Why funders, policy-makers, and researchers should care about data access and re-use policies

Simon Hodson

Executive Director CODATA

www.codata.org/blog

execdir@codata.org

@simonhodson99

What is CODATA?

Mission: To strengthen international science for the benefit of society by promoting improved scientific and technical data management and use.

An Interdisciplinary Body of ICSU, the International Council for Science: http://www.icsu.org/

CODATA has been working at the forefront of data science since 1966
http://www.codata.org/about/CODATA@45years.pdf

Not-for-profit, membership organisation (national members, scientific unions, affiliate members, members-at-large).

Genuinely global: national members include US, Canada, China, Japan, South Korea, Taiwan, India, Indonesia, South Africa…

Strong links with international scientific unions: 16 union members.

CODATA is…

An international community and network of expertise on data issues.

An influential and authoritative voice in national and international policy regarding scientific data management.

A focal point for international, cross- disciplinary collaboration and communication on key scientific data issues.

CODATA Strategic Plan

1. **Policy frameworks for data:** take the lead in defining a policy agenda for scientific data.

   - **Data Policy Committee.** Provide authority and support for data policy issues, addressing national, international, disciplinary and legal issues.

2. **Frontiers in data science and technology:** coordinate work in key frontiers of data science and interdisciplinary application areas.

   - **CODATA workshop series on Frontiers of Data Science and Technology:** looking to partner other organisations and build on work of task groups.

   - **Capacity building activities**: curriculum framework for data science, global outreach activities.

   - International **Science Data Conference, with WDS, New Delhi 2-5 Nov 2014.**

   - **Task Groups** http://www.codata.org/taskgroups/index.html ; **Data Science Journal**.

3. **Data strategies for international science:** support major ICSU scientific programmes like Future Earth to address data management needs.
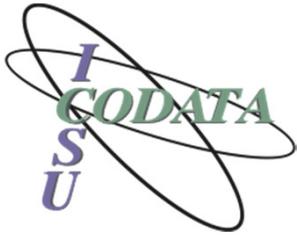
CODATA Data Policy Committee

Initiate and lead forums on open access, research funder policies, legislative frameworks, good practice, and the open data agenda.

Build on OECD principles to promote publicly access to the outputs of publicly funded research to achieve greater return on investments through reuse of research data.

Help address national agendas and promote international collaboration and agreement.

Encourage dialogue around national and disciplinary concerns WRT data policies.

Engage in a process of validation and iteration with National CODATA Committees (and academies, funders, transnational bodies …), with International Scientific Unions (and Learned Societies, publishers, journal editorial boards ….).

Work with stakeholders to formulate and implement policies with traction…

Supports ICSU initiatives: **Future Earth**, ICSU 'Report on Open Access and metrics for evaluation'

Policy
(general management statements)

Standards
(specific mandatory controls)

Guidelines
(recommendations / best practices)

Procedures
(step by step instructions)
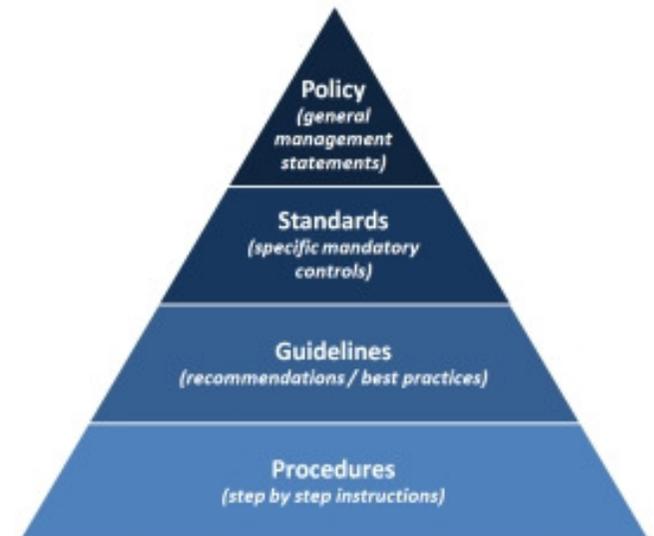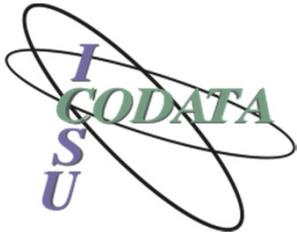
CODATA National Committees

Promote and enhance activities of CODATA National Committees.

