Enhancing discoverability of public health and epidemiology research data: Report Annexes

Contents

Annex A: Data Documentation and Access: Characterizing Current Practice

List of Tables

Review

Of the 49 studies/organisations previously identified, a random sample of 13 were reviewed to identify exemplar current practices in data documentation and access.

Initial key findings were:

- PDF is one of the most common formats.
- All provide links to and/or descriptions of publications.
- Of the 13 reviewed, SAS, STATA and SPSS are the statistical packages most commonly supported. Other such formats include ASCII and CSPro.
- Of the 13 reviewed, 4 offer online data visualisation/analysis tools.

1: The 49 studies and organisations

ID	Name
1	1958 National Child Development Study (UK)
2	1970 British Cohort Study (UK)
3	ALPHA Network
4	Analysis of a sample of type 2 diabetic patients with obesity or overweight and at cardiovascular risk: a cross sectional study in Spain
5	Australasian Association of Cancer Registries
6	Avon Longitudinal Study of Parents and Children (UK)
7	Birth to Twenty
8	Born in Bradford (UK)
9	Cancer Registries
10	CartaGene
11	Concord 2
12	ELFE, Growing up in France (France)
13	European Prospective Investigation into Cancer and Nutrition (EPIC)
14	European Social Survey
15	Generation Scotland (UK)
16	Growing up in New Zealand
17	ICPSR
18	INDEPTH Network
19	International Epidemiological Databases to Evaluate AIDS (IeDEA) in sub-Saharan Africa
20	IPUMS International Project
21	Longitudinal Study of Young People in England (UK)
22	Measure DHS, Demographic Health Surveys
23	MIDUS Midlife in the US
24	Millennium Cohort Study (UK)
25	Monitoring the efficacy and safety of three artemisinin based-combinations therapies in Senegal: results from two years surveillance
26	Norwegian Mother and Child Cohort Study (Norway)
27	Pelotas Birth Cohort Study
28	Population Health Metrics Research Consortium Gold Standard Verbal Autopsy Data 2005–2011
29	Prevalence of schistosome antibodies with hepatosplenic signs and symptoms among patients from Kaoma, Western Province, Zambia
30	RAND Centre for the Study of Ageing
31	SABE - Survey on Health, Well-Being and Aging in Latin America and the Caribbean
32	SABRE Southall and Brent Revisited
33	Scottish Longitudinal Study (Scotland)
34	Study of Environment on Aboriginal Resilience and Child Health
35	The China Kadoorie Biobank
36	The China-Anhui birth cohort study
37	The Epidemiology - France web portal A collaborative project in epidemiology
38	The International Collaboration of Incident HIV and Hepatitis C in Injecting Cohorts Study
39	The Limache birth cohort study

40 The Motorik-Modul Longitudinal Study (Germany)
41 The National Survey of Sexual Attitudes and Lifestyles (UK)
42 The spectrum of paediatric cardiac disease presenting to an outpatient clinic in Malawi
43 TwinsUK and Healthy Aging Twin Study (UK)
44 Udaipur health and Immunization studies
45 UK Biobank (UK)
46 Understanding Society (UK)
47 Whitehall Study (UK)
48 WHO Study on Global AGEing and Adult Health (SAGE)
49 Worldwide Antimalarial Resistance Network

To see the more-detailed analysis of the websites of the selected data sets, please see the attached Excel spreadsheet appended to the report as a separate data file.

1: Sample

ID	Name	Coverage	Source
6	Avon Longitudinal Study of Parents and Children (UK)	UK	http://www.bristol.ac.uk/alspac/
12	ELFE, Growing up in France (France)	France	http://www.elfe-france.fr/index.php/en/
13	European Prospective Investigation into Cancer and Nutrition (EPIC)	Lyon, UK (London), Denmark, France, Germany, Greece, Italy, Norway, Spain, Sweden, The Netherlands	http://epic.iarc.fr/index.php
14	European Social Survey	Albania, Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Kosovo, Latvia, Lithuania, Luxembourg, The Netherlands, Norway, Poland, Portugal, Romania, Russian federation, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, Ukraine, UK	http://www.europeansocialsurvey.org/
18	INDEPTH Network	South Africa, Guinea Bissau, Senegal, Ethiopia, Ghana, Gambia, Tanzania, Uganda, Malawi, Burkina Faso, Kenya, Mozambique, Nigeria, Cote d'Ivoire, India, Bangladesh, Vietnam, Cambodia, Thailand, Indonesia, Papua New Guinea	http://www.indepth-ishare.org/index.php/home
20	IPUMS International Project	Argentina, Armenia, Austria, Bangladesh, Belarus, Bolivia, Brazil, Burkina Faso, Cambodia, Cameroon, Canada, Chile, China, Colombia, Costa Rica, Ecuador, Egypt, El Salvador, Fiji, France, Germany, Ghana, Guinea, Greece, Haiti, Hungary, India, Indonesia, Iraq, Iran, Ireland, Israel, Italy, Jamaica, Jordan, Kenya, Kyrgyz Republic, Malawi, Malaysia, Mali, Mexico, Mongolia, Morocco, Nepal, Netherlands, Nicaragua, Pakistan, Palestine, Panama, Peru, Philippines, Portugal, Puerto Rico, Romania, Rwanda, Saint Lucia, Senegal, Sierra Leone, Slovenia, South Africa, South Sudan, Spain, Sudan, Switzerland, Tanzania, Thailand, Turkey, Uganda, United Kingdom, United States, Uruguay, Venezuela, and Vietnam	https://international.ipums.org/international/

22	Measure DHS, Demographic Health Surveys	Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Cape Verde, Central African Republic, Chad, Comoros, Congo (Brazzaville), Congo Democratic Republic, Cote d'Ivoire, Equatorial Guinea, Eritrea, Ethiopia, Gabon, Gambia, Ghana, Guinea, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Namibia, Niger, Nigeria, Nigeria (Ondo State), Rwanda, Sao Tome and Principe, Senegal, Sierra Leone, South Africa, Sudan, Swaziland, Tanzania, Togo, Uganda, Zambia, Zimbabwe, Albania, Armenia, Azerbaijan, Egypt, Jordan, Moldova, Morocco, Tunisia, Turkey, Ukraine, Yemen, Central Asia, Kazakhstan, Kyrgyz Republic, Tajikistan, Turkmenistan, Uzbekistan, Afghanistan, Bangladesh, Cambodia, India, Indonesia, Lao People's Democratic Republic, Maldives, Nepal, Pakistan, Philippines, Sri Lanka, Thailand, Timor-Leste, Vietnam, Samoa, Bolivia, Brazil, Colombia, Dominican Republic, Ecuador, El Salvador, Guatemala, Guyana, Haiti, Honduras, Mexico, Nicaragua, Paraguay, Peru, Trinidad and Tobago	http://dhsprogram.com/Data/
23	MIDUS Midlife in the US	United States of America	http://www.midus.wisc.edu/
26	Norwegian Mother and Child Cohort Study (Norway)	Norway	http://www.fhi.no/eway/default.aspx?pid=240&trg=Main_6664&Main_6664=6894:0:25,7372:1:0:0:::0:0
33	Scottish Longitudinal Study (Scotland)	Scotland	http://sls.lscs.ac.uk/
34	Study of Environment on Aboriginal Resilience and Child Health	Australia	https://www.saxinstitute.org.au/our-work/search/
48	WHO Study on Global AGEing and Adult Health (SAGE)	South Africa, China, Ghana, India, Mexico, Russian Federation	http://www.who.int/healthinfo/sage/en/
49	Worldwide Antimalarial Resistance Network	Thailand, Kenya, Brazil, Senegal	http://www.wwarn.org/

