GEOSPATIAL WORLD FORUM

25-29 May 2015
Lisbon Congress Center, Portugal

RECODE

Open Access to research data

Perspectives from the RECODE project

Lorenzo Bigagli, CNR-IIA
Jeroen Sondervan, Amsterdam University Press

Outline

- Introduction
- The RECODE project
  - Key findings
- Overarching recommendations
  - (Infrastructural &Technological challenges)
- Conclusions

"Sharing of Data Leads to Progress on Alzheimer's", and Parkinson's, and…

Sharing of Data Leads to Progress on Alzheimer's

By GINA KOLATA
Published: August 12, 2010

In 2003, a group of scientists and executives from the National Institutes of Health, the Food and Drug Administration, the drug and medical-imaging industries, universities and nonprofit groups joined in a project that experts say had no precedent: a collaborative effort to find the biological markers that show the progression of Alzheimer's disease in the human brain.

RECOMMEND
TWITTER
LINKEDIN
COMMENTS (155)
SIGN IN TO E-MAIL
PRINT
SINGLE PAGE
REPRINTS
SHARE

Enlarge This Image

Now, the effort is bearing fruit with a wealth of recent scientific papers on the early diagnosis of Alzheimer's using methods like PET scans and tests of spinal fluid. More than 100 studies are under way to test drugs that might slow or stop the disease.

And the collaboration is already serving as a model for similar efforts against Parkinson's disease. A $40 million project to look for biomarkers for Parkinson's, sponsored by the Michael J. Fox Foundation, plans to enroll 600 study subjects in the United States and Europe.

The work on Alzheimer's "is the precedent," said Holly Barkhymer, a spokeswoman for the foundation. "We're really excited."

Michael Temchine for The New York Times
Neil Buckholtz, chief of the Dementias of Aging Branch at the National Institute of Aging, in the National Institutes of Health

ACCESS DATA AND BIOSPECIMENS

AVAILABLE RESOURCES

CONTACT US

Parkinson's Progression Marker's Initiative (PPMI)

LRRK2 Cohort Consortium

BioFIND

Preclinical Tissue

"No one would own the data. No one could submit patent applications, though private companies would ultimately profit from any drugs or imaging tests developed as a result of the effort" (NYT, 12 Aug. 2010)

RECODE

Open Access – a dynamic period for policy & practice

- **2003 Directive 2003/98/EC: on the re-use of public sector information**
  - **Revised in June 2013**
- 2013 The RDA was launched
  - It is planning its 6th plenary and currently has more than 2500 members
- April 2013 The RCUK open access policy came into effect
- May 2013 White House launches open data policy
- **July 2013 EC Public Consultation on Open Research Data**
- **December 2013 European Commission launched the Open Data pilot in H2020**
- November 2014
  - Dutch government announces open access to data as a key policy aim
  - Gates Foundation announces world's strongest policy on open access research

Open Access – European Policy

- Europe 2020 Strategy
  - Digital Agenda for Europe
    - Sets out the role that information and communication technologies (ICT) must play in order to assist in meeting the goals for 2020
  - Linked with Horizon 2020 Framework programme
    - Europe and Member States should provide open access to scientific data generated by publicly funded research, particularly European Commission-funded research
- Free internet access to and use of publicly-funded (scientific publications and) data (EC, 2012:13)
  - Including original scientific research results, raw data & metadata, sources materials, digital representations of pictorial & graphical materials, scholarly multimedia material (Berlin Declaration)
- Data can be numerical/quantitative, descriptive/qualitative or visual, raw or analyzed, experimental or observational
  - Examples are digitized primary research data, photographs & images, films, etc. (EC 2012: 45)

RECODE

H2020 Pilot on Open Access to Research Data

- Requirements
  - Detailed data management plan covering individual datasets (within 6 months)
  - Deposit the research data, preferably into a research data repository
  - Take measures to enable for third parties to access, mine, exploit, reproduce and disseminate (free of charge for any user) this research data
    - E.g., Creative Commons License (CC-BY, CC0)
  - Provide information about tools and instruments at the disposal of the beneficiaries and necessary for validating the results
    - E.g., specialized software, algorithms, analysis protocols. Where possible, they should provide the tools and instruments themselves

RECODE

Definition - H2020 Guidelines on Data Management

- Scientific research data should be easily
  - Discoverable
  - Accessible
  - Assessable and intelligible
  - Useable beyond the original purpose for which it was collected
  - Interoperable to specific quality standards

RECODE

Benefits & challenges

BENEFITS

- Further research
- Solves global challenges
- Improves transparency & trust
- Reduces cost
- Facilitates inter-disciplinary enquiry
- Can help validate results
- Inform decision making
- Development of new products & services

CHALLENGES

- Poorly defined roles & responsibilities
- Lack of infrastructure
- Lack of career incentives
- Lack of skills/education
- Ethical considerations
- Intellectual property issues
- Disciplinary differences
- Policy fragmentation
- Funding
- Data-gap

RECODE

Project overview

Policy RECommendations for Open access to research Data in Europe

Duration
28 months (Feb 2013 – May 2015)

Total Budget
€1,147,484.70

EC DG Research contribution
€949,488.50

Grant agreement no: 321463

RECODE

RECODE consortium

© iMapbuilder

Objectives

- To reduce stakeholder fragmentation in the open access movement

- To produce **evidence-based** policy recommendations on supporting open access to research data