Updating model of national committees – includes various stakeholder representatives.

The time is ripe for a data forum to build dialogue between funder driven policy, data experts and the demands of particular research disciplines.

Opportunity for National Committees to take leading role in addressing data issues – relationships with funders, with National Academies, learned societies, key stakeholders. **Aim for National Committees to be a valued forum.**

Build dialogue and exchanges between national committees and with international CODATA community.

What issues and debates are being raised at national levels? To what extent can various stakeholders be involved (funders, research organisations, data experts, data centres, research libraries etc)?

Dialogue between funders, infrastructure providers, institutions, publishers, researchers…
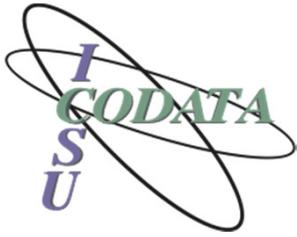
On Data Policies...

What is a data policy, what is it for?

Principles, and policies, and norms, and community agreements... Important to think about how communities reach agreements, relationship between practitioners and policy makers...

Should not be (just) top-down! Policies **should not** just be things that are imposed, 'done to people'.

To **protect** those who 'do the right thing' even if this may be costly or time consuming (analogous to the 'prisoners' dilemma') – **because benefits don't necessarily redound to the person making the 'extra' effort.**

Or a **contract**: this is what is expected in return for research funding...

'Compliance' is not an end in itself: data policies **should** aim to improve data quality, availability and improve research.

Objective to develop a culture of sharing/openness. Essential to have a combination of top down and bottom up.

A shared statement: 'this is how we conduct research'!

Policies are necessary, but not sufficient...
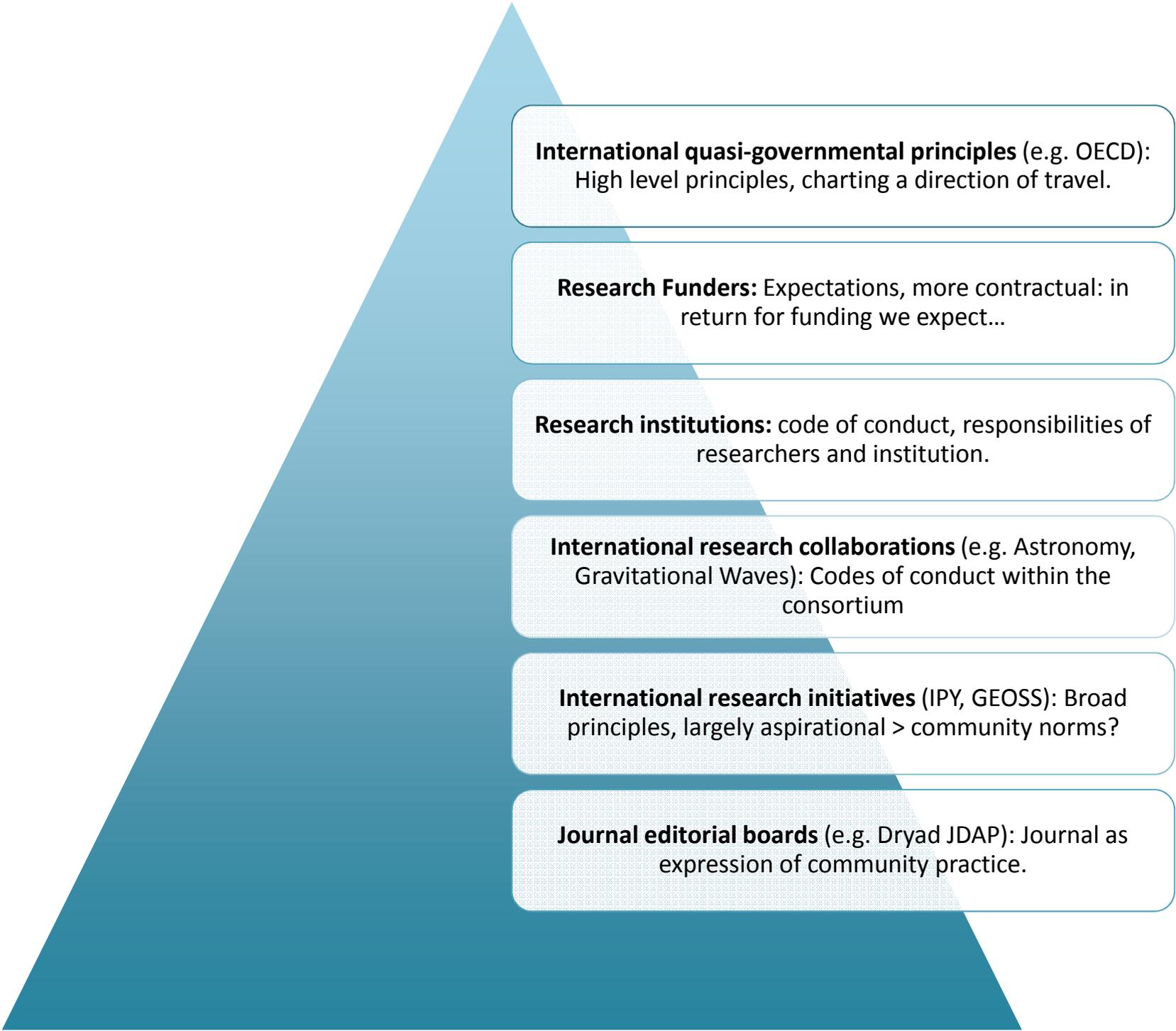
International quasi-governmental principles (e.g. OECD): High level principles, charting a direction of travel.