ID	Study protocol(s)	Data documentation	Data access	Online data visualisation/ analysis	Publication links/ descriptions	Social media/other forms of communication
6	Questionnaires available as PDFs	Downloadable data dictionary	Data access policy and guidance available online (PDF)	No	Yes	Facebook, Google+, Soundcloud, MyYahoo, YouTube, Twitter and QR code
12	Online description of key stages			No	Yes	Newsletter
13	Questionnaires and statistical methods	Questionnaires and descriptions of the cohort/anthropometric measurements available online. All measurements were standardised using EPIC-SOFT (available in multiple languages) to enable comparison.		No	Yes	RSS and LinkedIn
14	Online descriptions and PDFs	Available online by year, country and theme	Data is available for download online in SAS, SPSS and STATA	Yes	Yes	Email
18	Study overviews available online	Online data dictionaries available with variable names, labels and descriptions. DDI compliant metadata is available. Downloadable	Two types: 'Public use files' and 'Licensed files' and data can be filtered according to centre.	Yes	Yes	Email

		microdata (must register/login)				
20	Questionnaires available as PDFs and HTML	When data are extracted, codebooks are generated.	Online variable selection (integrated and harmonised variables options available). Users must be registered before data extracts may be downloaded. Extract system provides support for the import of generated ASCII files into SPSS, SAS and STATA.	Yes	Yes	Newsletter
22	Questionnaires are available for download (PDF)	Data are recoded (variable names, locations etc.) with all recording manuals available for download online. Use the DHS Recode.	Must be a registered user of the website. Application for data must include contact information, research project title and description of intended analysis. Certain data require users to sign additional agreements/terms of use. Files are distributed as compressed Zip files. List of potential data downloads is available online. Data are available in ASCII, STATA, SPSS, SAS and CSPro	Yes -HIV/AIDS Survey Indicators Database & STAT compiler	Yes	Email, Facebook, Twitter, YouTube, LinkedIn, Pinterest, Blog and Mobile phone app
23	Available in catalogue/ICPSR website	Categories, codes, variable grouping etc. available through Colectica catalogue. DDI Codebook/Lifecycle, Dublin Core and MARC21 XML metadata	Data is available for download online in SAS, SPSS, STATA, ASCII and Delimited from ICPSR website	No	Yes	LinkedIn, Facebook, Google or MyData
26	Online description and downloadable questionnaires	Basic participant response figures available online for version VIII	Applications must be submitted and approved before use of data and/or biological materials is enabled. Researcher(s) may need to sign a contract with the Norwegian Institute of Public Health	No	Yes	Facebook, Twitter, YouTube, RSS feed and newsletter

33	Online descriptions of creation/development available as individual 'Technical Working Paper'	Online data dictionary with searchable lists of tables and variables	2 methods - 'safe setting' in Edinburgh or remote access involving use of variable names and labels only for creation of syntaxes. These syntaxes are then returned to the Support Officer who will arrange for the analysis to be run and the results, once cleared returned. More details available online	No	Yes	Twitter, email, blog
34	Study protocol available as an academic paper		Describes 'SURE' Secure, Unified Research Environment' but does not make it clear if SEARCH data can be accessed	No	Yes	Email, Facebook, RSS, LinkedIn and Twitter
48	Summary of measures in questionnaires and the questionnaires themselves available in PDF	Related materials, study description, data dictionary (variables include name, label and question) and related citations all available. DDI compliant metadata available in PDF.	Through WHO Multi-Country Studies Data Archive.	No	Yes	RSS, YouTube, twitter, Facebook and Google+
49	Searchable procedures available for download	When sharing data (clinical, pharmacology, In vitro) data dictionaries should be accompany data	WWARN standardise data and then make these available including of audit trail and original dataset to the data contributor and nominated individuals. Researchers (under 'Third party data access') should contact the data owner(s) as they alone can grant access to the transformed data.	Yes	Yes	Facebook and twitter

Annex B: Profiles – Documenting Cohorts and Data Resources

Cohort Profiles

Cohort profiles, as published by the International Journal of Epidemiology[1], describe particular epidemiological studies. These concise publications of around 2500 to 3000 words, present key facts about cohorts using descriptors such as, '*Why was the cohort set up?'[2]* and *'What has been measured?'[2]* Cohort profiles also include generalised descriptions of any anthropometric measurements taken and provide an indication of the key findings to date.

A key part of cohort profile publications is the section, '*Can I get hold of the data? Where can I find out more'[2].* Having this kind of information in a standardised way has the potential to support research endeavours and promote adoption of the research data lifecycle. It is also in keeping with recent data sharing and open access initiatives.

Case Study

A search on PubMed[3] using the phrase 'Cohort Profile International Journal of Epidemiology' returned 224* results with the most recently indexed publication* being Cohort Profile: the Health and Retirement Study (HRS)[4]. This is a longitudinal, US-based study with over 37000 participants from 23000 households. Sister studies include, but are not limited to, the English Longitudinal Study of Aging[5] and The China Health and Retirement Longitudinal Study (CHARLS)[6]; both of which have cohort profiles.

Data Resource Profiles

A data resource profile describes epidemiological data available for use by researchers for purposes of testing hypotheses and performing analyses. Similarly to cohort profiles, these publications are short at around 2500 to 3000 words and include key information such as '*Data resource area and population coverage*'[2] and '*Strengths and weaknesses'*.[2]

Descriptions of how and where the data may be accessed are of particular importance to characterising the dataset especially to potential future users. Recently published examples include, Countdown to 2015: Maternal, Newborn and Child Survival[7] and United Nations Children's Fund (UNICEF)[8]

*correct 01/04/2014

1. *International Journal of Epidemiology. 2014 [cited 2014 April 01]; Available from: http://ije.oxfordjournals.org/.*
2. *Instructions to authors. 2014 [cited 2014 April 01]; Available from: http://www.oxfordjournals.org/our_journals/ije/for_authors/general.html.*

3. *PubMed. 2014 [cited 2014 April 01]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/#.*
4. *Sonnega, A., et al., Cohort Profile: the Health and Retirement Study (HRS). Int J Epidemiol, 2014.*
5. *Steptoe, A., et al., Cohort Profile: The English Longitudinal Study of Ageing. Int J Epidemiol, 2013.* ***42****(6): p. 1640-1648.*
6. *Zhao, Y., et al., Cohort Profile: The China Health and Retirement Longitudinal Study (CHARLS). Int J Epidemiol, 2014.* ***43****(1): p. 61-68.*
7. *Requejo, J., C. Victora, and J. Bryce, Data Resource Profile: Countdown to 2015: Maternal, Newborn and Child Survival. Int J Epidemiol, 2014.*
8. *Murray, C. and H. Newby, Data Resource Profile: United Nations Children's Fund (UNICEF). Int J Epidemiol, 2012.* ***41****(6): p. 1595-1601.*

Annex C: Survey Questionnaire

The questionnaire was conducted online, the questions are reproduced below:

Data Discoverability Survey

Thank you for your interest in completing this survey, which has been commissioned by Wellcome Trust to support the work of the Public Health Research Data Forum.

