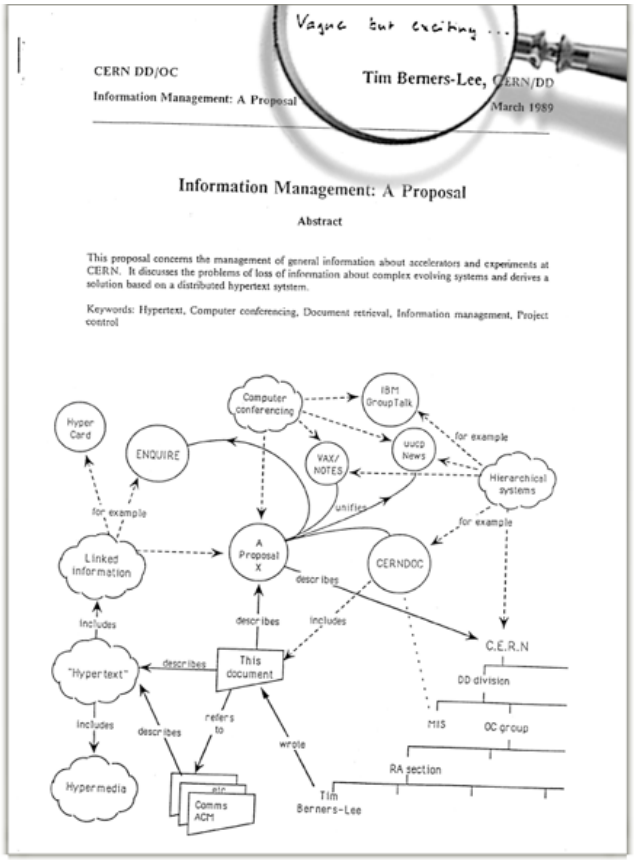
Repositories, Systems and Careers

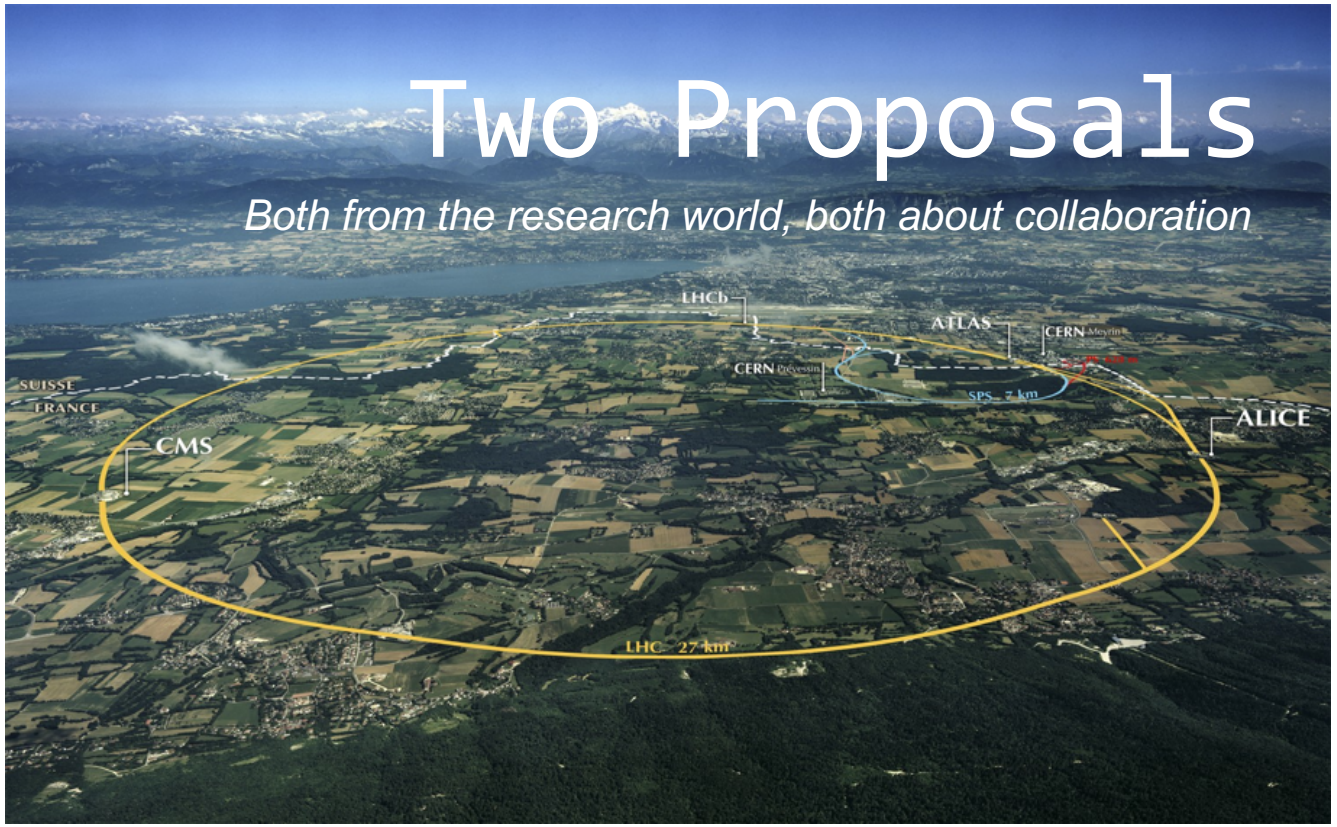
RDMF16: Data (and) Systems, Redux
Digital Curation Centre, University of Edinburgh