Research Funders: Expectations, more contractual: in return for funding we expect…

Research institutions: code of conduct, responsibilities of researchers and institution.

International research collaborations (e.g. Astronomy, Gravitational Waves): Codes of conduct within the consortium

International research initiatives (IPY, GEOSS): Broad principles, largely aspirational > community norms?

Journal editorial boards (e.g. Dryad JDAP): Journal as expression of community practice.
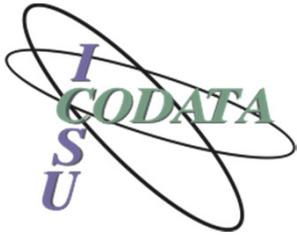
Research Data Policy Directions

Research funder policies, legislative frameworks, good practice, open data agenda

    The outputs of publicly funded research should be publicly available.

    The evidence underpinning research findings should be available for validation

    Greater return on investment through reuse of research data

Good data management is good for research

    More efficient research process, avoidance of data loss, research benefits of data reuse

Aligns with **university mission** to provide excellent research infrastructure.

Pressure on funders and universities to have better **oversight of research outputs and impacts**.

Some moves from **journals** and learned societies towards policies for availability of underlying research data.
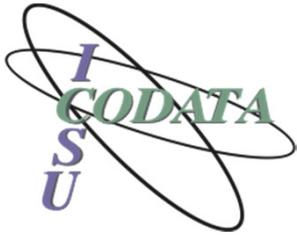
Data Policy Principles: OECD

Builds on 2004 **OECD *Declaration on Access to Research Data from Public Funding***
http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157&Lang=en&Book=False

2007 **OECD *Principles and Guidelines for Access to Research Data from Public Funding***:
http://www.oecd.org/sti/sci-tech/38500813.pdf

Data outputs of publicly funded research should be openly available with as little restriction and cost as necessary.

Asserts principles of Openness, Flexibility, Transparency, Legal Conformity, Protection of Intellectual Property, Formal Responsibility, Professionalism, Interoperability, Quality, Security, Efficiency, Accountability, Sustainability.
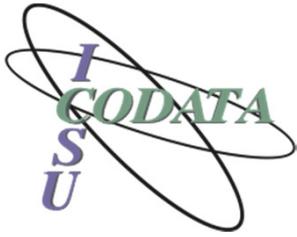
International Dimensions to Data Policy Directions

US Office of Science and Technology Policy Memorandum 'Expanding Public Access to the Results of Federally Funded Research': http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

> Funders required 'to develop a plan to support increased public access to the results of research', including data. Draft plans were due 22 Aug 2013.

EC Recommendation on Access and Preservation of Scientific Information:
http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf

G8 Science Ministers Statement:
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206801/G8_Science_Meeting_Statement_12_June_2013.pdf

- Addressing global challenges requires coordination, open data;

- Need to expand access to scientific research results.