The aim of this survey is to enhance our understanding of the current challenges facing everyone in making research datasets 'discoverable'. A great deal of emphasis is currently being placed on the promotion of data sharing, however this is only part of the challenge. The following questions seek to explore the issues in more detail but are necessarily brief.

We have tried to design this survey so that it is easy to complete within one sitting of no more than 10 minutes although it is possible for you to save and return if you run out of time.

Thank you again for taking the time.

Background information

Please use the following sections to tell us a bit about yourself.

Main employer or employment status

- ☐ University
- ☐ Government agency
- ☐ Non-governmental organization
- ☐ Charity
- ☐ Private company
- ☐ Self-employed
- ☐ Student
- ☐ Unemployed
- ☐ Retired

(Please select the option that applies best to you)

Please indicate the regions of the world where you carry out your work

- ☐ Southern Asia
- ☐ Eastern Asia
- ☐ Europe
- ☐ South-Eastern Asia
- ☐ South America
- ☐ Eastern Africa
- ☐ Northern America
- ☐ Western Africa
- ☐ Western Asia
- ☐ Northern Africa
- ☐ Central America
- ☐ Middle Africa
- ☐ Central Asia
- ☐ Southern Africa
- ☐ Caribbean
- ☐ Oceania

(Please tick all that apply)

How would you describe your role in public health research data

- ☐ Data provider
- ☐ Data user
- ☐ Archivist / Librarian
- ☐ Funding agency
- ☐ Policy maker
- ☐ Observer
- ☐ Other

(Please select all that apply)

Other role: _____

(Please briefly describe the role)

Are you in receipt of funding from any of the following agencies

- ☐ Agency for Healthcare Research and Quality (USA)
- ☐ Bill and Melinda Gates Foundation
- ☐ Canadian Institutes of Health Research
- ☐ Centres for Disease Control and Prevention
- ☐ Deutsche Forschungsgemeinschaft (DFG)
- ☐ Doris Duke Charitable Foundation
- ☐ Economic and Social Research Council (UK)
- ☐ Health Research Council of New Zealand
- ☐ Health Resources and Services Administration (USA)
- ☐ Hewlett Foundation
- ☐ INSERM
- ☐ Medical Research Council (UK)
- ☐ National Health and Medical Research Council
- ☐ (Australia)
- ☐ National Institutes of Health (USA)
- ☐ Substance Abuse and Mental Health Services
- ☐ Administration (USA)
- ☐ Wellcome Trust
- ☐ The World Bank
- ☐ NIHR (UK)
- ☐ Other(s)

(Please tick all that apply)

Other funders (please list one per line): _____

Please indicate the forms of data that you commonly handle

- ☐ Survey
- ☐ Healthcare records
- ☐ Disease registries
- ☐ Ethnographic
- ☐ Geospatial
- ☐ Environmental
- ☐ Genomic/Proteomic/Metabolomic
- ☐ Imaging
- ☐ Physiological measurement
- ☐ Other

(Please tick all that apply)

What other form(s) of data do you commonly work with? _____

Please indicate which areas of the research data life-cycle are you actively involved in

- ☐ Conceptualisation
- ☐ Creation or receipt
- ☐ Appraisal & Selection
- ☐ Analysis
- ☐ Metadata creation
- ☐ Preservation action
- ☐ Storage
- ☐ Access, use and reuse
- ☐ Transformation
- ☐ Data Destruction
- ☐ Archive management
- ☐ Administration

Data Discoverability

What are the most important things that are needed to promote data discoverability?

Please indicate below how important you consider each aspect of discoverable data:

	Not at all	Slightly	Fairly	Extremely	Essential
be on the web	☐	☐	☐	☐	☐
be provided in a machine-readable form	☐	☐	☐	☐	☐
be provided in a non-proprietary form	☐	☐	☐	☐	☐
conform with recognised data management standards	☐	☐	☐	☐	☐
be linked to an underlying conceptual framework or ontology	☐	☐	☐	☐	☐

Preferred search options

- ☐ Keyword
- ☐ Subject terms
- ☐ Concepts
- ☐ Related concepts

(Please tick all that apply)

What aspects of a research study should ideally be easily searchable?

- ☐ Research study question
- ☐ Research study protocol
- ☐ Research data management plan
- ☐ Consent form and associated information pack
- ☐ Funding details
- ☐ Data collection instrument designs
- ☐ Variables
- ☐ Code lists
- ☐ Concepts
- ☐ Research publications

(Please tick all that apply)

Is there any other aspect of data discoverability that you feel is important?

_____ (Please provide any other observation or concern)

How could discoverability of public health data be improved and do you see any immediate priorities?

_____

The following is a list of some repository services and classes of repositories that exist, mostly taken from Nature's website. The web-link gives examples of repositories for many of the categories below. Please indicate the repositories that you have used (data access, deposit or both) in the past or anticipate using in future.

	Already used	Intend to use
Dryad	☐	☐
Figshare	☐	☐
ClinicalTrials.gov	☐	☐
Social Science e.g. ICPSR / UK Data Archive	☐	☐
DNA protein sequences	☐	☐
Genetic association & genome variation	☐	☐
Functional genomics	☐	☐
Proteomics	☐	☐
Molecular interactions	☐	☐
Molecular structure	☐	☐
Taxonomy & species diversity	☐	☐
Organism or disease specific resources	☐	☐
Environmental & geoscience	☐	☐
Other	☐	☐

You selected "Other" in the list of repositories above. Please provide details.

_____

Controlled Vocabularies and Thesauri

Standardised taxonomies are commonly used approaches to enhance data discovery and facilitate comparison across datasets.

Please indicate which of the following terminologies, classifications, thesauri or metathesauri you are familiar with

- ☐ SNOMED CT
- ☐ OPCS-4
- ☐ International Classification of Disease (ICD)
- ☐ Logical Observation Identifiers Names and Codes (LOINC)
- ☐ Diagnostic and Statistical Manual of Mental Disorders (DSM 5)
- ☐ Read Codes
- ☐ Medical Subject Headings (MeSH)
- ☐ Humanities and Social Sciences Electronic
- ☐ Thesaurus (HASSET)
- ☐ European Language Social Science Thesaurus (ELSST)
- ☐ Unified Medical Language Service (UMLS)
- ☐ Other

(Please tick all that apply)

If your answer to the previous question included "Other" please specify

_____

Which tools do you use to assist with the management of controlled vocabularies? For example, UMLS, Library of Congress, WHO Global Health Observatory indicator registry

_____

Data documentation

In order that researchers can reuse research data meaningfully it is important to ensure that data are provided with detailed descriptors, typically this has been in the form of a code book plus ancillary documentation that describes the processes associated with data collection and any processing that has been carried out.