Objectives

Engage stakeholders

Identify values & inter-relationships

Evaluate 4 grand challenges using 5 case studies

Identify gaps, tensions & good practice solutions

Identify and promote over-arching good practice policy solutions

Produce policy guidelines

Grand challenges

- Stakeholder values & inter-relationships

- Infrastructure & technology

- Legal and ethical issues

- Institutional and policy issues

Case studies
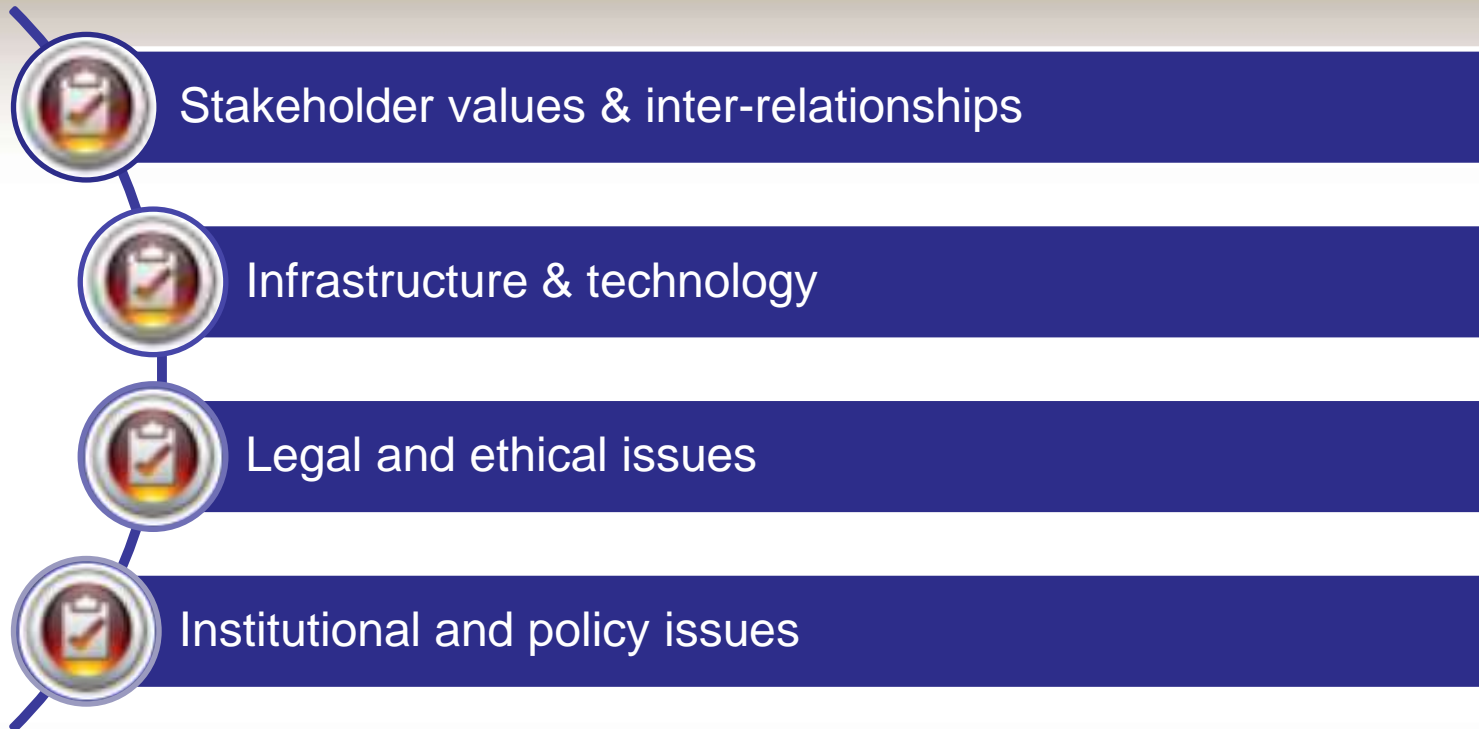
Physics	Health	Bioengineering	Earth Sciences	Archaeology
Particle Physics and Particle Astrophysics (PPPA) Group	FP7 Project EVA (Markers for emphysema versus airway disease in COPD)	Bioengineering Institute, Auckland NZ, and Virtual Physiological Human (VPH) Community	Global Earth Observation System of Systems (GEOSS)	Open Context

Grand challenges across case studies

	Physics	Health	Bioengineering	Earth Sciences	Archaeology
Stakeholder values & inter-relationships					
Infrastructure & technology					
Legal and ethical issues					
Institutional and policy issues					

Outcomes & Impacts

Key Deliverables

- Stakeholder values and ecosystems – Sept'13
- Infrastructure and technology challenges – Feb'14
- Legal and ethical issues in OA and data dissemination – April'14
- Institutional barriers and good practice solutions - June'14
- Policy guidelines for OA and data dissemination – Jan'15
- Feasibility of OA networks to support harmonization – Jan'15