- This requires Global Research Infrastructure; principles of Open Scientific Research Data.
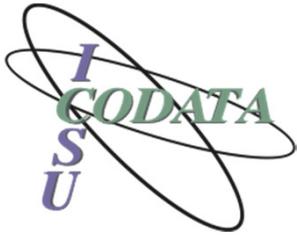
RCUK: Research Funder Principles, 2011

RCUK Common Principles on Data Policy:
http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx

Public good: Publicly funded research data are produced in the public interest should be made openly available with as few restrictions as possible

Planning for preservation: Institutional and project specific data management policies and plans needed to ensure valued data remains usable

Discovery: Metadata should be available and discoverable; Published results should indicate how to access supporting data

Confidentiality: Research organisation policies and practices to ensure legal, ethical and commercial constraints assessed; research process should not be damaged by inappropriate release

First use: Provision for a period of exclusive use, to enable research teams to publish results

Recognition: Data users should acknowledge data sources and terms & conditions of access

Public funding: Use of public funds for RDM infrastructure is appropriate and must be efficient and cost-effective.
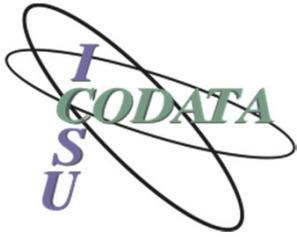
EPSRC Research Data Policy Expectations, 2011

Policy and expectations:

http://www.epsrc.ac.uk/about/standards/researchdata/Pages/policyframework.aspx

Research organisations to have RDM policy, advocacy and support functions. (i, iii)

Research data to be effectively managed and curated throughout the life-cycle (viii)

Research organisations to maintain public catalogue of research data holdings, adequate metadata and permanent identifier (v)

Publications to indicate how research data can be accessed (ii)

Data to be retained for 10 years from last access (vii)

Research data management to be adequately resourced from appropriate funding streams (ix)

Roadmap in place by 1 May 2012

Compliance by 1 May 2015

EPSRC

Pioneering research and skills
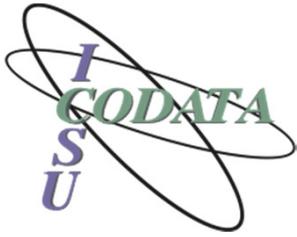
University Data Management Policies

University of Edinburgh led the way in May 2011; input from DCC.

High-level principles, aspirational.

Indicates responsibilities for researchers and the institution.

Widely imitated by other institutions: http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies/uk-institutional-data-policies

Exeter policies for researchers and for post graduates:

http://as.exeter.ac.uk/library/resources/rdm/whymanageyourdata/universityofexeteropenaccessresearchandresearchdatamanagementpolicies/

Policies and Regulations

Research Data Management Policy

This policy for managing research data was approved by the University Court on 16 May, 2011.

The University adopts the following policy on Research Data Management. It is acknowledged that this is an aspirational policy, and that implementation will take some years.

1. Research data will be managed to the highest standards throughout the research data lifecycle as part of the University's commitment to research excellence.

2. Responsibility for research data management through a sound research data management plan during any research project or programme lies primarily with Principal Investigators (PIs).

3. All new research proposals [from date of adoption] must include research data management plans or protocols that explicitly address data capture, management, integrity, confidentiality, retention, sharing and publication.

4. The University will provide training, support, advice and where appropriate guidelines and templates for the research data management and research data management plans.

5. The University will provide mechanisms and services for storage, backup, registration, deposit and retention of research data assets in support of current and future access, during and after completion of research projects.

6. Any data which is retained elsewhere, for example in an international data service or domain repository should be registered with the University.

7. Research data management plans must ensure that research data are available for access and re-use where appropriate and under appropriate safeguards.

8. The legitimate interests of the subjects of research data must be protected.

9. Research data of future historical interest, and all research data that represent records of the University, including data that substantiate research findings, will be offered and assessed for deposit and retention in an appropriate national or international data service or domain repository, or a University repository.

10. Exclusive rights to reuse or publish research data should not be handed over to commercial publishers or agents without retaining the rights to make the data openly available for re-use, unless this is a condition of funding.

Royal Society *Science as an Open Enterprise* Report, 2012

'how the conduct and communication of science needs to adapt to this new era of information technology'.

Intelligent Openness: data should be accessible, assessable, intelligible, usable.