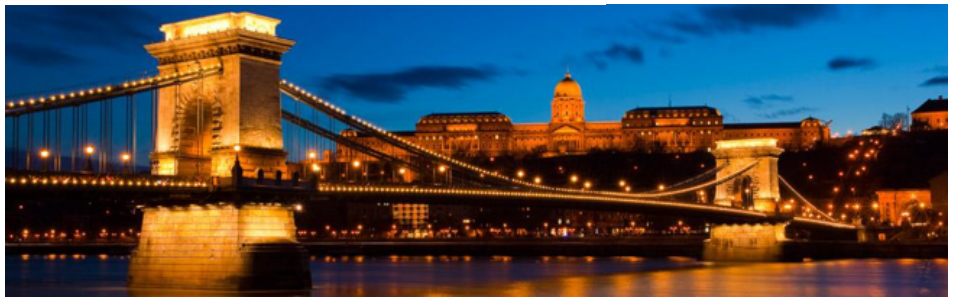
Professor Leslie Carr, University of Southampton

Leslie Carr

- Professor of Web Science, Department of Electronics and Computer Science, University of Southampton
  - Director of EPrints Services & onetime Repository Manager
  - Director of Web Science Institute
- The one is the reason for the other
  - Open Access as a technical possibility *vs* cultural, political and economic realities





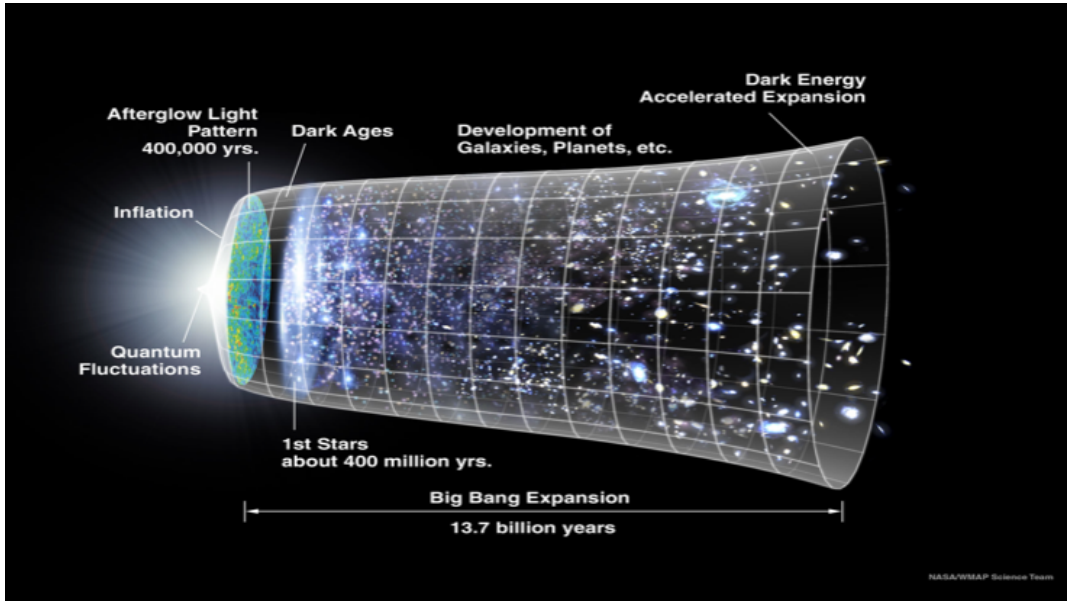


An old tradition and a new technology have converged to make possible an unprecedented public good. (BOAI 2001)

