

# Enabling Scientific Data Sharing and Re-use

R. Darby, S. Lambert, B. Matthews, **Michael Wilson**

Science and Technology Facilities Council, Didcot, UK

K. Gitmans

Alfred Wegener Institute for Polar and Marine Research, Germany

S. Dallmeier-Tiessen, S. Mele

CERN, Geneva, Switzerland

J. Suhonen

CSC - IT Center for Science, Espoo, Finland



*Opportunities for Data Exchange*

Scientific Computing  
Department



Science & Technology  
Facilities Council

**“We must give taxpayers more bang for their buck. Open access to ... data is an important means of achieving this.”**



Máire Geoghegan-Quinn,  
European commissioner for research,  
innovation and science  
July 2012.



*Opportunities for Data Exchange*

Scientific Computing  
Department



Science & Technology  
Facilities Council

# The study

- 1) collect 21 examples of data sharing
- 2) create baseline model of sharing;
- 3) test model completeness in workshop;
- 4) identify drivers, barriers and enablers
- 5) prioritise through interviews with 55 experts;
- 6) derive recommendations for stakeholders.



*Opportunities for Data Exchange*

Scientific Computing  
Department



Science & Technology  
Facilities Council

# Stakeholders

- Policy-makers: National, Regional
- Funders: Research, Infrastructure
- Researchers: Data producers, Data consumers
- Research and education organisations
- Data management and infrastructure service providers (librarians, publishers)