Milestone Events

- Stakeholder engagement workshops
  - Infrastructural and technological challenges – Jan'14
  - Legal and ethical barriers and solutions – March'14
  - Institutional barriers and solutions – May'14
  - Policy recommendations – July'14
- Final conference Jan'15

Expected Impacts

- Support the Commission's policies on open access to scientific data
- Contribute to network-building among concerned stakeholders at the European and international levels
- Support the development of joint or common policy agendas and activities in the area of scientific data

Other key accomplishments

- 5 stakeholder workshops
- 168 workshop participants from 35 countries
- 65 Interviews with academics, researchers, policy-makers, data centres, legal experts, publishers and others
- 1 Book proposal – *Mobilising Data in a Knowledge Society*
- Specific recommendations for 4 stakeholder groups
  - Funders,
  - Research institutions
  - Data managers
  - Publishers

RECODE

Key findings

- **Stakeholder Values and Ecosystems**
  - Shared perception of OA of a public good, but conflicting value chains
    - Costs of research data

- **Infrastructure and Technology Challenges**
  - Less of a concern compared to financial, cultural and legal ones

- **Legal and Ethical Challenges**
  - Strategies already adopted to address potential legal and ethical issues, but stakeholders often subject to conflicted legal obligations, not all data can be open

- **Institutional Challenges**
  - Data quality essential for reuse and long-term preservation, need clear guidelines and mechanisms that contribute to evaluation, such as data journals and peer review; training and rewarding is paramount

RECODE

Overarching recommendation (1/2)

1. Develop aligned and comprehensive policies for open access to research data (aligned policies)

2. Ensure appropriate funding for open access to research data (funding)

3. **Develop policies and initiatives that offer researchers rewards for open access to high quality data (rewards)**

4. Identify key stakeholders and relevant networks and foster collaborative work for a sustainable ecosystem for open access to research data (collaboration)

5. Plan for the long-term, sustainable curation and preservation of open access data (long-term preservation)

RECODE

Overarching recommendation (2/2)

6. Develop comprehensive and collaborative technical and infrastructure solutions that afford open access to and long-term preservation of high-quality research data (infrastructure)

7. **Develop technical and scientific quality standards for research data (quality standards)**

8. Require the use of harmonized open licensing frameworks (licensing)

9. Systematically address legal and ethical issues arising from open access to research data (legal/ethical aspects)

10. Support the transition to open research data through curriculum-development and training (training/education)

- **Promote the concept of fitness for use**
  - **Instead of quality statements inherent to data**
  - **Integrate user feedback in the metadata**
- Recognise different levels of Open Access
  - Implement appropriate access control
  - Enforce security policies (authentication, authorisation, auditing) automatically

RECODE

Conclusions

- Two overarching issues in the mobilisation of open access to research data
  - Lack of a coherent open data ecosystem
    - Challenges and stakeholder needs are interrelated, must be addressed holistically
  - Lack of attention to the specificity of research practice, processes and data collections (including PSI)
- Build on experience and on the existing
  - Fill in the gaps: interoperability
  - No specific solutions: mediation
- Sharing is not necessarily giving for free
  - Discuss new Business Model and professional roles

RECODE

Stay engaged

VISIT OUR WEBSITE

http://recodeproject.eu

FOLLOW US ON TWITTER

• @RECODE_Project

SIGN UP TO OUR NEWSLETTER

• kush.wadhwa@trilateralresearch.com

Stay engaged

- European Geosciences Union General Assembly Meeting 2016
  - Vienna, 17-22 April 2016
- ESSI Division - Open Access to Research Data and Public Sector Information: perspective, drivers, and barriers
  - Convener: Lorenzo Bigagli
  - Co-Conveners: José Miguel Rubio Iglesias, Berit Arheimer, Stefano Nativi
- Call for Abstract deadline
  - Sometimes in January 2016

GEOSPATIAL WORLD FORUM

25-29 May 2015
Lisbon Congress Center, Portugal

RECODE

Open Access to research data

Perspectives from the RECODE project

Lorenzo Bigagli, CNR-IIA
Jeroen Sondervan, Amsterdam University Press

Grand challenges – WP1

Stakeholder values & inter-relationships

Infrastructure & technology

Legal and ethical issues

Institutional and policy issues

RECODE

Objectives and Methodology

- **Objectives**
  - To identify and map the diverse range of stakeholder values in open access and data dissemination and preservation
  - To map stakeholder values on to scientific ecosystems using case studies from different disciplinary perspectives
- **Methodology**
  - **Two stage document review**
    - 1st Stage – Broad Scoping of material and synthesis from stakeholder literature
    - 2nd Stage – Thematic analysis of a smaller sample of documents
  - **Case study research within five scientific fields**
    - Archaeology
    - Bioengineering
    - Environmental research
    - Health and clinical research
    - Particle Physics and Particle Astrophysics
  - **Stakeholder validation workshop**

RECODE

Document Review – Key Findings

- Overall drive for Open Data access within the policy documents, which is part of a wider driver for open science in general

- The values underpinning this move are the view of science as an open enterprise, where knowledge is sought and where discovery rests on scientists working together to solve specific challenges, which increasingly are becoming interdisciplinary in nature

- The argument for publicly funded science to be open to the public is also strong, although it is not entirely clear how often this openness should be operationalized

- When discussing Open Data there is a clear tendency to refer to science as a whole sector, thus there is a danger that differences between disciplines are ignored in further policy making

Stakeholder values as identified in document review

- Open Access is that it is addressed differently by stakeholders in the research ecosystem.
  - **High level policy makers** focus at the very general level and argue for Open Access in terms of very broad social and economic benefits as well as seeing it as a development that will improve science.
  - **Funders** are increasingly motivated to ensure that the allocation of publically funded research yields good value for money.
  - **Stakeholders from within the infrastructural, libraries, repositories,** see value in Open Access to Data as a way of improving the means by which data is made more accessible, and they are motivated to meet the needs of Open Access within their business cases and service provision.
  - **Publishers** are adapting to the open publishing environment and are developing new types of business models to facilitate that. Here the question of where the cost for Open Access publishing will rest is still undecided.

