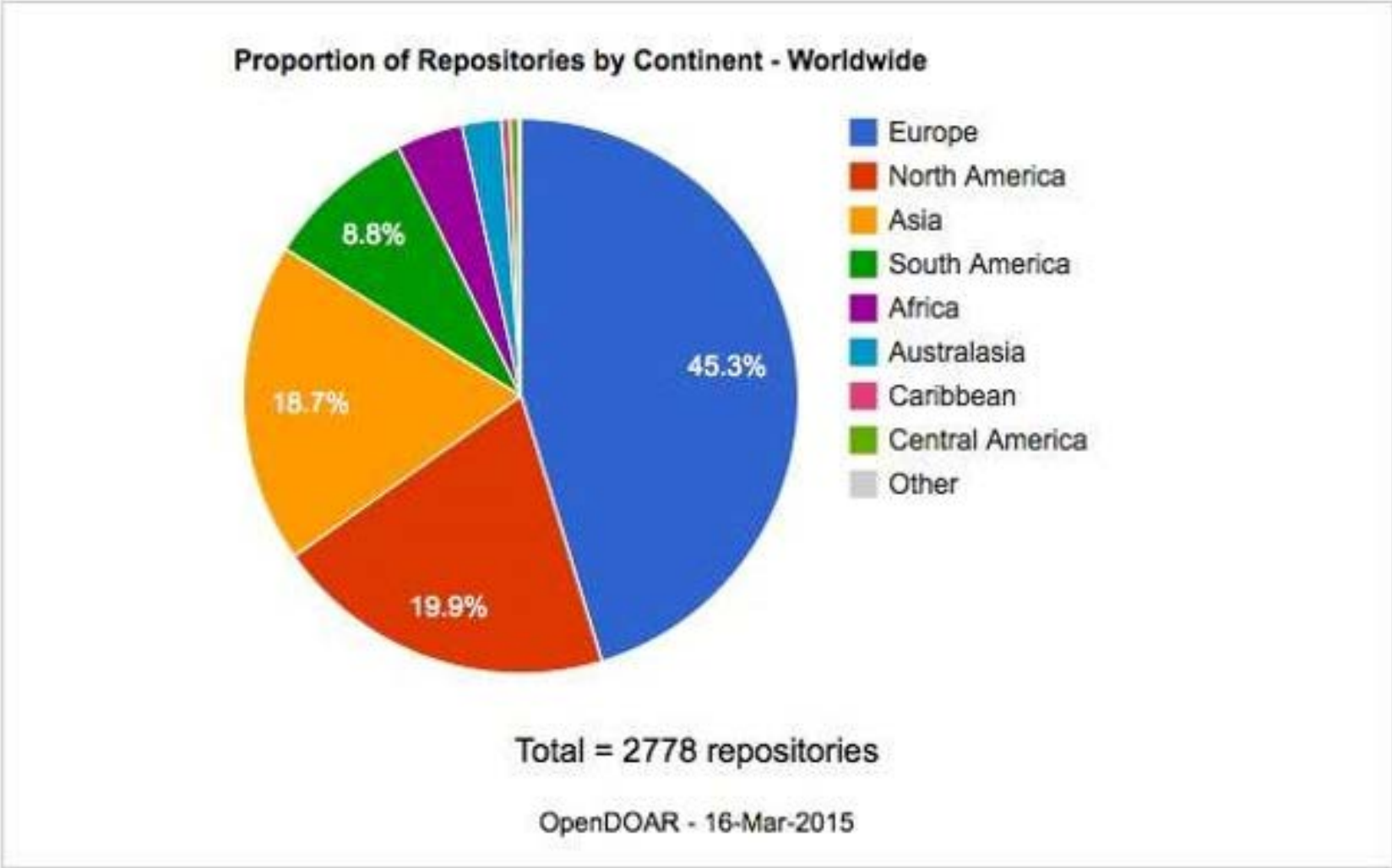
- Large investments in data, resources, and knowledge infrastructures.

PRESENT COST vs FUTURE BENEFIT

- Safeguarding the value of open data.
  - The capacity to use is limited to elites.
- Protecting commercial interest and IP.
- Protecting privacy of research subjects.

FOSTER

Where are the Open Data Elites?



FOSTER

KEY LESSONS LEARNT

The value of open scientific data lies primarily in their quality and the potential for future use and re-use.

More open scientific data does not necessarily mean more science, reproducible science or data-driven science.

Open Data requires an Open Mindset.

FOSTER

Key observations

- Open scientific data requires strong leadership and change management skills within research organisations.
- Unclear meaning of 'research data' acts as a barrier to sharing.
- Incentives for researchers and their organisations to curate, share and re-use data are required.
- Resources are finite and choices must be made about what data to curate and maintain. Need to develop 'data packages/products' rather than 'open data'.
- The value of open scientific data lies in their quality and potential for future (re)use. Metadata are the key to improved quality.