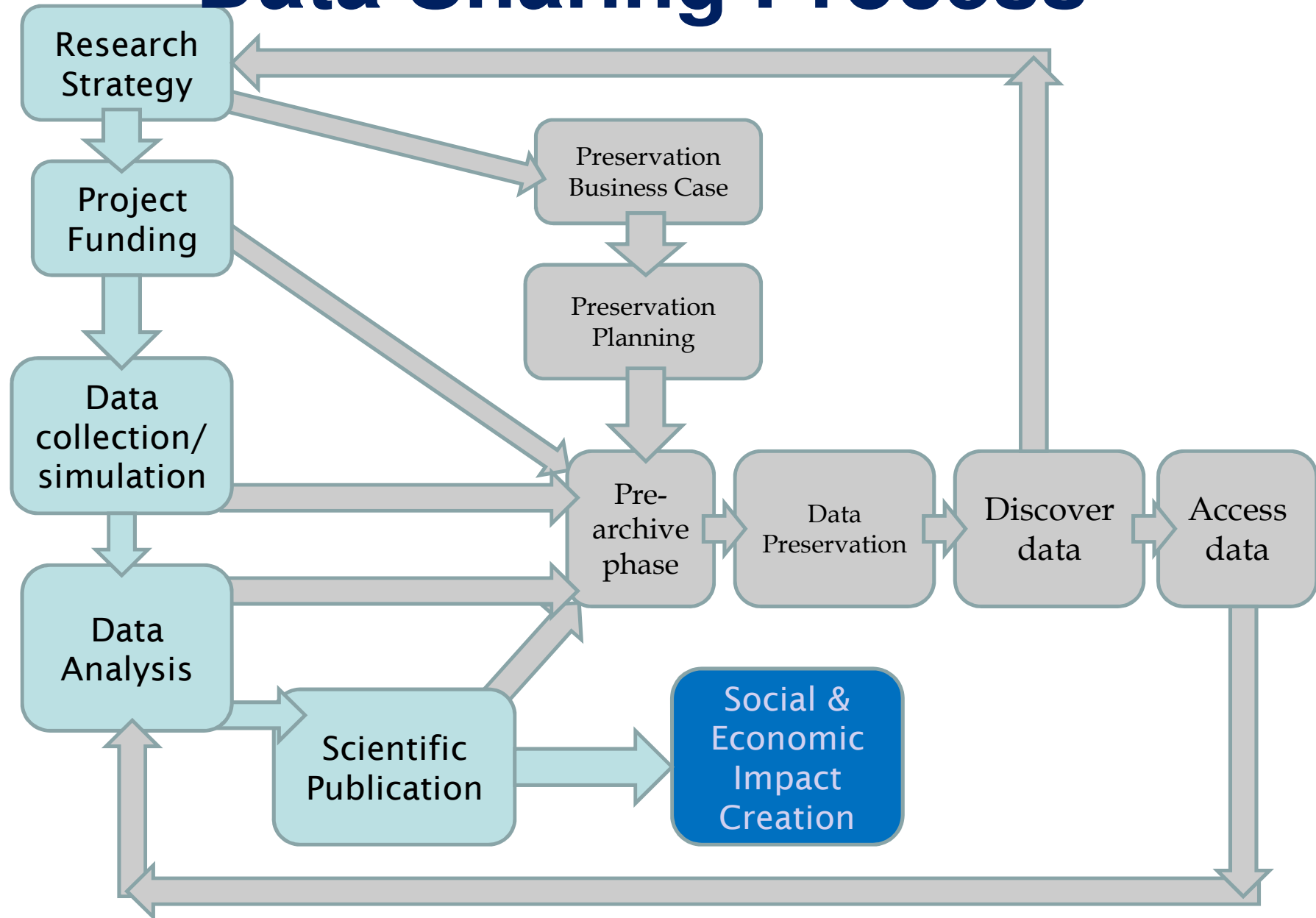
*Opportunities for Data Exchange*

Scientific Computing  
Department



Science & Technology  
Facilities Council

# Data Sharing Process



# Context Factors

- Academic discipline:
  - Source of data, Cost of data collection, Possibility to collect data again, Complexity of data analysis
- Country:
  - Legislation, Infrastructure, Funding
- Age of researcher:
  - Willingness to invest effort for possible long-term benefit
- Data Re-use Sector :
  - Non-commercial research, Commercial research, Education



*Opportunities for Data Exchange*

Scientific Computing  
Department



Science & Technology  
Facilities Council

# Drivers

- a) Societal benefits
- b) Academic Benefits
- c) Research Benefits
- d) Organisational Incentives
- e) Individual Contributor Incentives



*Opportunities for Data Exchange*

Scientific Computing  
Department



Science & Technology  
Facilities Council

# Barriers

- f) Individual Contributor Incentives
- g) Availability of a Sustainable Preservation Infrastructure
- h) Trustworthiness of the data, Data Usability, Pre-archive activities
- i) Data Discovery
- j) Academic Defensiveness
- k) Finance
- l) Subject Anonymity and Personal Data Confidentiality
- m) Legislation/Regulation



*Opportunities for Data Exchange*

Scientific Computing  
Department



Science & Technology  
Facilities Council



# Enablers

- n) Individual Contributor barriers
- o) Availability of a Sustainable Preservation Infrastructure
- p) Trustworthiness of the data, Data usability, Pre-archive activities
- q) Data Discovery
- r) Academic Defensiveness
- s) Finance
- t) Subject Anonymity and Personal Data Confidentiality
- u) Legislation/Regulation



*Opportunities for Data Exchange*

Scientific Computing  
Department



Science & Technology  
Facilities Council

# ***Data Publication Practice Themes***

- *The role of publishers in data sharing*
- *Data citation and description for discovery and use*
  - Granularity of data to be cited



*Opportunities for Data Exchange*

Scientific Computing  
Department



Science & Technology  
Facilities Council

**nature** International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive | Volume 467 | Issue 7319 | Articles | Article

NATURE | ARTICLE **OPEN** previous article next article >

## A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium

Affiliations | Contributions | Corresponding author

Nature 467, 1061–1073 (28 October 2010) | doi:10.1038/nature09534

Received 29 July 2010 | Accepted 30 September 2010 | Published online 27 October 2010

Corrected online 25 May 2011

Corrigendum (May, 2011)

**Abstract**

**Abstract** • Introduction • Data generation, alignment and variant discovery • Power to detect variants • Genotype accuracy • Putative functional variants • Application to association studies • Mutation, recombination and natural selection • Discussion • Methods • Change history • References • Acknowledgements • Author information • Supplementary information • Comments

