Public Health Research Data Forum

Enhancing Discoverability of Public Health and Epidemiology Research Data

July 2014

wellcome trust

Acknowledgements

This report is the work of a project team, whose membership is as follows:

Tito Castillo: Honorary Senior Research Associate, University College London and Cambridge University

Arofan Gregory: Open Data Foundation

Samuel Moore and Brian Hole: Ubiquity Press, London

Christiana McMahon and Spiros Denaxas: Clinical Epidemiology, Farr Institute at UCL Partners, University College London,

Veerle Van Den Eynden, Hervé L'Hours, Lucy Bell, Jack Kneeshaw and Matthew Woollard: UK Data Archive, University of Essex, Colchester

Chifundo Kanjala, Gareth Knight and Basia Zaba: London School of Hygiene and Tropical Medicine

The project team would like to take this opportunity to acknowledge the assistance of the members of the Public Health Research Data Forum for commissioning this work and their helpful comments and suggestions. We would also like to thank all of the contributors to focus groups sessions, survey respondents and contributors to the special collection data journal, without whose help the report would not have been possible. Additionally, we are grateful to Rich Hutchinson, the epiLab Service Manager, for hosting the data collection service at UCL and Paul Seabright, at Cambridge Enterprise, for assisting with the contractual process. Finally, we would like to thank David Carr and Jane Simmonds at the Wellcome Trust for their patience and encouragement throughout this process.

Contents

Executive Summary

The project "Enhancing Discoverability of Public Health and Epidemiology Research Data" was commissioned by the Wellcome Trust on behalf of the Public Health Research Data Forum. The work focused on assessing the discovery and use of major data sets in the public health and epidemiology research domain. Further, it aimed to identify relevant models which could be used to enhance data discoverability and re-use, and to explore the feasibility of these models.

The project was international in scope, and analyzed best practice not only within the public health and epidemiology research domain, but also in related, data-intensive research domains. It sought to investigate the perspectives of our major groups of stakeholders: researchers and secondary users of data; data producers; data archives, libraries, and other data disseminators; and funding agencies.

The study was conducted using several investigative techniques: a review of significant data sets within the public health and epidemiology research domain; an online survey; focus groups with researchers; and an assessment of relevant models for improving data discovery and supporting re-use, within the public health and epidemiology research domain, and in similar domains.

Key findings

Our findings suggest that the public health and epidemiology research domain could enhance data discoverability, access, and re-use by adopting best practice as it exists in some other data-intensive research domains (social and behavioural sciences, economics research). Existing practices around data management, support for researchers, data archiving, and documentation are extremely varied across the field. The establishment of best practices and adoption of standards would enable significant enhancement of infrastructure related to data discovery and re-use.

Three dominant models for enhancing data discovery were identified, based on the input gathered in the focus groups and the online survey, and on the examination of practice for significant public health and epidemiology research data sets:

1. *The Centralized Portal Model* – This model has a domain-focused catalogue of all available data, well-documented to the variable level, so that researchers know what data exist and are of interest before applying for access.
2. *The Data Journal Model* – This model uses peer-reviewed open-access journals which focus on data articles: descriptions of high-value data sets which are useful for research, and link to the place where the data are disseminated.
3. *The Linked Data Model* – A decentralized approach based on the machine-searchable inter-linking of data and documentation published on the web, using current standards from W3C.

The Centralized Portal Model was the preferred approach among researchers. This is also a model which requires a high degree of coordinated infrastructure across organizational boundaries, both for the cataloguing of data sets and for the reliable archiving of data. The production of standard, rich metadata on the part of data producers or archives is required. This is a relatively expensive model, but was clearly the most useful and intuitive model from the researchers' perspective. The technology for implementing this model is mature, and has been in production and use for more than a decade.

The Data Journal Model was also seen as very useful by researchers. Peer-reviewed, citable publication is a model which researchers understand. When combined with good, standard documentation about

the data sets described in data articles, this could be a very attractive model. This model presents us with a requirement for good archiving infrastructure for data sets, and a standard mechanism for their citation. It is perhaps less resource-intensive than the Centralized Portal Model, but it is still fairly demanding.