Expansion of the Web

- The Web spread the conditions of its initial creation throughout the whole of society as it underwent an initial inflationary phase.
- The academy
  - government patronage
  - large-scale co-operation
  - sharing of intellectual property

A physics-based CERN metaphor



Society is Diverse. One Web Fits All?

Institution	Objective
Academy	Create and transmit knowledge
Commerce	Make and trade goods
Press	Investigate and report news
Media	Create and broadcast content
Police	Maintain order and public surveillance
Judiciary	Apply law and resolve disputes
Government	Control society and share resources

The development of society as a whole (nuanced and structured and refined) is inextricably related to the technology of information provision, consumption and dissemination (e.g. writing, reading, printing, education). *Different parts of society have different objectives and hence incompatible Web requirements, e.g. openness, security, transparency, privacy.*

Historic Attempts at Making a Web

Sponsor	System	Scope	Real	Date	Important Properties
Finance / Press	**Reuters**	Professional, centralised	✔	1850	News & stock information (originally carrier pigeon and subsequently telegraph)
Private Institution	**Mundaneum**	Public, centralised	✔	1920	Based on indexing technology (the library card)
Military	**Memex**	Scholarly, individual, centralised	✗	1945	Aimed at Scientists and Technologists in WWII
Media	**Xanadu**	Public, decentralised	✗	1960	Focused on DRM, reuse and writing for “creatives”
Media	**CEEFAX**	Public, national, centralised	✔	1970	Broadcast, linked, not participatory
Government	**Minitel**	Public, national, centralised	✔	1980	Commercial services and information
Academy (CS & HEP)	**FTP / Archie / Anarchie**	Public, decentralised	✔	1985	Download resources (papers, reports) to hard drives and print them on LaserWriters.
Commerce	**Hypercard, HyperTIES**	Private, centralised	✔	1988	Personal applications, sometimes tied to multimedia resources on CDROMs / video disks
Academy (HEP)	**WWW**	Public, global, decentralised	✔	1990	Universal naming, linking, interoperability, participative. However no writing, no indexing in public version.
Academy (CS)	**Microcosm**	Private, centralised	✔	1990	Sophisticated linking and openness for personal information stores
Academy (CS)	**HyperG**	Public, centralised	✔	1990	Extension of Web for with support for writing, indexing and consistency management.
Commerce	**AOL, CompuServ**	Public, centralised	✔	1990	Dialup walled garden access to email, forums, chat rooms and information resources

RDMF16 Big Questions

competitive tool ecosystem

functional tool ecosystem

global tool ecosystem

How will data repositories evolve and compete with CRIS, file sharing and related platforms?

What tools, strategies and workflows enable institutions to connect open data, safe sharing, and confidential collaboration – and manage the risks?

How can research data services connect with broader research infrastructures and information, engage with non-academic data sources, and promote reuse beyond the research audience?

- Developed in research lab, serves library
  - Web and Internet Science, University of Southampton
- Universities and researchers
  - knowledge producers
  - knowledge consumers
- The Web has radically altered the potential for knowledge dissemination in society
- We want to **understand** and **facilitate** that change
  - Research *and* development

EPrints Software Releases

- Announcement: Oct 1999
- Version 1: June 2000
- Version 2: Feb 2002
- Version 3.0: Jan 2007
- Version 3.3: June 2011
- Version 3.4: Oct 2016

EPrints for Research Data

Building on the long running success of EPrints, historically used for Open Access publications

- Supplies researchers with a sharable, citable resource.
- DataCite DOI minting support
- Search and browsing facilities refined for the research data.
- Highly configurable and customisable workflows, disciplinary-specific tailoring
- Functionality extensible from EPrints Bazaar (one-click install community app store)
- Customisable look and feel
- Integration with large-scale back end storage solutions *e.g.* Arkivum's storage appliances
- Cross-linking with other Institutional EPrints repositories: publications linked to research data.
- Familiar EPrints integration and aggregation points: OAI-PMH, Dublin Core metadata, numerous metadata import/export formats.
- Data ingest options, such as Sword deposit and EPrints' extensible plugin architecture.