Values and motivations from Case Studies

- There is some mapping of values and motivations from stakeholder review to case studies
    - Data has a definite value for scientists (knowledge production, hypothesis and model testing etc.)
    - Access to more data = opportunities for testing, linking, integrating ➔ faster advancement within their disciplines
    - Helps to avoid duplication of effort (clear benefits to health researchers and patient groups)
    - Open data is seen to fosters multidisciplinary research and allows for the tackling of new research challenges

RECODE

Operationalising open data

- Although values and motivations mapped on to those identified in the stakeholder review, scientists have reservations when it comes to operationalizing open data

    - Data must be 'meaningful' before it is made open, this may include a lot of work from scientists with unclear benefit
    - Currently no reward for 'data work' – peer reviewed publications
    - Data comes in different forms,
    - Lack of standardisation within many disciplines, idiosyncratic and individual ways of managing and annotating data
    - Ethical and legal issues of opening up patient and location data
    - No 'one size fits all' – data is embedded within different research cultures, traditions and practices
    - Sustainable infrastructure is needed to host data, current short term funding models are unable to ensure this
    - Data Licencing standards need to be considered
    - Peer review mechanism for data to ensure accuracy, validity and reliability

RECODE

Infrastructure & Technology challenge

Stakeholder values & inter-relationships

Infrastructure & technology

Legal and ethical issues

Institutional and policy issues

Scope

- Infrastructural and technological factors that may hinder Open Access and preservation of research data

- *"Infrastructure"* includes

    - Technological assets (hardware and software)
    - Human resources
    - Procedures for management, training and support its continuous operation and evolution

RECODE

I&T Stakeholders

- **Producers**
  - The source of research data
  - E.g. researchers elaborating raw sensor datasets
- **Disseminators/Curators**
  - The actors in charge of the distribution and preservation infrastructure (information systems, e-infrastructure) for storage, access, and maintenance of research data
  - E.g. publisher, library
- **Funders**
  - The parties providing financial and policy support to data collection activities in research
  - E.g. research councils, funding agencies
- **End users**
  - The generic final recipient of research data
  - E.g. researchers, the industry, governmental agencies, data users at large

RECODE

I&T Challenges

- **Heterogeneity** – relates to the variety of data at any level, e.g. format and encoding issues, data accessibility, protocol interoperability, but also high level issues

- **Accessibility** – relates to the volume of data and to its impact on the infrastructure capabilities and architecture

- **Sustainability** – relates to the long-term impact of maintaining and operating an open infrastructure for research data

- **Quality** – relates to the technological support for the evaluation of data suitability and appropriateness

- **Security** – relates to the restrictions on the usage, access, and consultation of data and metadata, and their enforcement from a technical viewpoint

I&T Challenges Overview

	Creator	Disseminator / Curator	Funder	End user
Heterogeneity	standardization, encoding, semantics	interoperability, reuse, data cross-walk, internationalisation		standardization, encoding, semantics
Accessibility	bandwidth	standardization, storage, scalability, distribution		discoverability, storage, bandwidth
Sustainability	persistent identification	obsolescence, reuse, data migration, persistent identification	governance	obsolescence
Quality	provenance, training, fitness for use	provenance, training, completeness, peer review	certification	provenance, peer review, fitness for use
Security	authorization, attribution, licensing	authentication, authorization, accounting, privacy, obfuscation	licensing	authentication, privacy, trust

RECODE

Recommendations on I&T for Open Access to Research Data

- Promote culture of standard
  - Both in education and research practice
  - Reinforce importance of metadata and data standardisation (e.g. common models and encodings)
- Adopt System-of-Systems approach
  - Mitigate heterogeneity and accessibility issues
  - Cf. GEOSS
- Enforce persistent digital identifiers (PID)
 - Both data and users
  - Ensure interoperability of PIDs

RECODE

- Promote culture of data management
  - Cf. physics, social sciences, libraries
  - Almost absent in Public Sector Information
  - Investigate new professional roles
- Resort to virtualization technologies
  - Periodical migration (e.g. format convertion, transcoding)
- Enforce presence of complete and accurate metadata
  - In particular provenance information

RECODE

- **Promote the concept of fitness for use**
  - **Instead of quality statements inherent to data**
  - **Integrate user feedback in the metadata**
- Recognise different levels of Open Access
  - Implement appropriate access control
  - Enforce security policies (authentication, authorisation, auditing) automatically