The Linked Data Model was perceived as less useful by researchers, in part because it relies on the creation of client applications, operating on the "smart" linkages published on the web by the disseminators and users of the data. These do not exist today in a sufficient form for us to be confident that this approach will provide the optimal result. However, this technology is increasingly being used in other domains, and may become more important in future. It also requires rich metadata published in a standard form. It is difficult to estimate required resources, because the costs – like the technology itself – are not applied in a centralized fashion.

It is important to note that these approaches are complementary, and not mutually exclusive. In other domains, they are often employed together by a single organization such as an archive, to optimize the discoverability of the data sets they disseminate.

As a long-term goal, all three approaches might be considered in combination. This is not likely to be feasible in the short to medium term.

Recommendations

(1) Focus on the creation of a centralized domain portal for public health and epidemiology research, taking the following steps:

   (A) Develop a search portal, with an interface similar to the examples described (such as the CESSDA and UK Data Service portals) with a mechanism for harvesting metadata exposed by data producers and archives.

   (B) Identify technical standards and protocols based on the DDI standard and an analysis of the various harvesting protocols such as the OAI-PMH protocol used by CESSDA (and others), and the DwB WP 12 Prototype. Other networks (such as the MRC Gateway and the INDEPTH Network) should also be considered.

   (C) Establish guidelines and best practices for the use of technical standards and protocols for exposing data holdings to the domain portal.

   (D) Establish best practices and guidelines for archiving data holdings, based on any of the archival best practices found in the public health and epidemiology domain, the behavioural and social sciences, and the economics domain. Engage with existing archival infrastructure where possible, rather than trying to create wholly new archives, and provide support for researchers looking for secondary data to use following existing good practice.

   (E) Develop tools and guidelines for researchers where required to encourage good practices around data management and documentation. Tools should be DDI-based, so that data can easily be exposed to the centralized portal and archived.

   (F) Create incentives for research projects to follow established best practice for data management, documentation, archiving, and sharing. Funders must recognize that these activities do require additional resources on the part of research projects which produce data.

(2) Encourage the use of data journals and further publication of data articles in the public health and epidemiology research domain. Archival practices established for the centralized portal should

include dissemination of data sets which are citable, to allow for easy linking into the same data sets catalogued in the portal. A standard such as DataCite might be considered here. Also, standards and best practices for data documentation should be established (the DDI documentation used by the centralized portal could be re-used for this purpose, or a direct link to the portal could be used from the data article).

(3) Continue to monitor the potential of the Web of Linked Data regarding public health and epidemiology research data. The data journals, the archives, and the centralized portal might wish to leverage this technology approach in the medium term, so agreed ontologies (based on the DDI ontologies and other data-related ones) should be established and promoted.

1. Introduction and Overview

This report is the result of a project commissioned by the Wellcome Trust on behalf of the Public Health Research Data Forum[1], examining the ways in which the discoverability of data could be enhanced within the domain of public health and epidemiology research. This summary report briefly describes the work undertaken by the project team. The findings of the project are presented, and future activities moving forward are proposed.

A. Discoverability of Public Health Research Data: The Challenge

The infrastructure and support for discovering, accessing, and using research data for public health and epidemiology does not compare favourably to that found in some other domains, such as genetics, behavioural sciences, environmental science, and social science. For example, in genetics research, well-established data sharing infrastructures exist (Genbank, EMBL), complemented by international data sharing agreements (1996 Bermuda Principles[2], 2003 Fort Lauderdale agreement[3]) and journal data policies that mandate data deposit before publication. For the social sciences and environmental sciences in the UK, ESRC and NERC mandate via their respective data policies the deposit of research data in data centres they fund; these data centres provide central points of discovery for data[4,5].

The public health and epidemiology research community has focused on the collection of data for specific research purposes, and has not focused as much on ensuring that the data collected through funded projects is made available for secondary use by researchers coming from outside the project. While in some cases there is support for finding and understanding data which might be useful for secondary use, in many other cases there is little or no support. Researchers looking for existing data may need to use personal contacts to access the data, or the amount of information available about a data set of interest is insufficient or non-optimal.

Since the research community does realize the potential of public health research data for future and integrated research, various initiatives to make public health data accessible and discoverable exist, but are often disconnected and isolated.