The following is a list of standards to assist with data documentation. Please indicate which of these you have experience of knowingly using.

- ☐ DC
- ☐ SDMX
- ☐ EAD
- ☐ METS
- ☐ DCAT
- ☐ CKAN
- ☐ eGMS
- ☐ INSPIRE
- ☐ ADMS
- ☐ DDI 2/3

(Please tick all that apply)

Which tools do you use to assist you with data documentation? _____

Which are the key challenges in creating/using documenting data? _____

Data Citation and Data Publications

Citation of data is becoming an important tool to promote and track data reuse. Data publication offers a mechanism to promote data citation. The following section explores your views and understanding of these options.

It is possible to publish articles that describe research datasets independently of conventional research publications. Are you familiar with this form of data publication?

- ☐ Yes
- ☐ No

Where did you first hear about data publications?

- ☐ Colleague
- ☐ Journal
- ☐ Conference/workshop
- ☐ Search engine
- ☐ Other

(Please tick all that apply)

If your answer to the previous question included 'other' please specify.

_____

Please indicate which benefit(s) of data citation are most important to you

- ☐ Easier for readers to locate data
- ☐ Proper credit given to data contributors
- ☐ Links between datasets and associated methodology publication provide context for reader
- ☐ Links between datasets and publications describing their use can demonstrate impact.
- ☐ Infrastructure can support long-term reference and reuse
- ☐ Less danger of data plagiarism
- ☐ Promotes professional recognition and rewards
- ☐ Other

(Please tick all that apply)

What other benefits do you see in data citation? _____

How granular should data citations be:

- ☐ Dataset collections
- ☐ Single datasets (or sweep)
- ☐ Files within datasets
- ☐ Individual items of data
- ☐ Other

(Please tick all that apply)

Since you selected other above, please give your suggestion here: _____

Ideally, how should longitudinal and regularly changing datasets be handled?

- ☐ New identifier assigned at each update
- ☐ Publish revisions at regular intervals
- ☐ Time series data should be published as complete 'snapshots'
- ☐ Time series data should be published in instalments
- ☐ All published versions of the datasets must be stored
- ☐ Other

(Please tick all that apply)

Other mechanism suggested for handling longitudinal and changing datasets:

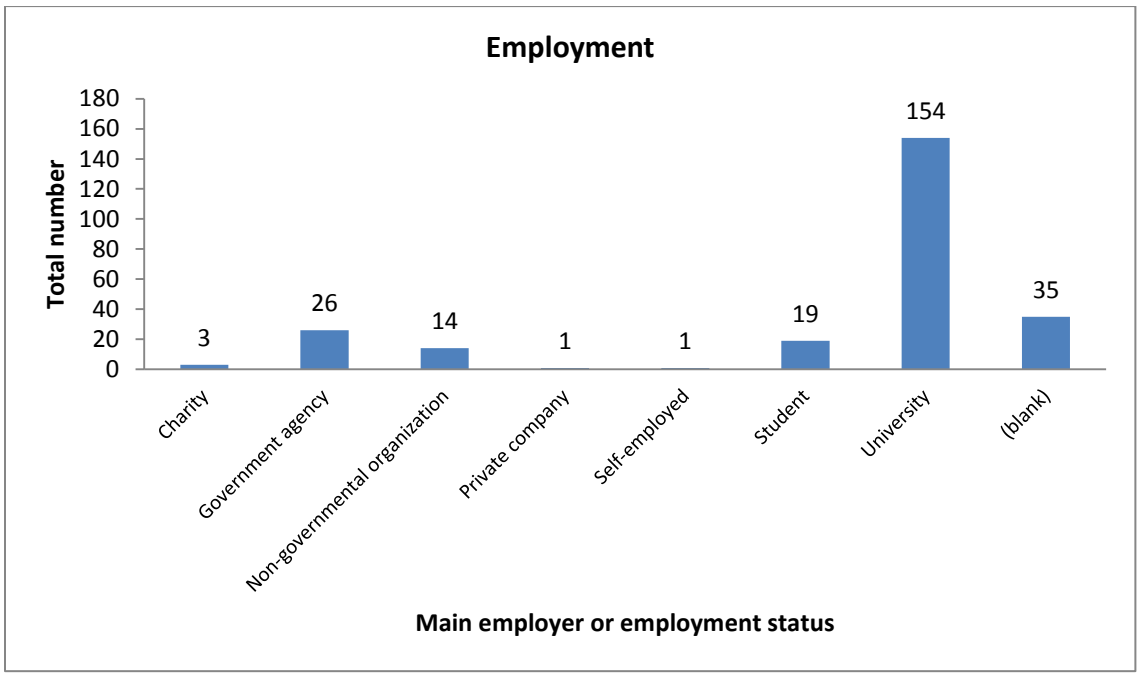
_____

Which would you say, if any, are the key challenges affecting the widespread adoption of data publications? _____

Annex D: Results of Online Survey

The following graphs and tables present the findings of the online survey. Note that multiple selections were allowed for each respondent.