'**As a first step towards this intelligent openness, data that underpin a journal article should be made concurrently available in an accessible database**. We are now on the brink of an achievable aim: for all science literature to be online, for all of the data to be online and for the two to be interoperable.'

A first step...

Royal Society June 2012, Science as an Open Enterprise,
http://royalsociety.org/policy/projects/science-public-enterprise/report/

Science as an open enterprise

June 2012

THE ROYAL SOCIETY
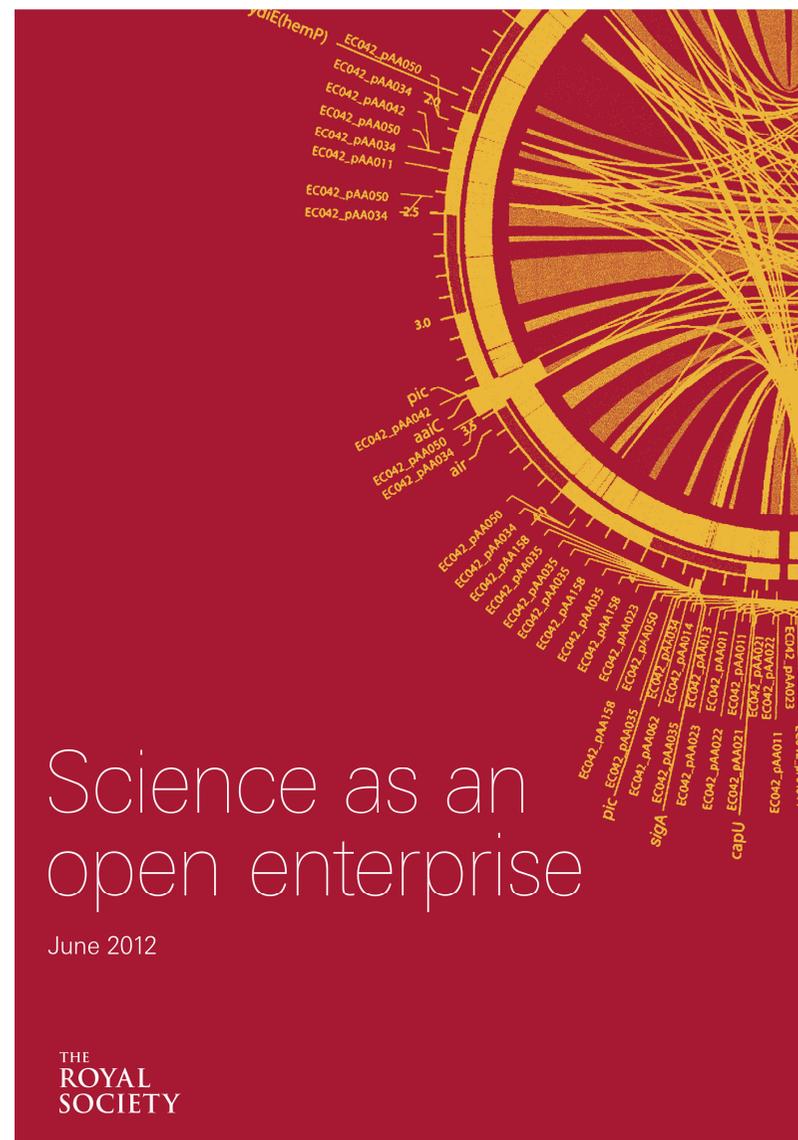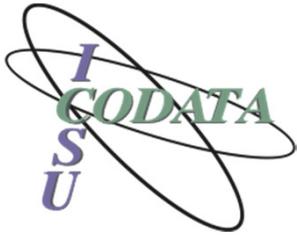
Science as an Open Enterprise
Report: six key changes

1. **a shift away from a research culture where data is viewed as a private preserve;**

2. expanding the criteria used to evaluate research to give **credit for useful data communication** and novel ways of collaborating;

3. the development of **common standards** for communicating data;

4. **mandating intelligent openness for data relevant to published scientific papers;**

5. strengthening the cohort of **data scientists** needed to manage and support the use of digital data (which will also be crucial to the success of private sector data analysis and the government's Open Data strategy);

6. the development and use **of new software tools** to automate and simplify the creation and exploitation of datasets.

Royal Society 2012, Science as an Open Enterprise,
http://royalsociety.org/policy/projects/science-public-enterprise/report/

DU
CONTRACT SOCIAL;
OU,
PRINCIPES
DU
DROIT POLITIQUE.