Often, it is the principal investigator of a project, and other project researchers, who are responsible for documenting and disseminating data collected for their own project to those who wish to re-use it. This is not the primary task of any of the project researchers, and the data may not be managed, documented, or archived sufficiently to be suitable for easy re-use. Data may be made accessible by individual studies, through their own collaboration agreements (see, for example, the data access policy for the ALSPAC study[6]), but a centralized point of discovery does not exist.

---

[1] For more information on the Public Health Research Data Forum, see http://www.wellcome.ac.uk/publichealthdata

[2] Marshall,E. Bermuda Rules: Community Spirit, With Teeth. Science *2001;*291(5507):1192. http://www.sciencemag.org/content/291/5507/1192.full

[3] The Wellcome Trust. Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility. The Wellcome Trust; 2003. http://www.sanger.ac.uk/datasharing/assets/fortlauderdalereport.pdf

[4] http://data-search.nerc.ac.uk/

[5] http://discover.dataservice.ac.uk

[6] http://www.bristol.ac.uk/alspac/researchers/data-access/

In other cases, data is catalogued in centralized portals, which offer rich metadata descriptions and practise good data management, so that researchers can see exactly which variables exist within a dataset, and can be confident that if they apply for access, they will get the persistently-identified version of the data set which they expect. Some of this data is held in national archives (such as the UK Data Service[7] or the Interuniversity Consortium for Political and Social Research[8]), and sometimes in portals with a medical research focus (e.g., the Medical Research Council Gateway project[9] in the UK).

Often, it is the principal investigator of a project, and other project researchers, who are responsible for documenting and disseminating data collected for their own project to those who wish to re-use it. This is not the primary task of any of the project researchers, and the data may not be managed, documented, or archived sufficiently to be suitable for easy re-use.

In other cases, data is catalogued in centralized portals, which offer rich metadata descriptions and practice good data management, so that researchers can see exactly which variables exist within a dataset, and can be confident that if they apply for access, they will get the version of the data set which they expect. Some of this data is held in national archives (such as the UK Data Service), and sometimes in portals with a medical research focus (e.g. the Medical Research Council Gateway).

If we look at the best examples of models for discovery in domains which are similar to public health and epidemiology, it becomes clear that improvements are possible. The challenge is to identify the correct actions to take, and to make sure that they are feasible and affordable within the research culture and funding structure of the domain.[10]

B. Existing Technology Approaches and Standards

When issues and approaches around data discoverability and re-use are assessed with a broad view – that is, across various data-intensive domains which are in some way similar to public health and epidemiology – we find different approaches in terms of how information technology is employed. We also find that there are many potentially relevant standards which can be used to achieve large-scale solutions involving many organizations, and which can be established across national borders.

Within public health and epidemiology research today, we find that there is no single, dominant approach. Other domains such as social science research show that coordinated approaches, once identified, can be applied in large-scale solutions across entire regions, Europe perhaps providing the best example of this in the CESSDA network of data archives[11]. Using the correct standards for describing and citing data sets is key to supporting large-scale data discoverability, and this is true across almost all of the successful technology approaches. Technology infrastructure for data discovery and re-use must be coordinated, however, and the public health and epidemiology research domain does not have as mature an infrastructure, in this regard, as that seen in some other domains.

Typically, modern approaches to the implementation of information technology rely on agreed models and standards for describing relevant resources.[12] There are many different standards which are used in

[7] http://www.ukdataservice.ac.uk
[8] http://www.icpsr.umich.edu/
[9] https://www.datagateway.mrc.ac.uk/
[10] For an interesting perspective on data sharing, see: Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L et al. Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 2011;6(6):e21101.
[11] http://www.cessda.net/
[12] Kush R, Goldman M. Fostering Responsible Data Sharing through Standards. New England Journal of Medicine 2014;370:23.

the public health and epidemiology research domain, and in related domains (**Box 1**). These are not always used to support data discoverability, but that is often one of their functions.

Standards are, however, fundamental to data discovery, because a consistent description of the data allows for modern information technology to be effectively applied to the problem. Many different standards were considered here, both those specific to data management and discovery, and the more generic standards which are being used for health research applications.