RECODE

RECODE Stakeholder Taxonomy

Users

Curators

Disseminators

Creators

Funders
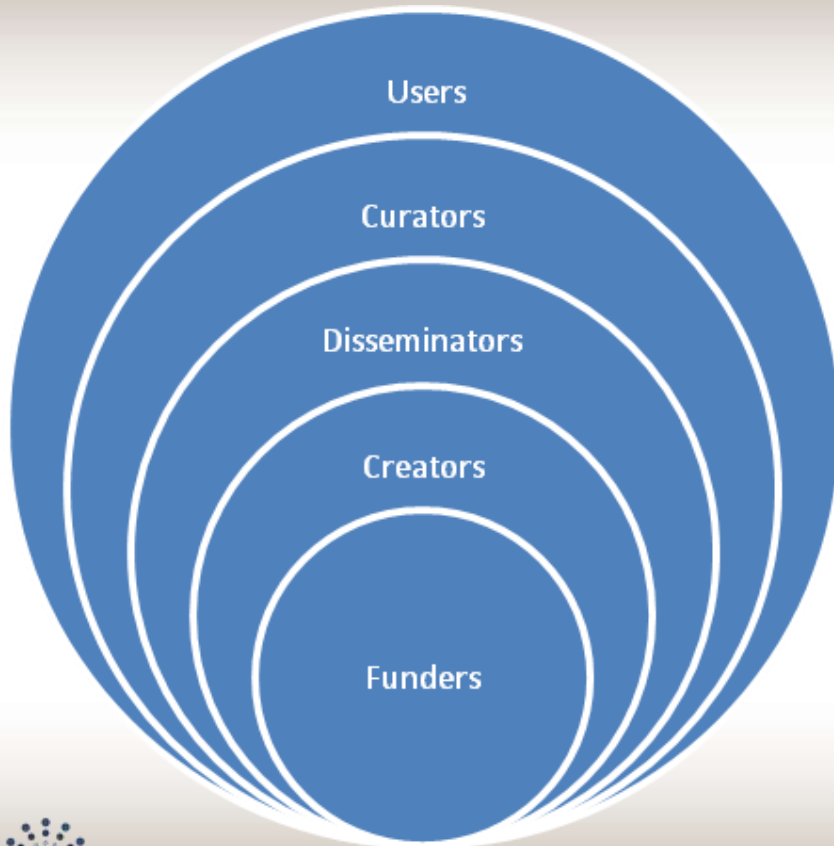
- We congregate the disseminator and curator roles
  - We assume they share similar concerns for our purpose
- Citizens may be considered as research data user/producers (cf. Citizen Science)
  - We assume their involvement in use and production of data is mediated by appropriate applications (e.g. mobile apps) that practically isolate them from the implied technological and infrastructural issues

Methodology

- Literature review

- Online survey questionnaire

- Case-study interviews

- Validation workshop

- Advisory board comments

Case-study Interviews

- Experience with OA publications, but not with data publications or data preservation

- Metadata are considered crucial to enable retrieval, re-use and preservation of research data

- Financial and legal barriers are considered higher priority then technical ones

- Data management plans are being developed, but still at an early stage

- Solutions for data management and preservation are not common nor centralized
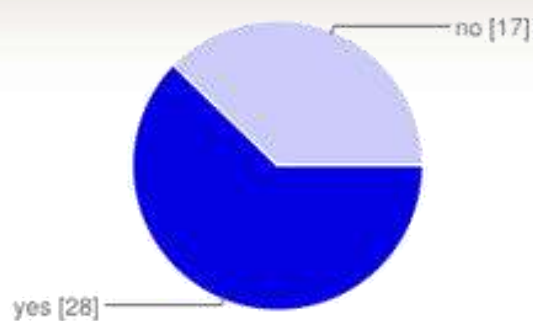
Online questionnaire
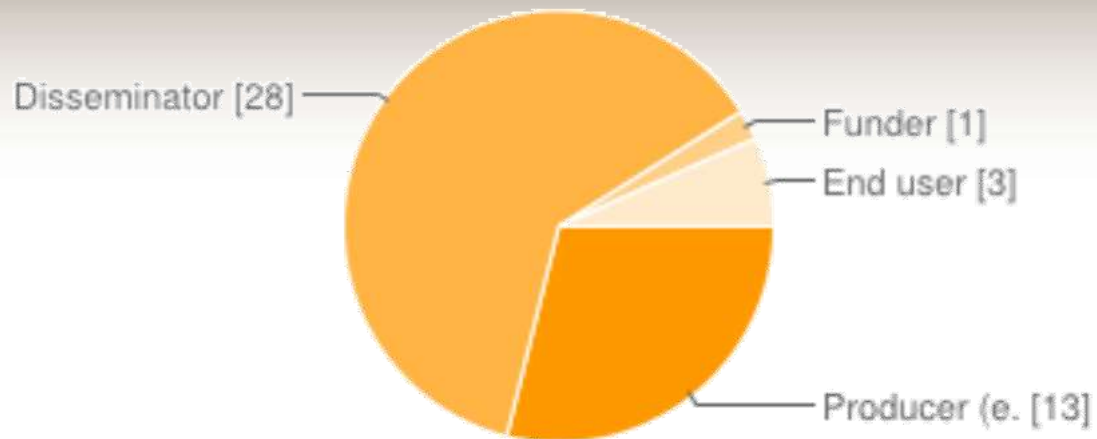
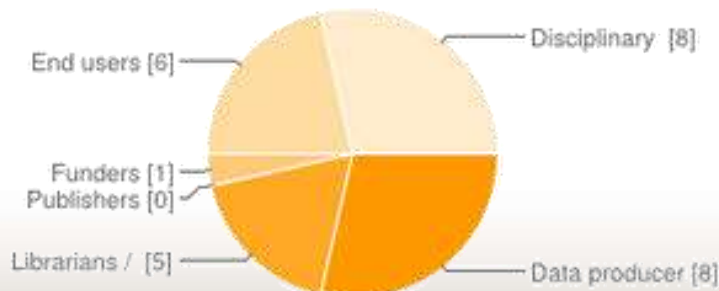|		
Producer (e.g. researcher)	**13**	29%
Disseminator/Curator (e.g. publisher, librarian)	**28**	62%
Funder	**1**	2%
End user	**3**	7%