PAR J. J. ROUSSEAU,
CITOYEN DE GENEVE.

Dicamus leges. — fœderis aquas
Æneid. XI.

A AMSTERDAM,
Chez MARC MICHEL REY.
MDCCLXII.

The credibility and effectiveness of the research enterprise is due in large part to the social contract behind scholarly publishing. Researchers disclose their work to their peers in return for professional credit. In so doing, they also expose their findings to be confirmed or refuted, and enable other researchers to build upon their results. **Dryad seeks to extend this social contract to research data by providing a model for how a disciplinary repository can motivate researchers to disclose the data that is of the greatest value for scientific reuse, that associated with publications**, and realize the manifold benefits of free access to scientific data in perpetuity.

http://datadryad.org/pages/about

Vision 'Open Data and the Social Contract of Scientific Publishing'
http://www.bioone.org/doi/full/10.1525/bio.2010.60.5.2
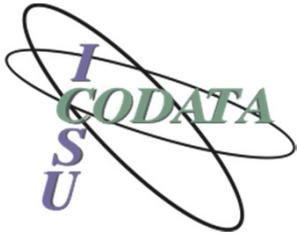
Dryad Joint Data Archiving Policy

Dryad Data Repository: http://datadryad.org/

Not-for-profit organisation whose mission is to maintain a data repository providing a home for data underpinning peer-reviewed research articles.

Joint Data Archiving Policy: http://datadryad.org/jdap

Joint declarations from journal editorial boards, Feb 2010, in American Naturalist, Evolution, the Journal of Evolutionary Biology, Molecular Ecology, Heredity, and other key journals in evolution and ecology:
http://www.journals.uchicago.edu/doi/full/10.1086/650340

This journal requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive, such as GenBank, TreeBASE, Dryad, or the Knowledge Network for Biocomplexity.

Many other journals have developed policies: e.g. **20 BioMed Central** titles adopted data policies between Aug 2011 and Mar 2012.
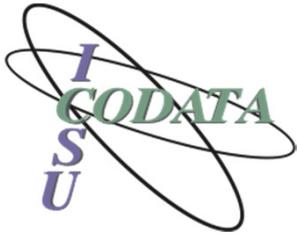
Policies / Standards / Repositories

Journal data availability policies:

JorD project
http://jordproject.wordpress.com/ reports that nearly 50% of journals sampled have a data availability policy of some sort (though only 25% of these can be characterised as 'strong').

FIGURE 5: LOCATIONS OF DATA SUGGESTED BY JOURNAL DATA SHARING POLICIES
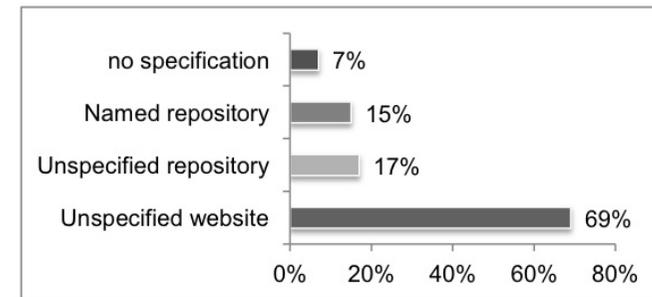
Proposal to build an information resource – a database like RoMEO detailing journal data availability policies.

Journal policies generally not clear or specific about repository, standards etc.

Clarify relations between **Policies > Standards > Data repositories**

What are the policy requirements from funders, journals, what is said if anything about standards and repositories?

What standards have community uptake, are used in repositories?

What repositories have accreditation, employ given standards, are recommended in policies.