The 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. Here we present results of the pilot phase of the project, designed to develop and compare different strategies for genome-wide sequencing with high-throughput platforms. We undertook three projects: low-coverage whole-genome sequencing of 179 individuals from four populations; high-coverage sequencing of two mother–father–child trios; and exon-targeted sequencing of 697 individuals from seven populations. We describe the location, allele frequency and local haplotype structure of approximately 15 million single nucleotide polymorphisms, 1 million short insertions and deletions, and 20,000 structural variants, most of which were previously undescribed. We show that, because we have catalogued the vast majority of common variation, over 95% of the currently accessible variants found in any individual are present in this data set. On average, each person is found to carry approximately 250 to 300 loss-of-function variants in annotated genes and 50 to 100 variants previously implicated in inherited disorders. We demonstrate how these results can be used to inform association and functional studies. From the two trios, we directly estimate the rate of *de novo* germline base substitution mutations to be approximately  $10^{-8}$  per base pair per generation. We explore the data with regard to signatures of natural selection, and identify a marked reduction of genetic variation in the neighbourhood of genes, due to selection at linked sites. These methods and public data will support the next phase of human genetic research.

**Subject terms:** Genetics and genomics

**Figures at a glance**

**Introduction**

**Abstract** • Introduction • Data generation, alignment and variant discovery • Power to detect variants • Genotype accuracy • Putative functional variants • Application to association studies • Mutation, recombination and natural selection • Discussion • Methods • Change history • References • Acknowledgements • Author information • Supplementary information • Comments

Understanding the relationship between genotype and phenotype is one of the central goals in biology and medicine. The reference human genome sequence<sup>1</sup> provides a foundation for the study of human genetics, but systematic investigation of human variation requires full knowledge of DNA

Journal home | Subscribe | Current issue | E-alert sign up | For authors | RSS feed

**Spotlight On Latin America** EXPLORE NOW

Contents to this article  
Crossref (215) | Scopus (315) | Web of Science (263)

**Selected feature**

**After the ice**  
A special issue of Nature explores the role of science in the changing Arctic.  
See complete feature >

**Editor's summary**

1000 Genomes Project pilots published  
This issue of Nature contains the first publication from The 1000 Genomes Project, an international collaboration that will produce an extensive genetic catalogue of human genetic variation. The plan...

**News & Views**  
by Nielsen  
The 1000 Genomes Project has completed its pilot phase, sequencing the whole genomes of 179 individuals and characterizing all the protein-coding sequences of many others. Welcome to the third phase 0...  
Continue >

**Science jobs from natureJobs**

**Post-Doctoral Scientist**  
University of Michigan

**Scientist Bioinformatics**  
Polyclone Bioservices Pvt. Ltd

**Research Fellowships in Cancer and Diseases of Immune Regulation**  
University of Queensland

Post a free job > | More science jobs >

Open Innovation challenges

new publishing models

author information  
doi

live updates

article-level metrics

Tool box to print,  
download reference,  
share: email, social media,  
bookmark

Related content

Figure previewer

Collapsible sections

Scientific Computing  
Department



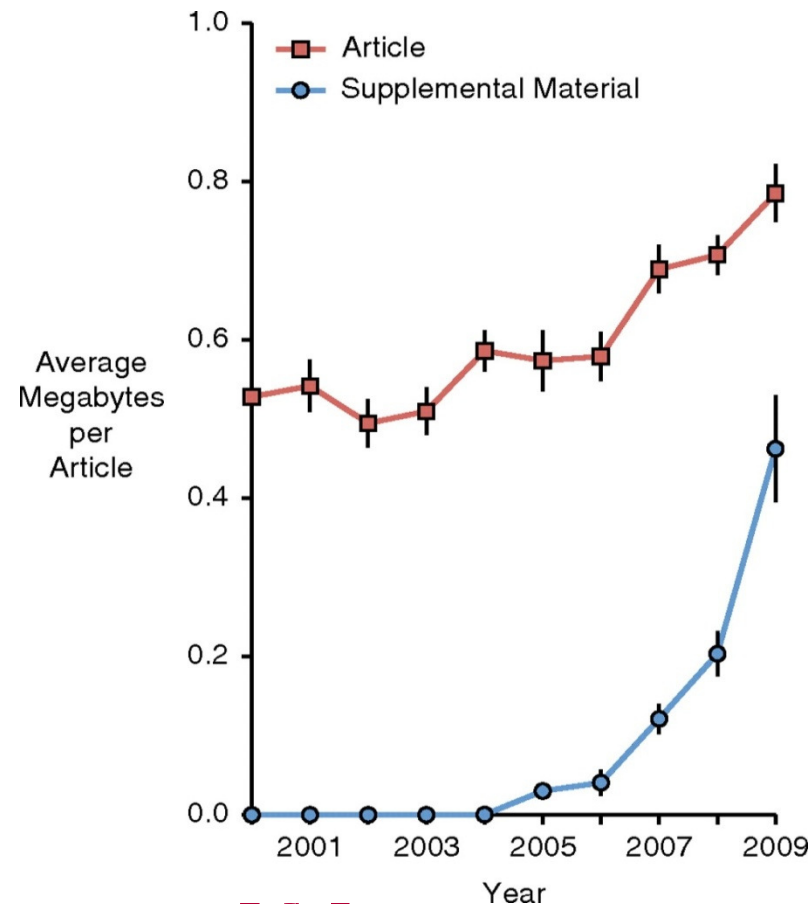
Science & Technology  
Facilities Council