Online questionnaire

Do you think research data should all be preserved indefinitely, in principle?

yes	28	62%
no	17	38%

If not, who should decide what to preserve and until when?

Data producers	8	29%
Librarians / repository managers	5	18%
Publishers	0	0%
Funders	1	4%
End users	6	21%
Disciplinary associations / peer review	8	29%

RECODE

Online questionnaire

Who should be primarily responsible for storing European research data and making them accessible?

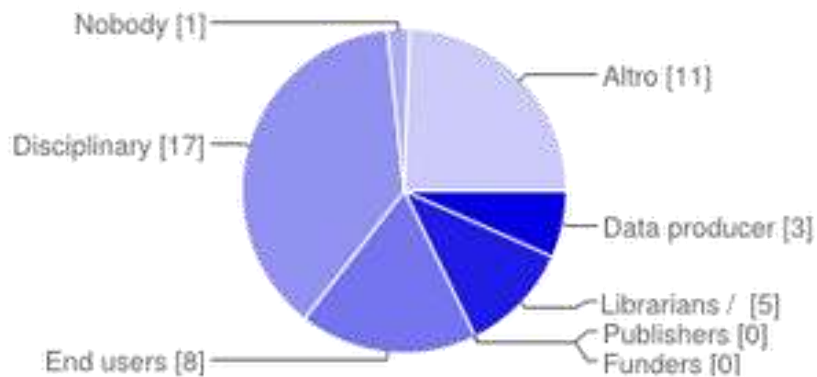|		
Data producers	11	12%
Digital libraries / Institutions (e.g. Universities)	22	25%
National institutionalized repositories	22	25%
Centralized European repository	18	20%
Publishers	3	3%
Funders	6	7%
Altro	7	8%
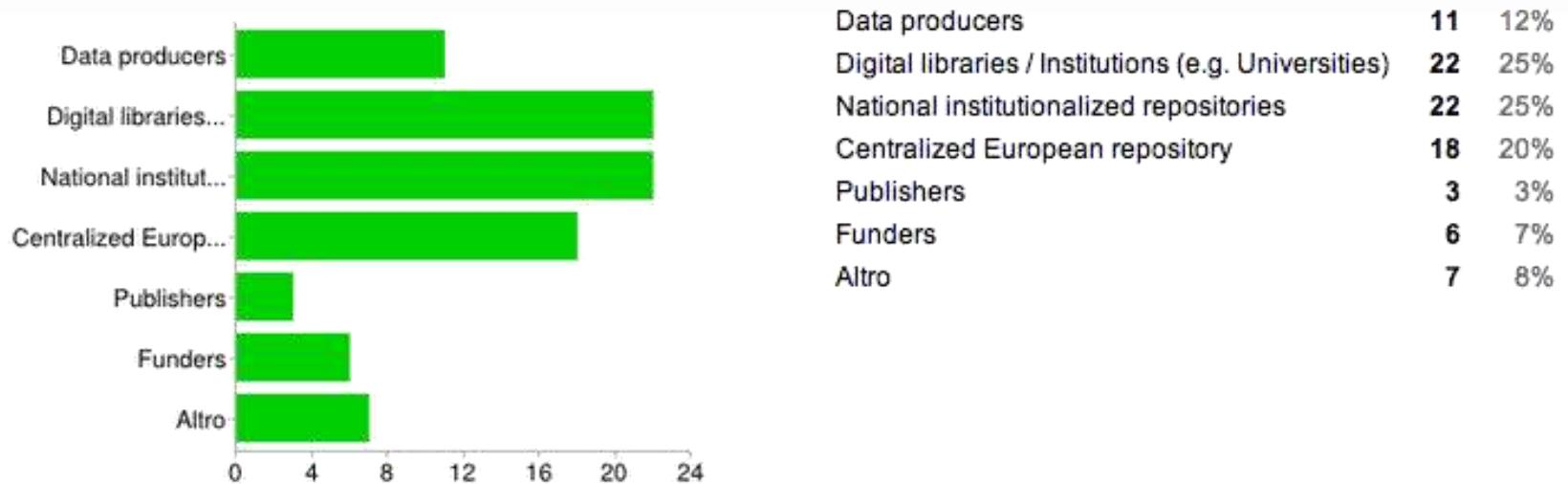
Online questionnaire

With the increase of openly accessible data, what factors do you think will have the greatest impact?