Employment of respondents



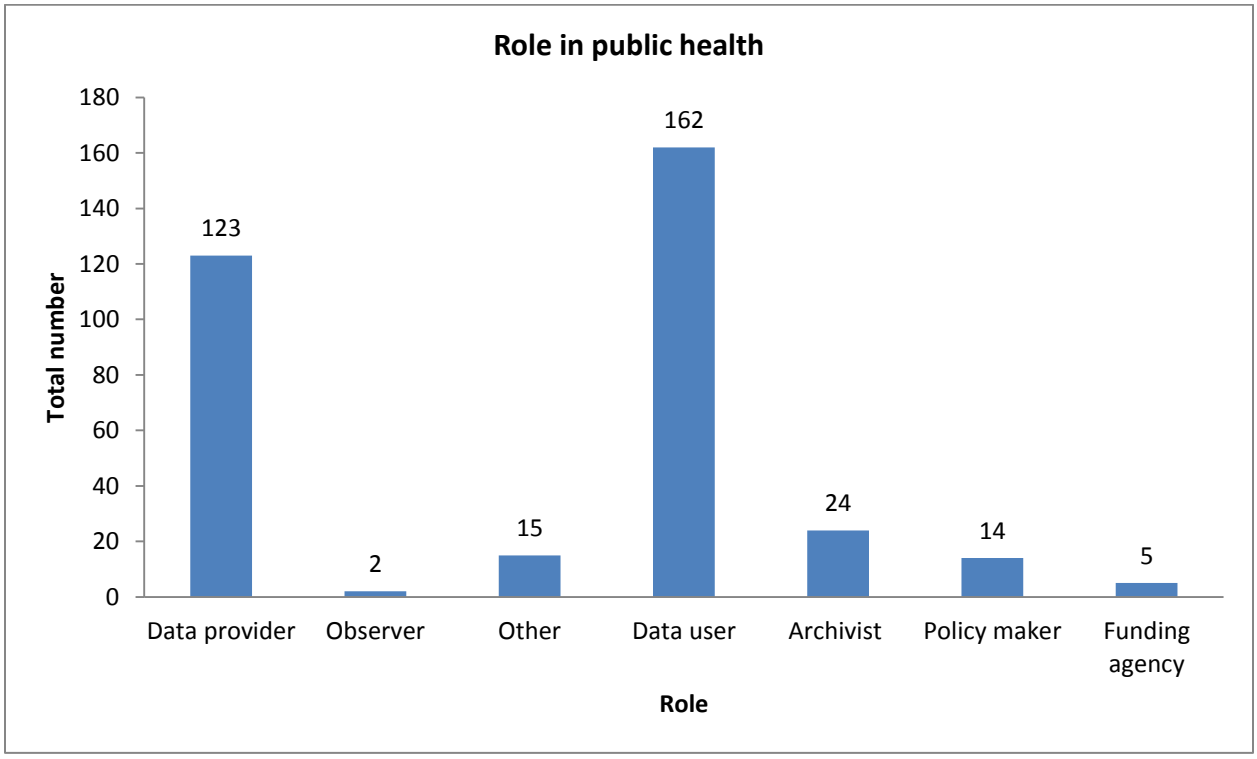
218 respondents

Geographical regions of respondents

Southern Asia	20
Eastern Asia	12
Europe	108
South-Eastern Asia	18
South America	11
Eastern Africa	30
Northern America	31
Western Africa	28
Western Asia	6
Northern Africa	9
Central America	5
Middle Africa	9
Central Asia	5
Southern Africa	29
Caribbean	4
Oceania	49