EPrints 3.4 from EPrints Services

Refactored and simplified configuration for repository innovation

Publication domain specific features have been separated out, making a pure RD repository cleaner and more maintainable.

Improved record summary page rendering options.

Support for more advanced search and browsing features such as guided/faceted searching.

Added support for alternative user interfaces.

EPrints 3.4 Flavours

Purpose: What's it for? Who wants it? Why is it important? Who are the community?

Contents: Metadata – support an existing schema? Application profile? DIY?

Data – what kind of Objects / Documents / Files / Media? Where will the metadata / data come from? Import from another system? User contributed?

Users: Do you need users? Some editors? Just one administrator? Who can be a user? Do they need rich profiles? How are they authenticated? What workflows are needed to interact with the contents – what can users do and how can they do it?

Access to Contents: Human visualisation – what kinds of searches and presentation views for individual items and collections. Data access API export formats? OAI? REST? Policies for access – who gets to see what?

Ecology: What other systems does the repository need to work with?

Next Generation Repositories - Now

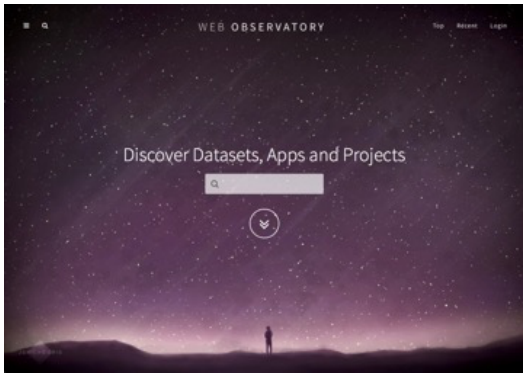
- **We have** a global network of repositories that allows frictionless access to open content and encourages the creation of cross-repository added-value services
- **But** not fully realized their potential and function mainly as passive, siloed recipients of the final versions of their users' conventionally published research outputs
  - few individual repositories are important in and of themselves
  - collectively have the potential to offer a comprehensive view of the research of the whole world
  - while also enabling each scholar and institution to participate in the global network of scientific and scholarly enquiry

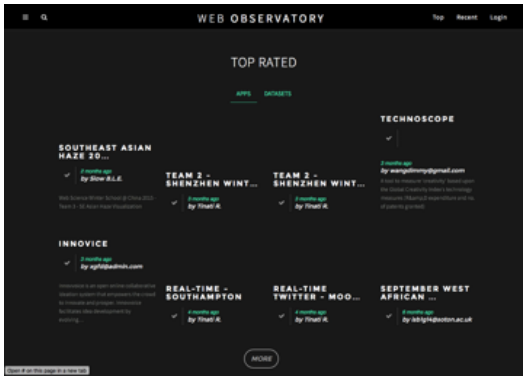
Next Generation Repositories – Soon?

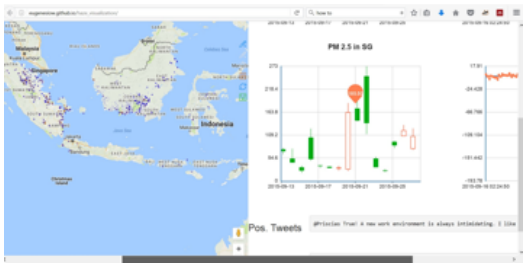
- the distributed network of repositories can and should be a powerful tool to promote the transformation of the scholarly communication ecosystem
- research-centric, innovative, managed by the scholarly community

- **In short**, make the repository network more *of the Web*, facilitating a global community.
- Think Web 2 & Web 3, not Web 1!

http://webobservatory.soton.ac.uk/







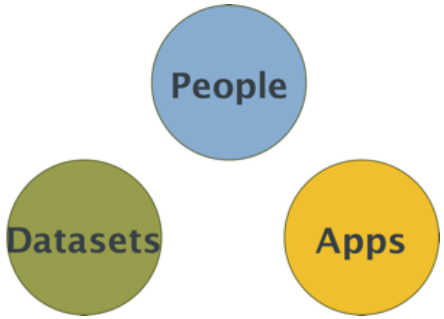
Web Observatory

An initiative supported by the Web Science Trust & W3C
http://www.webscience.org/