Heterogeneity of data formats	14	9%
Application interoperability	18	11%
Data access / download	15	9%
Catalogues and search engines	11	7%
Storage capacity on the client side	2	1%
Obsolescence of data formats	7	4%
Data preservation	21	13%
Energy footprint	0	0%
Bandwidth	1	1%
Data completeness (gaps, etc.)	5	3%
Data documentation (metadata)	36	22%
Data quality	21	13%
Privacy and security issues	10	6%
Altro	2	1%

RECODE

Online questionnaire

Who should evaluate the quality of research data?

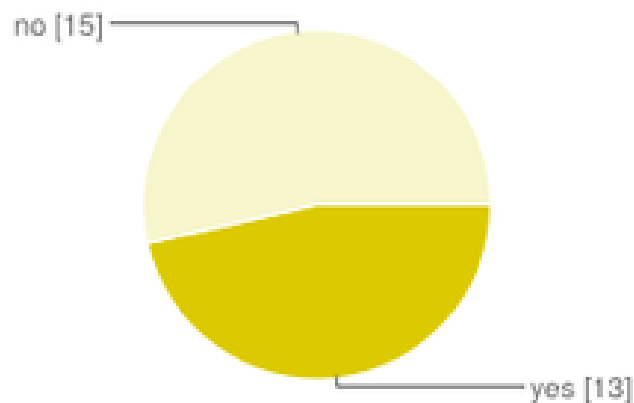|		
Data producers	3	7%
Librarians / repository managers	5	11%
Publishers	0	0%
Funders	0	0%
End users	8	18%
Disciplinary associations / peer review	17	38%
Nobody	1	2%
Altro	11	24%

Online questionnaire

Disseminator/Curator of research data

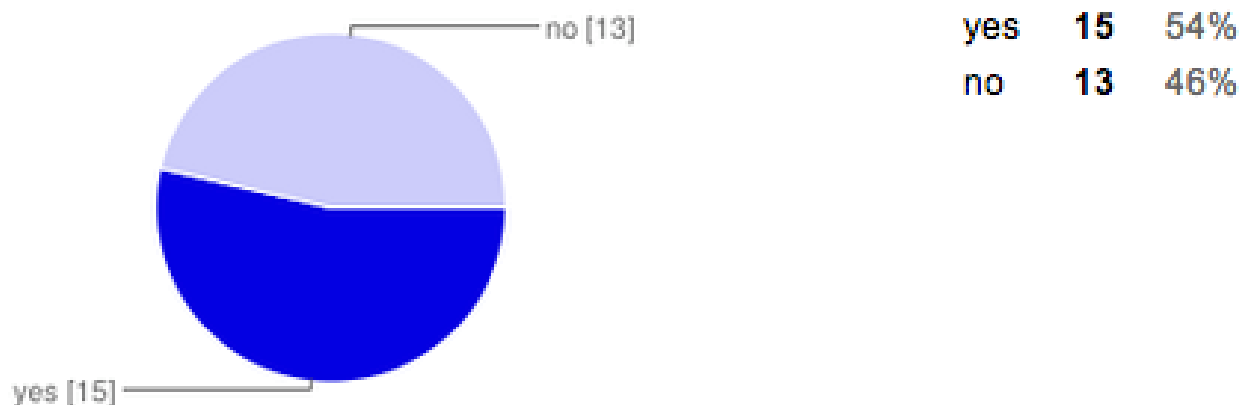Do you have direct experience of implementing Open Access to research data?

yes	**13**	46%
no	**15**	54%

no [15]

yes [13]

RECODE

Online questionnaire

Disseminator/Curator of research data

Do you take user feedback into account for improving the quality of your data?

yes	**15**	54%
no	**13**	46%

no [13]

yes [15]

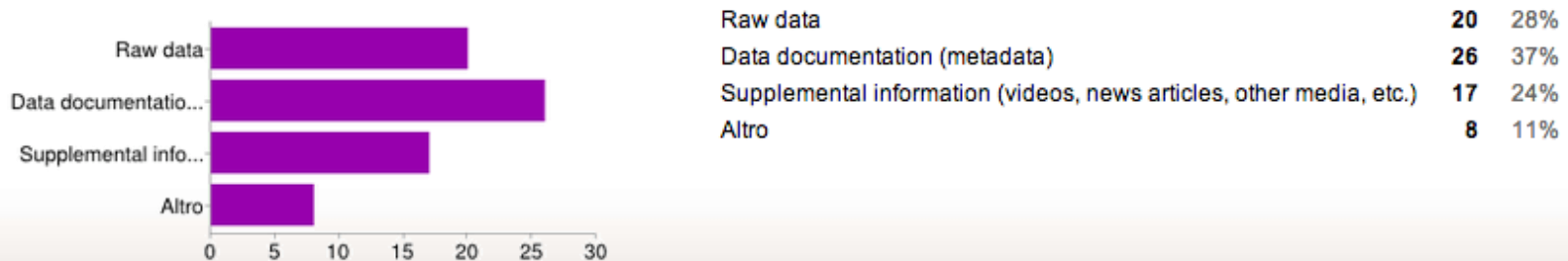# Online questionnaire

Disseminator/Curator of research data

As a disseminator/curator of research data, are you offering tools in order to associate research data to scientific publications?