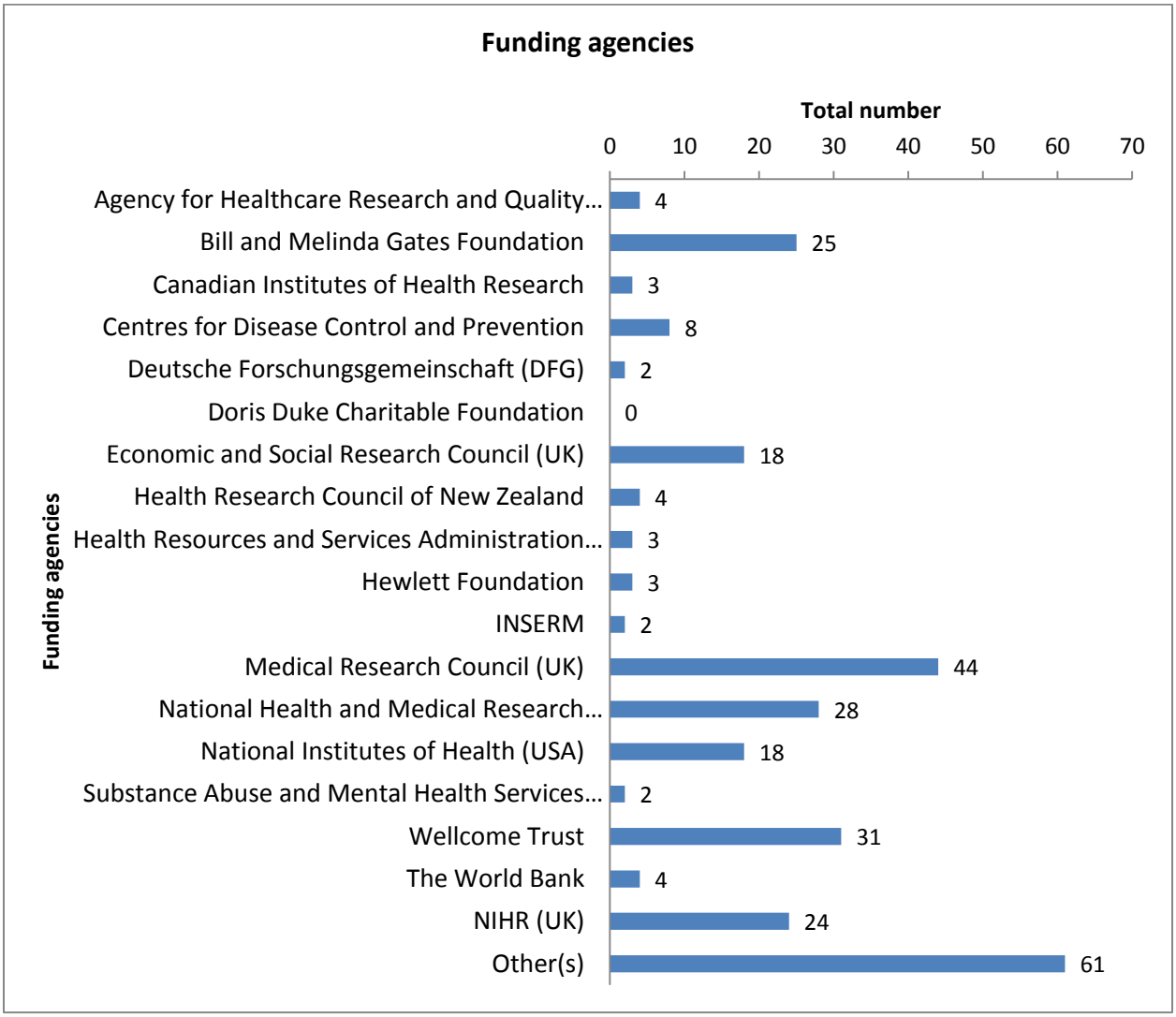
245 Respondents

Role in public health



213 Respondents

Most common funders



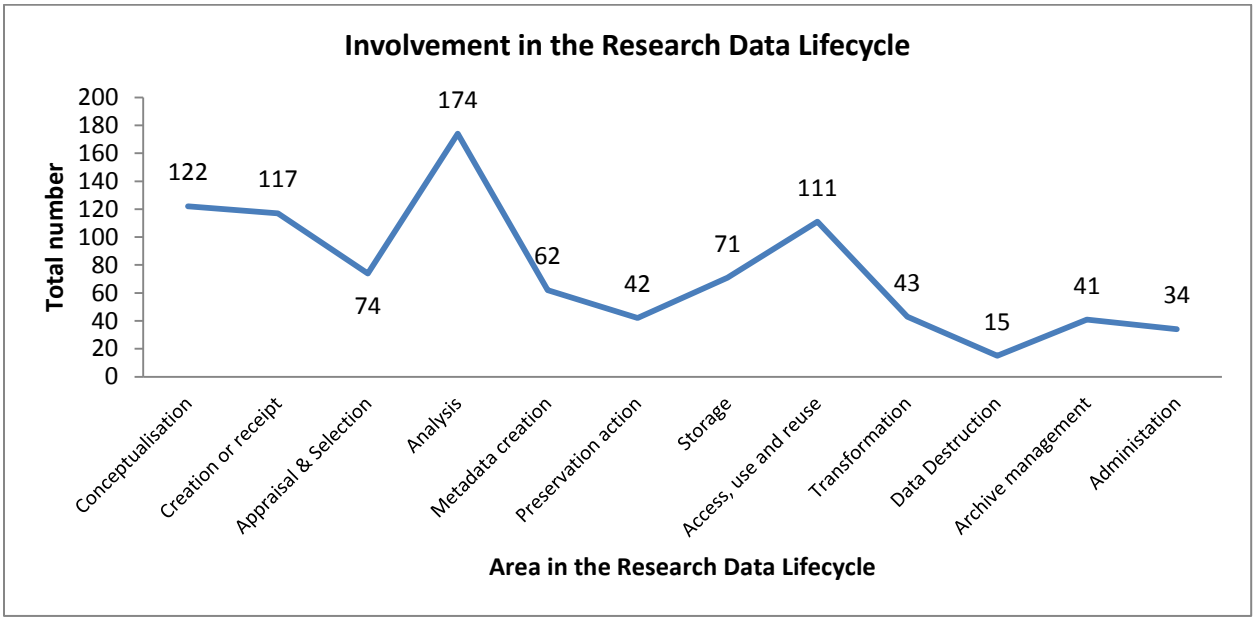
175 Respondents

Forms of data

Survey	157
Healthcare records	125
Disease registries	76
Ethnographic	24
Geospatial	46
Environmental	31
Genomic/Proteomic/Metabolomic	30
Imaging	19
Physiological measurement	47
Other	37

211 respondents

Involvement in the data lifecycle



214 respondents

How best to improve data discoverability – importance

Aspect	Opinion						
	Essential	Extremely	Fairly	Slightly	Not at all	(blank)	Total number
be on the web	83	69	31	7	9	54	253
be provided in a machine-readable format	94	74	19	3	6	57	253
be provided in a non-proprietary form	44	61	64	5	14	65	253
conform with recognised data management standards	55	82	50	1	5	60	253
be linked to an underlying conceptual framework or ontology	20	45	78	14	34	62	253

201 Respondents

Preferred search techniques

Concepts	66
Related concepts	33
Subject terms	133
Keyword	181

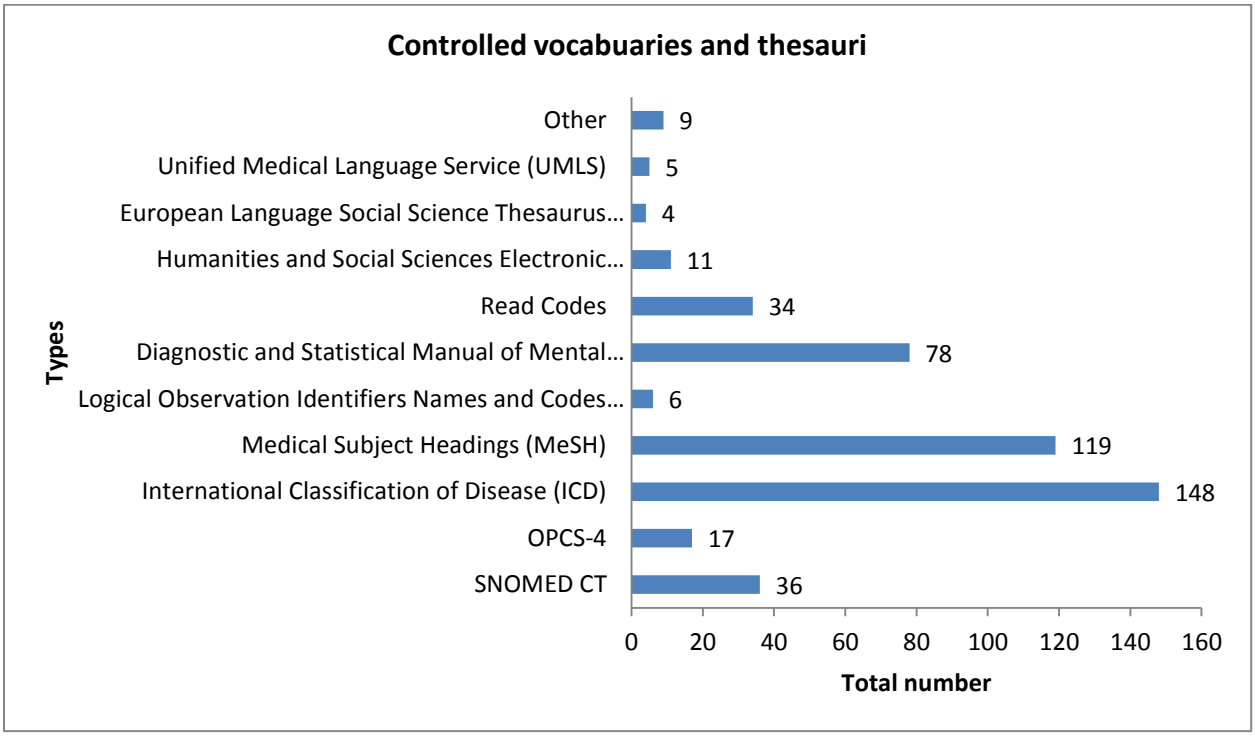
207 Respondents

Use of repositories

Repositories	Use (total number of people)				
	Already used		Intended to use		
	Number of respondents	% of respondents	Number of respondents	% of respondents	(blank)
Dryad	2	0.79	6	2.37	0
Figshare	5	1.98	9	3.56	0
ClinicalTrials.gov	58	22.92	18	7.11	0
Social Science	47	18.58	1	0.40	0
DNA protein sequences	6	2.37	2	0.79	0
Genetic association & genome variation	17	6.72	5	1.98	0
Functional genomics	8	3.16	4	1.58	0
Proteomics	1	0.40	5	1.98	0
Molecular interactions	4	1.58	2	0.79	0
Molecular structure	3	1.19	2	0.79	0
Taxonomy & species diversity	3	1.19	5	1.98	0
Organism or disease specific resources	13	5.14	7	2.77	0
Environmental & geoscience	16	6.32	9	3.56	0
Other	20	7.91	6	2.37	0
Total	203	80.24	81	32.02	0