A experimental distributed infrastructure using common metadata (schema.org) for listed (possibly hosted) datasets and apps
http://index.webobservatory.org/

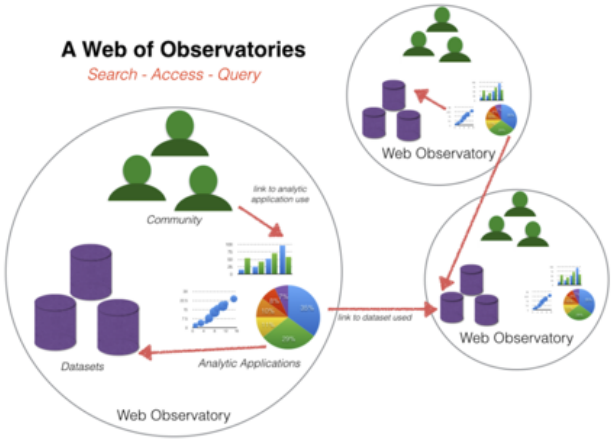
Some WO sites use purpose-built software that:

- Allows their community members to list and share public or private datasets and apps
- Provides for discovery and access to listed datasets and apps across WO sites (OpenIDConnect)
- Provides APIs for app development using listed datasets



Web Observatory

No Datasets = App Store
No Apps = Data Repository
No People = Personal Observatory



Not all datasets or applications need be public

Web Observatories list two main types of resources:

datasets and analytic applications, including visualisations.

Not all listed resources need to be locally hosted

Metadata describing the listed resources and projects are published.

Web Observatory

- Technical infrastructure is a space to negotiate
  - trusted access, legal and ethical responsibilities (data protection, privacy, anonymysation)
- Distributed ethics is hard!
  - In a small analysis of 145 different attributes identified from 10 university ethics forms (UK, US, EU and Asia), only the name of the PI, and whether informed consent was sought were common to all forms (Hutton and Henderson 2015)

Social Media Data for Research

Speaking of ethically and legally problematic data…




- From social practices and effects … to data
- Social data is no longer generated and owned by social scientists

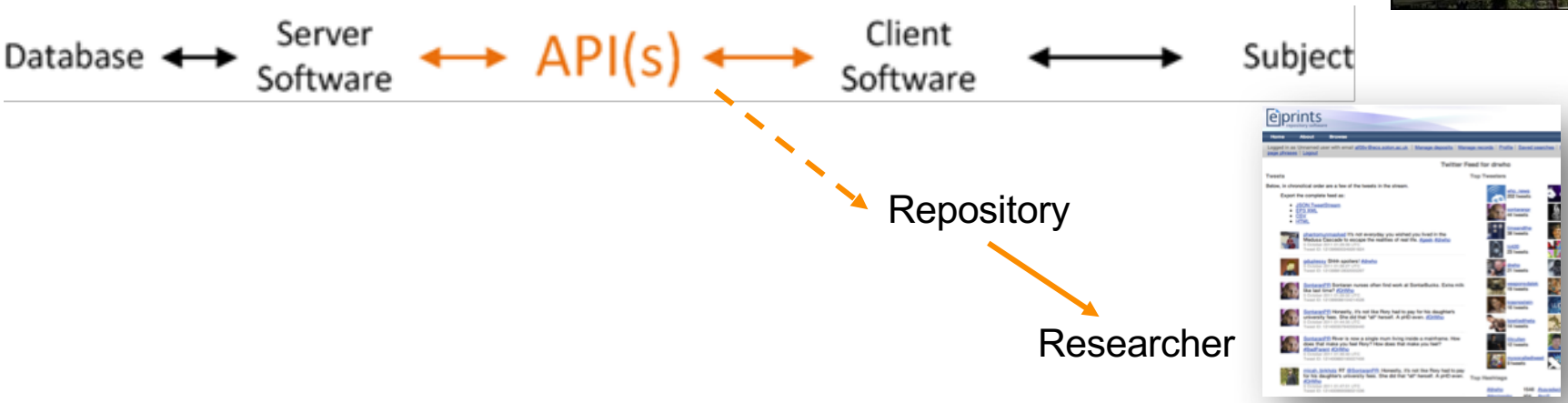
also transaction data, IoT sensor data, health data, behavioural data

- Enthusiasm:

  '… as if the inner workings of private worlds have been pried open' (Latour 2007)
- Scepticism:

  'whatever value big data may have for "knowing capitalism", its value to social science has … for the present at least, to remain very much open to question' (Goldthorpe 2016)

The Social Media Data Pipeline



- Owned and controlled by a commercial company
- Invisible and unknowable and variable shaping of the data
  - Rate limiting, data culling, realtime vs historic
- Researchers worry about populations, sampling, methods
- What are we “archiving” and why (*e.g.* Twitter at the LoC)

Repositories are more than Software

- The strength of the repository infrastructure is not that of a technical network of repository systems
- It is the knowledge, experience and activities of a network of skilled professionals using that network
- Librarians, repository managers, open access advocates
- At their best, repositories become *trusted stewards* of researchers' intellectual property

The Software Sustainability Institute

www.software.ac.uk

A national facility for building better software

- Better software enables better research
- Software reaches boundaries in its development cycle that prevent improvement, growth and adoption
- Providing the expertise and services needed to negotiate to the next stage

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

EPSRC Pioneering research

BBSRC

THE UNIVERSITY OF EDINBURGH

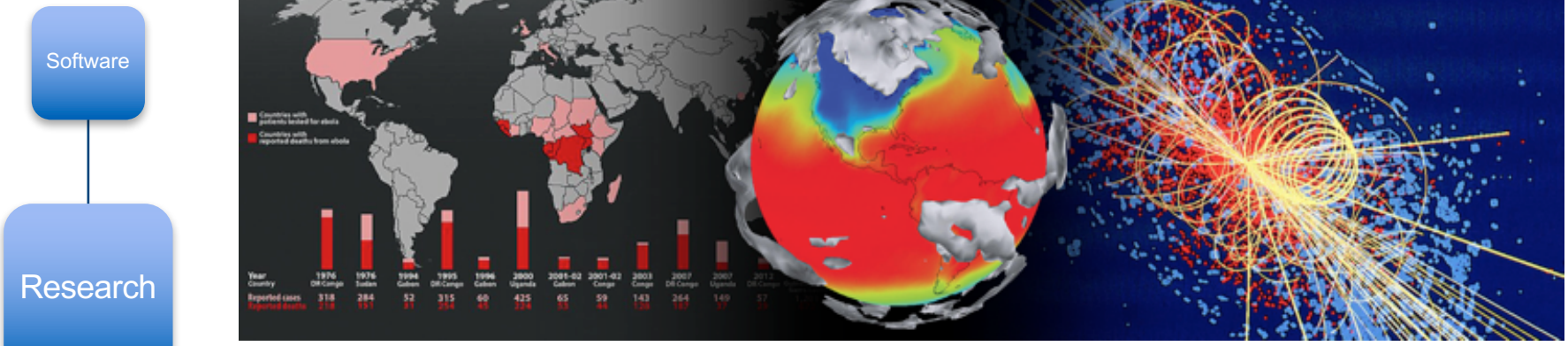
MANCHESTER 1824

UNIVERSITY OF Southampton

UNIVERSITY OF OXFORD

info@software.ac.uk

SSI: Research is impossible without software



From thrown-together scripts, through an abundance of complex spreadsheets, to the millions of lines of code behind large-scale infrastructure, there are few areas where software does not play a fundamental part in research, partly because so much of life is mediated online.