Towards a combined information resource: http://www.biosharing.org/

Contact: Susanna Sansone (OERC) and Rebecca Lawrence (F1000)

biosharing.org

POLICIES	STANDARDS	DATABASES
A catalogue of data preservation, management and sharing policies from international funding agencies and regulators.	FORMATS TERMINOLOGIES CHECKLISTS BioPortal OBO mibbi equator network — A catalogue of reporting standards (minimum reporting guidelines, exchange formats and terminologies) and organizations that develop these.	A catalogue of databases, described according to the BioDBcore guidelines, along with the standards used within them; compiled in collaboration with 2012 NAR Database.
JISC jord SHERPA/JULIET		BioDBcore ISB International Society for Biocuration re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES OXFORD UNIVERSITY PRESS

- Works to be a *coherent, curated* and *searchable registry* of **standards** for describing, reporting and sharing experiments in life science, environmental, biomedical and biotechnological domains – collaborate with existing efforts and registries to
  - *Develop* assessment **criteria** for usability and popularity of standards
  - *Associate* standards **to journals' and funders' data policies** and **databases**
- With the objective to help stakeholders to make informed decisions *e.g.*
  - funders: which standards or databases to use or recommend in their policies; identify efforts they have been or should be funded, or require integration
  - developers: facilitate selection, use and harmonization of standards, avoiding duplications

Crystallographic Data

Raises the question of which data should be available and described.

Raw data (few MB-few GB)

Reduced data (tens of kB-few MB)

Structure data (few kB – ~1 MB)

Data flow in crystallography

Raw experimental data
Reduced/processed data
Derived data

Experiment (synchrotron or laboratory)

Data reduction

Structure solution and refinement (laboratory)

IUCr journals

Other journals

Chemistry databases (CCDC)

Biological structure databases (PDB)

● retained by scientist
▮ archived at facility (~6 months)
■ deposited
◆ published/disseminated
✚ validated
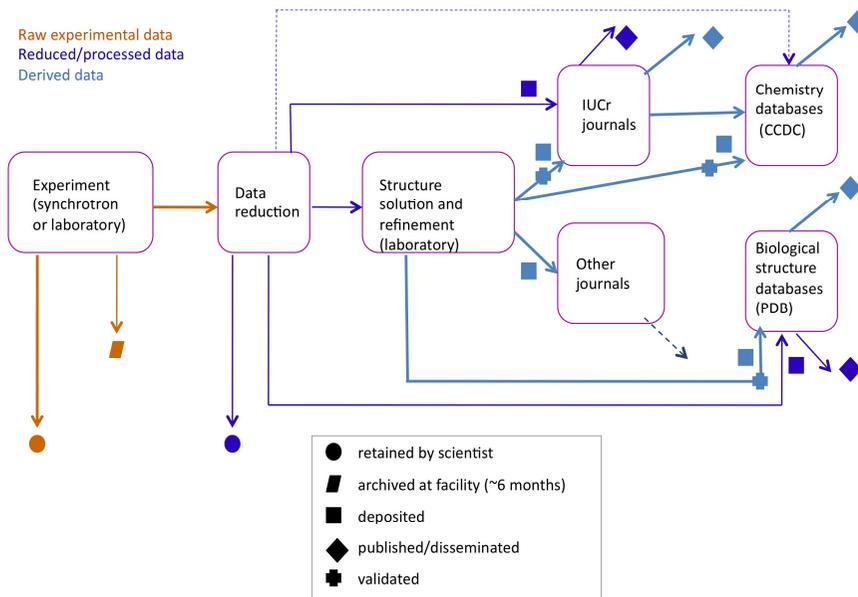
Established workflow and format for sharing structure data; but don't always share derived data, and less frequently raw data.

Slide credits: Brian McMahon, John Helliwell, Michael Hoyland.
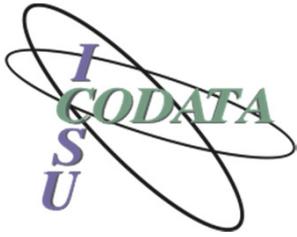
Why Publish Raw Crystallographic Data?

Increasing use of validation for structure data: found, retrospectively that **over 100 fraudulent structures had been published in Acta Crys. E from 2007-2009**:

http://www.iucr.org/__data/assets/pdf_file/0020/80273/MH_Publication-of-Small-Molecules.pdf

Publishing raw data allows validation and checking.

Above all has benefits for improving techniques for reduction and structure identification.

Helliwell: publishing **raw data** allows improved refined software use, new results and wider uptake of improved approach:

http://www.iucr.org/__data/assets/pdf_file/0004/80266/JRHelliwel_ECM28.pdf

IUCr Diffraction Data Deposition Working Group working towards a recommendations on data deposit and for a federation of repositories.

Information challenges likely to include: federated search portals, extraction of subsets of large data sets, establishment of automated procedures for expiring data sets, linking to publications, sorting by different criteria ...