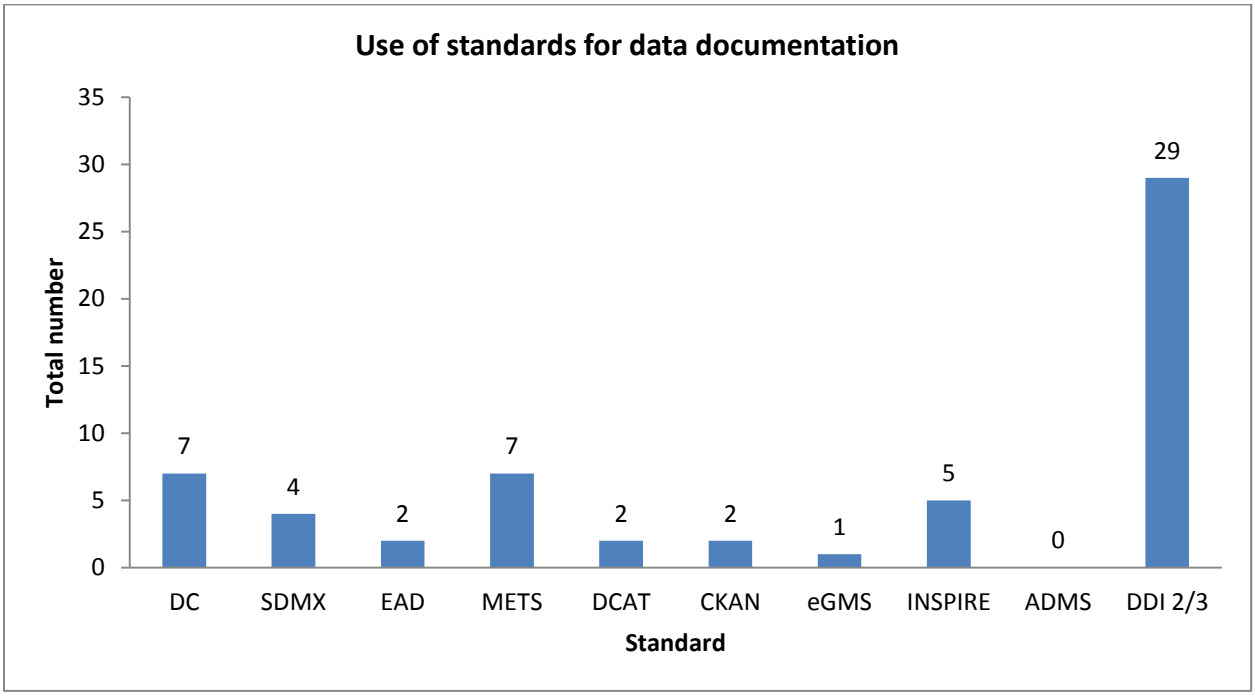
138 Respondents

Controlled vocabularies and thesauri



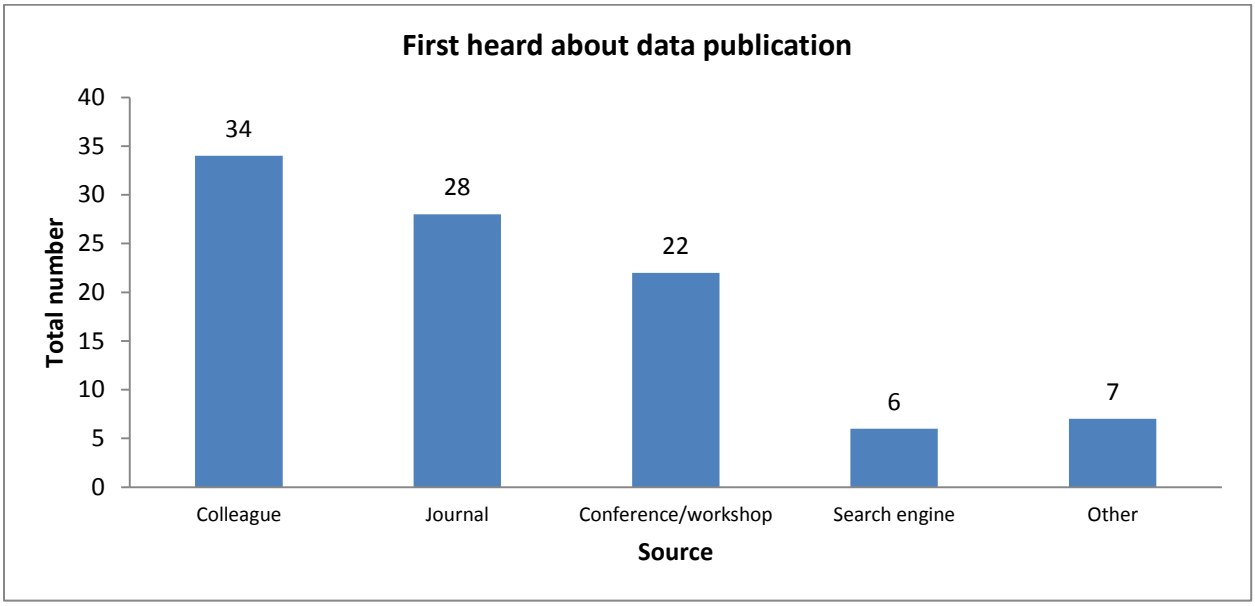
163 Respondents

Data documentation – use of standards



43 Respondents

First heard about data journals



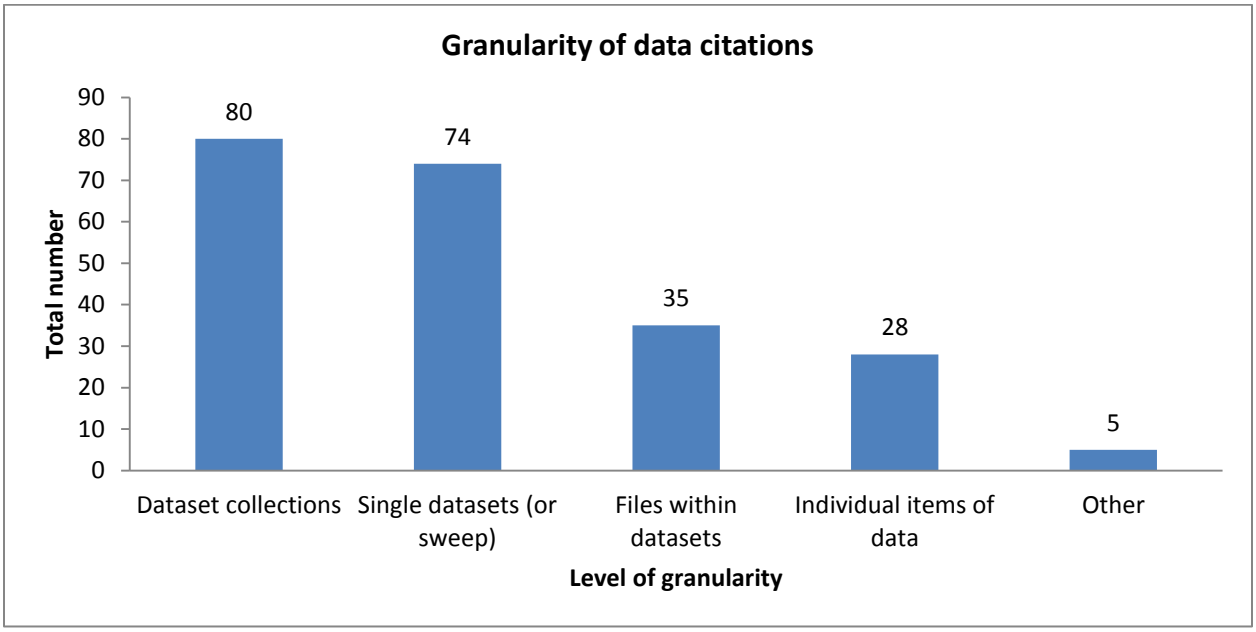
79 Respondents

Benefits of data citation

Easier for readers to locate data	137
Proper credit given to data contributors	113
Links between datasets and associated methodology publication provide context for reader	114
Links between datasets and publications describing their use can demonstrate impact	77
Infrastructure can support long-term reference and reuse	69
Less danger of data plagiarism	40
Promotes professional recognition and rewards	64
Other	4