SSI Goals

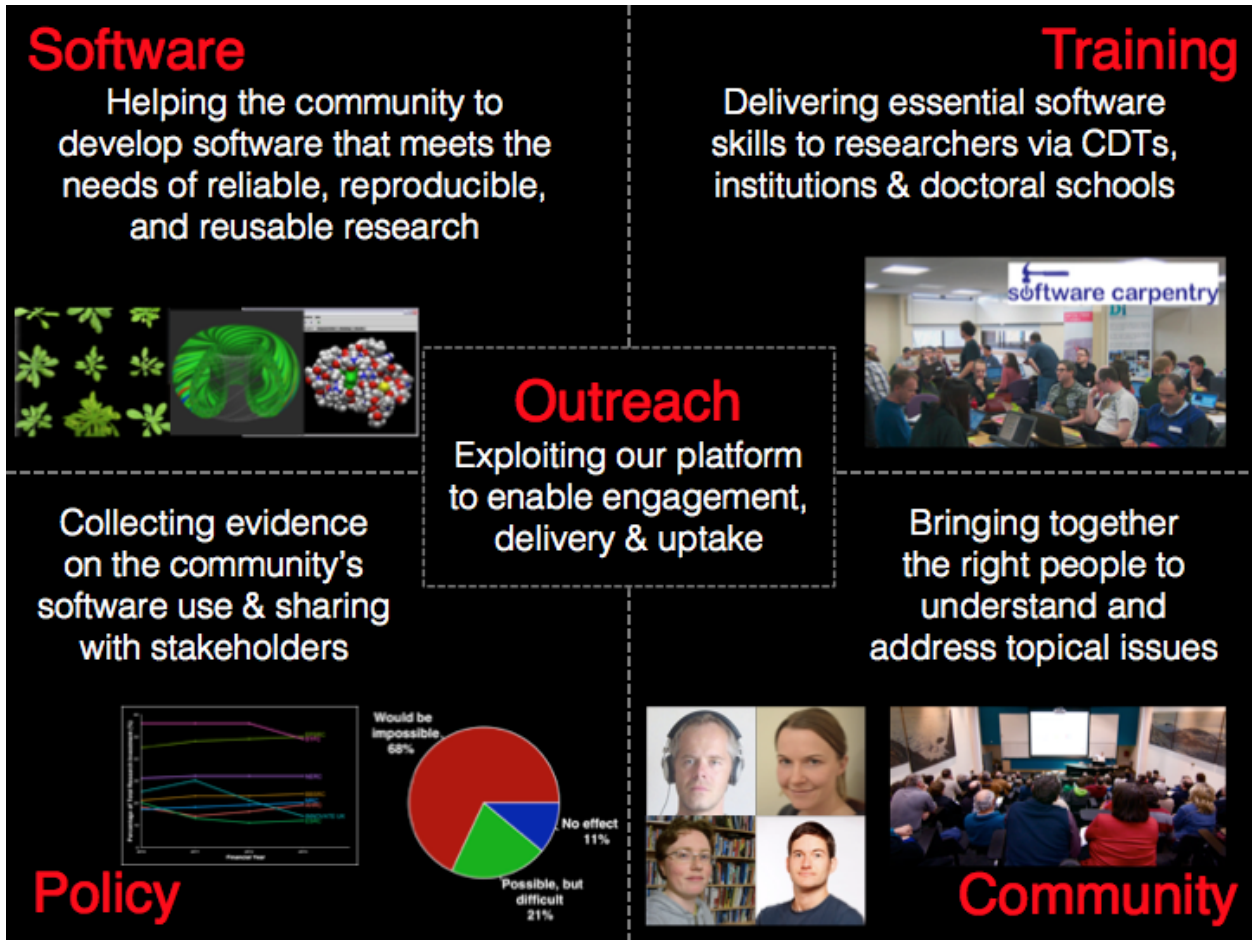
A skilled researcher base in the UK

Recognition of the importance of software to research

Professionalisation of the research software area

Increased scientific integrity

Protection of investment



Organisation	Group/Fellow
University of Bristol	Advanced Computing Research Centre
University of Cambridge	High Performance Computing Service
University College London	Research Software Development Group
Culham Centre for Fusion Energy	Data and Coding team (CODAS & IT)
University of Edinburgh	Edinburgh Parallel Computing Centre
Francis Crick Institute	Application Integration and Migration
Imperial College London	Research Software Engineering Community
ISIS	Mantid Group
University of Manchester	Research Software Engineering Group
University of Oxford	Advanced research Computing
University of Oxford	Research Software Developers Network
University of Sheffield	Research Software Engineering at Sheffield
University of Southampton	Research Software Group
STFC	Software Engineering Group

Combine expertise in programming with an intricate understanding of research.

Lack a formal place in the academic system - no easy way to recognise their contribution, to reward them, or to represent their views.

Working to raise awareness of the role and bring the community together.

Some start off as researchers who spend time developing software to progress their research.