Opportunities for Data Exchange

Source: V. Kiermer, Nature Publishing Group, 2011

# How big is the Data Problem for journals?

## Too big for the Jnl of Neuroscience and Cell:



Maunsell J J. *Neurosci.* 2010;30:10599-10600  
©2010 by Society for Neuroscience

*Opportunities for Data Exchange*

### Jnl of NeuroScience:

The Graph depicts the average size of a Journal of Neuroscience article and supplemental material in megabytes.

As a consequence, the Journal no longer accepts supplementary files to manuscripts, soon the supplementary material would outgrow the article volume. The burden on the peer review process became simply too large.

### Journal Cell:

Editors suspect researchers to treat supplements as data dumping grounds (Emily Markus, Cell)

### General:

Publishers cannot guarantee proper preservation and future accessibility of supp files.

# ***Data Management Infrastructure Themes***

- *Finance: funding infrastructure and data services;*
- *Quality assurance of research data*
- *Standards and interoperability*



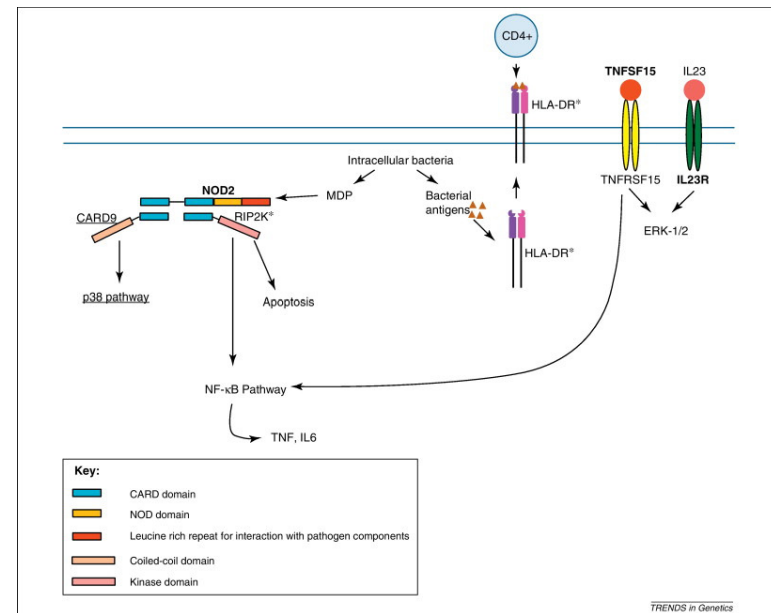
*Opportunities for Data Exchange*

Scientific Computing  
Department



Science & Technology  