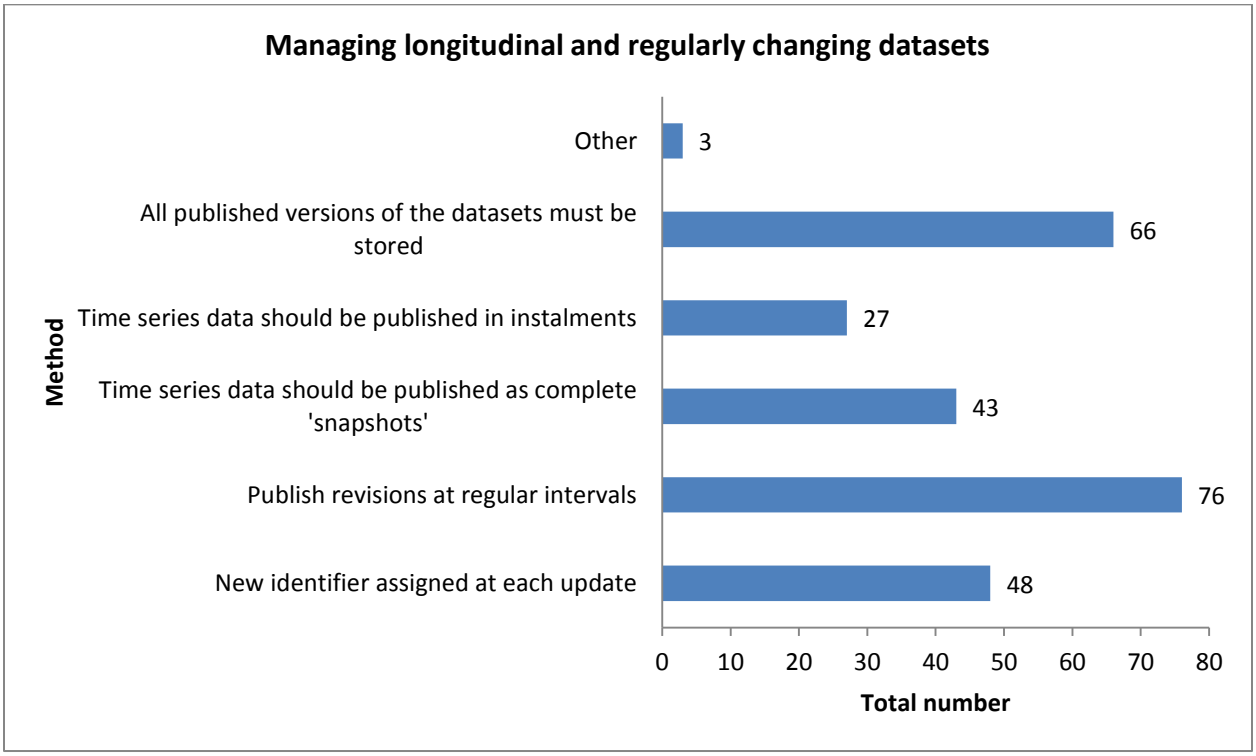
160 Respondents

Needed granularity of data citation



134 Respondents

Mechanisms for managing longitudinal and regularly changing data sets



140 Respondents

Annex E: Topic Guidelines for Focus Groups

Overview

- Introductions
- Experience with data
  - Discovering data
  - Using data
- Technology and standards
  - Data portals
  - Data journals
  - Web of Linked Data
  - Standards

Introductions

- Who are you?
- What is your institutional affiliation?
- Where are you in your professional career?
- What is the focus of your research?
- What publications are the most important to you professionally?
- What data sets do you work with most often?

Experience with Data

- How do you discover good data?
- What is “good data”?
- What information is most important when you are looking for data?
- Are you satisfied with the quality of available data sets?
- What issues do you have with the quality of the data available to you?
- What information is most important when working with data?
- Are you satisfied with the quality of the documentation/metadata available to you?
- What are the main issues you face when working with data?

Technology and Standards – Examples Provided for Discussion

- Data Portals
- Data journals
- Semantic Web/Web of Linked Data
- Standards for data and metadata

Annex F: Project Data Journal – Progress to Date

The 'Exemplar Public Health Datasets' collection (http://openhealthdata.metajnl.com/collections/special/exemplar-public-health-datasets) will launch in late-summer 2014 and feature invited papers describing datasets that are outstanding in the field of public health. Published in the journal Open Health Data, the collection will launch with five papers in the first instance, with 10-15 more expected thereafter. We originally invited roughly 150 authors and received 22 positive responses, resulting in five submissions thus far and the remainder will be submitted over the next six months. For such a new concept, this response rate can be considered quite high, especially given the short deadline we set authors.

Whilst we are unable to report the exact subjects of the submitted papers (as they are currently being peer-reviewed), the majority describe data from longitudinal/cohort studies from a mixture of publicly available data and data with accessibility criteria. This in itself created a few problems because the original article submission template was designed for single, openly accessible datasets, rather than studies with numerous files and conditions for access. Consequently, the template was revised to incorporate longitudinal studies and studies with access criteria.

Data journals have a number of purposes: facilitating discoverability, crediting data producers, allowing data to be cited, and so on. However, until this particular collection we had not appreciated the potential for data journals to formalise the various access criteria associated with different studies. As such, the collection will showcase data for which, until now, the accessibility criteria were not readily available in a centralised location. We see this as a positive step and another selling point of the data journal model.

Finally, we expect to publish papers where the data itself is not publicly available, due to consent restrictions, but where additional materials surrounding the data collection have been uploaded in an openly accessible fashion, such as consent forms, protocols, data management plans and data access policies, all of which will benefit by wider circulation. This helps broaden the scope of the data journal and makes it more than just a tool for sharing data.

While the uptake of the collection could have been higher, the uptake does reflect the current research culture of data sharing. As far as the ordinary researcher is concerned, there are perceived disincentives for sharing one's data that outweigh the perceived benefits. Collections such as 'Exemplar Public Health Datasets' will help to change this by demonstrating that data journals are being taken up by senior scholars to showcase their well-known studies.