Slide credits: Brian McMahon, John Helliwell, Michael Hoyland.
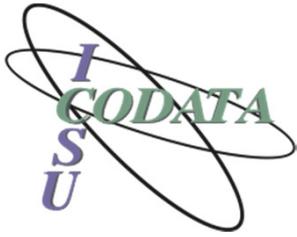
ICSU Consultation on Open Access and Metrics

ICSU (International Council of Science) consultation on Open access to scientific data and literature and assessment of research by metrics.

Contribute towards a report which will provide providing 'an analysis of the current situation and thinking on open access and the use of metrics and a statement of ICSU's overall policies' with regard to these things.'

ICSU feels the need to clarify position, work with scientific unions on this.

Concerns about the Gold OA business model (for data and publications) and where this leaves researchers without grant funding.

What type of of metrics, to what end? **What sort of metrics would provide appropriate incentives for OA for literature and data?**

Consultation workshop on 25 September involved ICSU members, International Scientific Unions, CODATA and WDS.

Short report and position statement to be produced by late spring 2014, released before or at ICSU General Assembly in August.
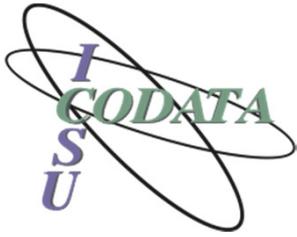
International Research Programmes: Lessons from IPY

IPY, **International Polar Year 2007-2008:** http://www.icsu.org/news-centre/publications/reports-and-reviews/ipy-summary

Mark Parsons et al, The State of Polar Data
http://www.arctic.gov/downloads/Arctic_SAP/Misc/State_of_Polar_Data20100514_distribute.pdf

Particular challenge in interdisciplinary projects.

Disparate data sharing cultures in research disciplines: some disciplines have established cultures of data sharing; others not.

Some disciplines have established data infrastructures others not. Dependency on funding streams can lead to orphan data.

How readily can established data centres expand to cover other areas in interdisciplinary projects? **Need for data archives that can deal with diverse data types, long tail…**

Culture and priorities: Despite good intentions a lack of follow through and integration: essential to involve data scientists directly in the science at all levels early in the process and throughout. This means funding data science as part of the scientific effort.

As a result of ' "naive assumptions", [a] lack of planning and other unanticipated obstacles, properly managing the IPY data will require another decade of work.' [Parsons in Nature Article on Data Sharing, 2009, doi:10.1038/461160a]

Cultural approach: polar commons norms

PIC (Polar Information Commons) Norms:
http://www.polarcommons.org/ethics-and-norms-of-data-sharing.php

IASC (International Arctic Science Committee):
http://www.iasc.info/home/iasc/datahttp://www.scar.org/researchgroups/#Data

SCAR (ICSU Scientific Committee on Antarctic Research):
http://www.scar.org/researchgroups/#Data

Lessons to be learnt for international programmes like Future Earth

Proposals for Future Earth

Future Earth: a ten-year overarching programme to coordinate research on global environmental change and sustainability.

Aims: 'to provide the scientific evidence needed for a sustainable future.'

Research for evidence-based policy making.

Data issues will be very important.

CODATA and WDS working with the Future Earth team to raise the data agenda.

Proposed a set of high level principles and a policy requiring DMPs, use of accredited infrastructure and reporting/monitoring of implementation.

futurearth
research for global sustainability
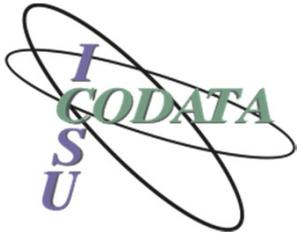
Challenges for Data Policy Makers

Policies are necessary but not sufficient. How to develop policies that promote good practice – and then support that practice?

How to make policies that fuse government objectives, funders demands but also engage researchers?

Who represents researchers and can develop norms that are appropriate and will be recognised?

Is it possible to encourage policy development from journal editorial boards, international scientific unions, learned societies?

Thank for your attention!

Simon Hodson
Executive Director CODATA
www.codata.org/blog
Email: execdir@codata.org
Twitter: @simonhodson99
Tel (Office): +33 1 45 25 04 96 | Tel (Cell): +33 6 86 30 42 59

CODATA (ICSU Committee on Data for Science and Technology), 5 rue
Auguste Vacquerie, 75016 Paris, FRANCE