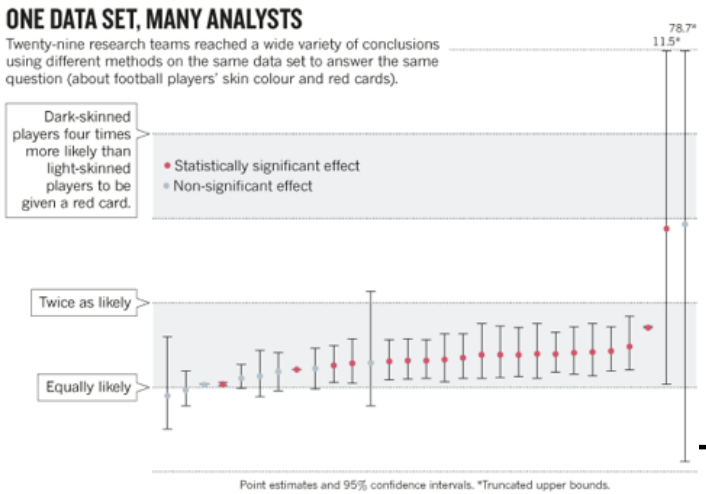
Others start off from a software-development background and are drawn to research.

Without them, research software will fail to meet the demands of researchers.

Challenge of Research Software

> “It's not just about data. Storing data is one step towards reproducibility. If the software used to interpret the data is lost, then what's the point in having the data? This, of course, raises all the extra questions about related to software, how do you store it, how do you associate it with the relevant data (persistent identifiers and citation) and how do you provide the skills needed to deal with software properly (training, careers for RSEs).

Challenge for/of Data Science



Nature 526, 189–191 (08 October 2015)
doi:10.1038/526189a

Hypothesis, same data, multiple (29) research teams' analysis

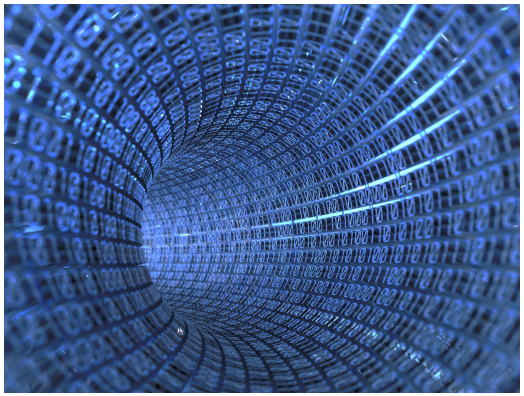
Each team's results strongly influenced by subjective choices.

Conclusions ranged from no race bias in referee decisions to a huge bias.

Where the Web Went Wrong re Data

- W3C defines the Web – URLs HTTP HTML – to be an abstract information space

Knowledge → Information → Data

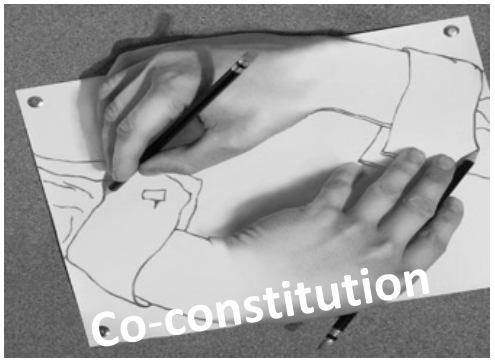


- These are not neutral / abstract things
- Human to machine, value to commodity.
- Political economics of the web
  - companies stealing our data *vs* value exchange

The Web: An Opportunity

- The Web isn't a thing but an activity
  - the creation of a network of information by a network of individuals.
- The Web wasn't invented by Tim Berners-Lee,
  - it is still being invented by all of us as we gradually adapt our tools and change our practices

The Web is an extraordinary change in the ability of humankind to build value and to be valuable
The product of people with open license, open standards, open systems



Missing Question

"Research is the by-product of researchers getting promoted"
David Barron, Professor of Computer Science

How do we make researchers careers more successful?

- Corollary: why is telling researchers what they ought to be doing such hard work?

Concluding Thoughts

- NGM agenda
  - acknowledges that repositories are stuck in old-web thinking
  - that the world and the Web have moved on
  - it is necessary to contextualise the technology in a global network of knowledge producers

- When we consider data management technologies, we need to think
  - vital material of communities of research expertise
  - support for career-long identities with valuable and active back catalogues

Others' Thoughts

- Susan H – access and privacy, J drive, invisible infrastructure good/bad, qualitative vs quantative

Librarians Transforming Practice

Not just enforcing historic norms, but stimulating new practice to emerge

- Copyright
- Openness
- Intellectual Property
- Privacy
- Creativity
- Science 2.0