Facilities Council

# Interdisciplinary data sharing standards: common semantics



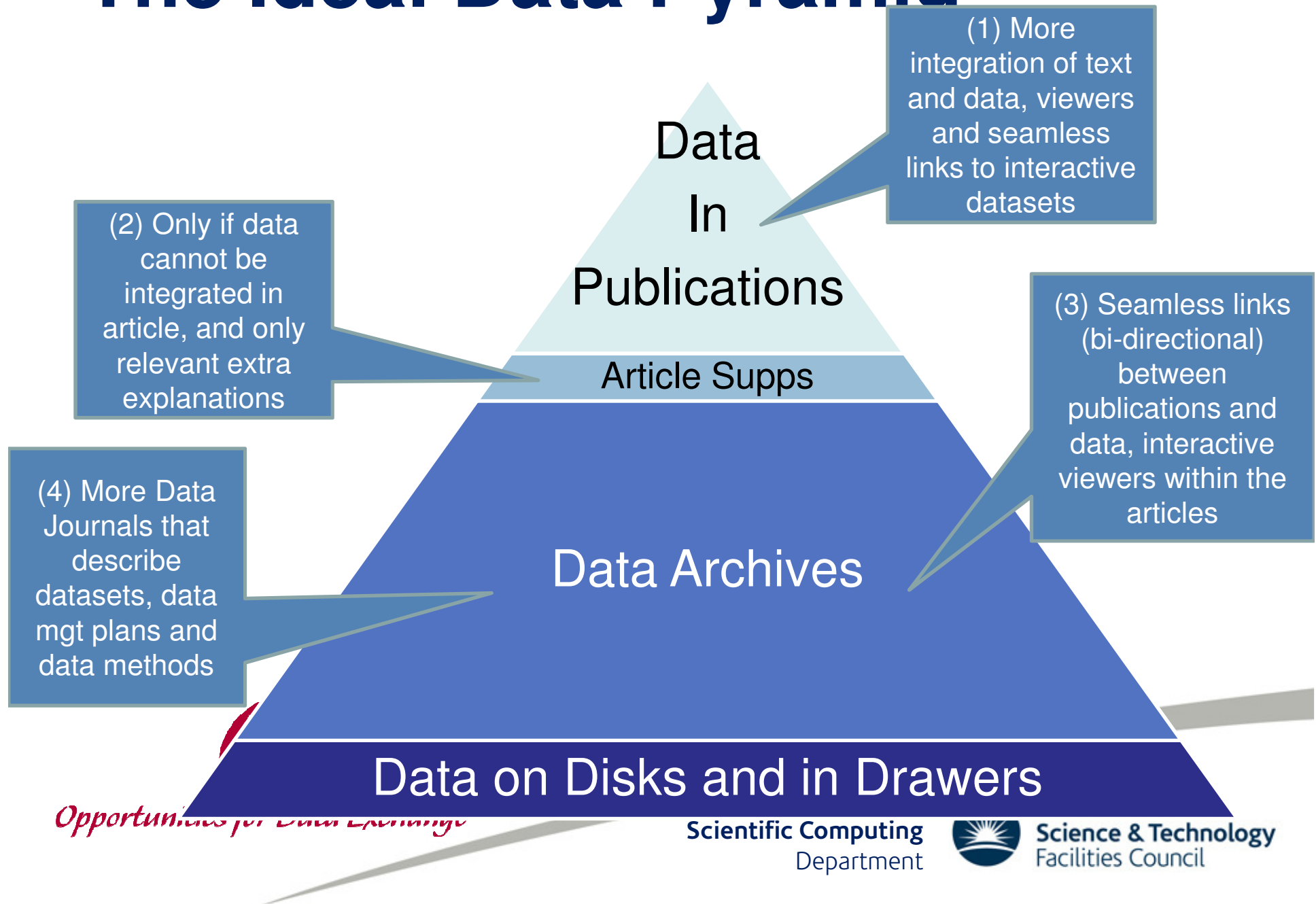
*Opportunities for Data Exchange*

Scientific Computing  
Department



Science & Technology  
Facilities Council

# The Ideal Data Pyramid





# ***Culture and Policy Themes***

- *Data sharing culture*
- *Public visibility of research data*
- *National and regional policy and legal frameworks*
- *Incentives in the academic reward system for good data practice*



*Opportunities for Data Exchange*

Scientific Computing  
Department



Science & Technology  
Facilities Council



# *Incentives in the academic reward system*

- “data sharing could work against individual scientists' need for recognition”- Gerrit Hirschfeld· Nature  
Volume: 487, Page: 302 Date published: (19 July 2012) DOI: doi:10.1038/487302c
- a common system of credit and recognition for data production and sharing is needed
- provide researchers with clear instructions on how to cite data
- Include data publication & citation metrics in researcher appraisal

# Questions ?



## Opportunities for Data Exchange

Scientific Computing  
Department



**Science & Technology**  
Facilities Council