yes	**13**	46%
no	**15**	54%
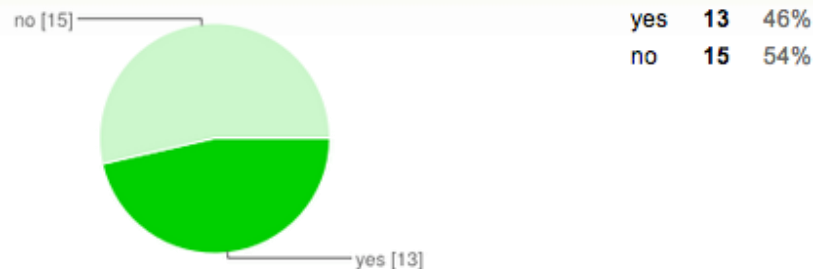
What kind of additional information would you consider relevant to complement scientific publications?

Raw data	**20**	28%
Data documentation (metadata)	**26**	37%
Supplemental information (videos, news articles, other media, etc.)	**17**	24%
Altro	**8**	11%

Online questionnaire

Producer of research data

Does your institution/organization have a data management plan?

don't know / [3]

no [6]
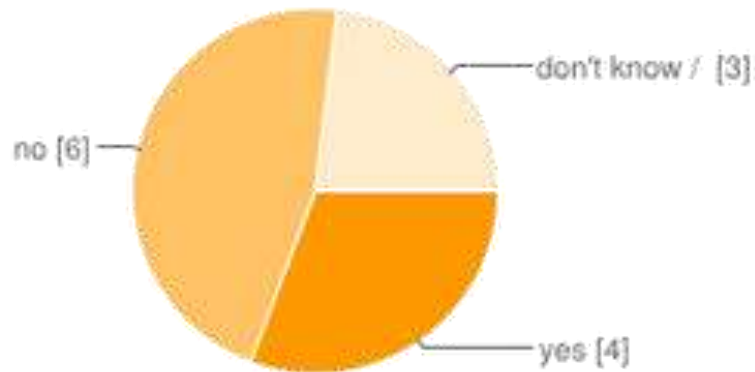
yes [4]

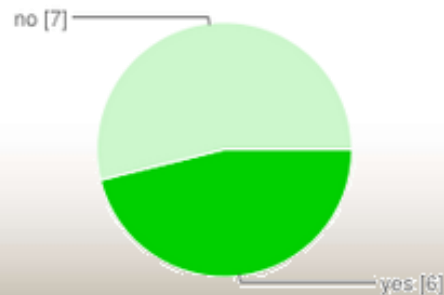|		
yes	**4**	31%
no	**6**	46%
don't know / not applicable	**3**	23%

RECODE

Online questionnaire

Producer of research data

How would you categorize the level of adoption of Open Access in your field?
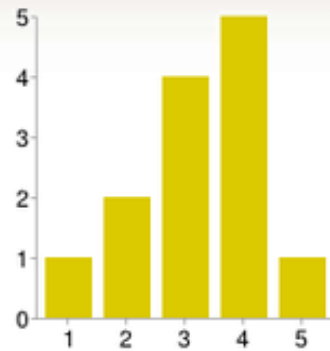
1	1	8%
2	2	15%
3	4	31%
4	5	38%
5	1	8%

Do you have direct experience of releasing your data according to an Open Access policy?

no [7]

yes [6]

yes	6	46%
no	7	54%

RECODE

EC Public Consultation on Open Research Data

- 2nd July 2013, in Brussels
- ~130 attendants stakeholders from the research community, industry, funders, libraries, publishers, infrastructure developers, etc.
- Focused on five questions
  - How can we define research data and what types of research data should be open?
  - When and how does openness need to be limited?
  - How should the issue of data re-use be addressed?
  - Where should research data be stored and made accessible?
  - How can we enhance data awareness and a culture of sharing?

RECODE

EC Public Consultation on Open Research Data

- How can we define research data and what types of research data should be open?
  - Definitions vary, with some contributions defining research data as potentially all data (including public sector information), and some limiting it to data that is the product of research
    - For researchers, research data includes all data from an experiment, study or measurement, including metadata and details on processing data
    - For publishers, data linked to publications is part of the publication

EC Public Consultation on Open Research Data

- **How should the issue of data re-use be addressed?**
  - Discussions about licensing, but also about technical aspects of open research data
    - Not just on whether and how data should be re-used, but also on the adequacy of e-infrastructures for data re-use
  - Directive on the re-use of public sector information (2003/98/EC, currently under revision) was mentioned several times
    - While public sector information (PSI) is distinct from research data and governed by a specific directive, it is important to remember that this type of information can also be useful for research

EC Public Consultation on Open Research Data

- Where should research data be stored and made accessible?
  - The need for improved data management practices and better data accessibility is a key concern
  - Issues closely linked with data preservation and sustainability of data repositories
  - The readiness of professionals to engage in data curation was also highlighted
    - All stakeholders agreed that any funding body policy on open research data must call on researchers to take the issue of data management seriously by developing data management plans (DMPs) for their research projects

RECODE