Box 1 – Examples of relevant standards

1. The Data Documentation Initiative (DDI)[13]
2. ISO 17369 – The Statistical Data and Metadata Exchange Initiative (SDMX)[14]
3. Clinical Data Interchange Standards Consortium (CDISC)[15]
4. Dublin Core
5. Health Level Seven (HL7)
6. The BRDIG group (HL7, CDISC and others)[16]
7. Various archival and library standards (EAD, METS, etc.)
8. ISO 11179 – Metadata Repositories
9. CaBIG/CaCore (an implementation of ISO 11179)[17]
10. The Resource Description Framework family of standards from W3C for the Web of Linked Data/Semantic Web
11. DataCite (an implementation of the Digital Object Identifier [DOI] standard for referencing data sets)[18]

In many cases, these standards are not focused on data discovery, or even data management, but are primarily used for exchanging data between different applications or systems. This may not seem to be a major distinction, but in fact it proved to be significant: data discovery is a metadata-intensive activity, requiring a high degree of detail, so that the data can be understood by those considering it for their research. This information is not necessarily relevant for simple data interchange, or other types of use. The standards designed for data management seemed to have more in common with standards designed explicitly for discovery. However, data discovery is an exacting form of resource discovery, and generic standards such as Dublin Core – which are very popular for discovery of many types of information – are insufficient when data is being sought. Similarly, the generic standards used by many archives and libraries were not specific enough regarding data and often did not seem a good fit for our analysis.

The set of standards considered was looked at from a perspective which was not narrowly focused on discovery – indeed, the standards which seemed to be used most often for discovery were in some cases designed to support not only discovery but also other uses, typically those related to data management. Standards such as DataCite (DOIs) designed for data citation proved to be quite important, given that

[13] http://www.ddialliance.org
[14] http://www.sdmx.org
[15] Notably, the recent CDISC annual report emphasized data sharing: http://www.cdisc.org/cdisc-annual-report
[16] http://www.bridgmodel.org/
[17] Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, Fragoso G, Coronado SD, Reeves DM, Hadfield JB, Ludet C, Covitz PA, caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability. J Biomed Inform 2008;41(1):106-23.
[18] http://www.datacite.org/

citation can be understood as a sub-set of the discovery function, leading from research to the data upon which it is based.

This report does not contain a description of how each standard was assessed, but it will be clear from the various models used to support data discovery which standards are most popular, and commonly used.

C. Incentives for Data Sharing and the Public Health Research Data Forum

Clearly, the research culture within a domain must be taken into account when large-scale infrastructure developments are to be undertaken. Solutions might make sense from the perspective of the application of information technology, but might not be feasible within a particular research culture for non-technological reasons.

When we look at other domains which do have good infrastructure for data discovery and re-use, such as genetics or environmental sciences, we see that cultural changes have been implemented over time, and involve not only the researchers, but also other players in the overall picture, notably data archives, data libraries, and the funders of research. Examples of such infrastructure typically include catalogues or portals for discovering data, and detailed documentation of the data holdings of archives available on the websites of those disseminating data. This type of infrastructure is typically based on standard models for metadata and exchange protocols within the domain. This type of infrastructure takes time to socialize among the members of a community. The social sciences are a good example of this type of infrastructure within a research community.

Cultural change is not possible without incentivizing the researchers, however, as other disseminators of data will be more naturally focused on re-use and data sharing (e.g. archives and data libraries). The UK Expert Advisory Group on Data Access (EADGA) produced a report in May 2014 on incentives for data sharing indicating that data management funding is needed, as well as recognition for data sharing in research careers[19]. The Public Health Research Data Forum brings together many of the major global funders of research in the public health and epidemiology research domain[20]. Enhancing the discoverability of research data has emerged as a key challenge for Forum partners and other research funders, and several funders have initiated new initiatives in this area. For example, the US National Institutes of Health (NIH) is working towards the establishment of an NIH Data Discovery Index as part of its Big Data to Knowledge Initiative[21].

Funders are in a powerful position to incentivize researchers, and other funded organizations and projects, to address cultural change within the domain on the part of researchers, and to establish and influence infrastructure for data discoverability and re-use for other funded organizations.

D. Project Scope

The scope of this project was broad – although the data sets analyzed were ones used by researchers in public health and epidemiology, they were not always data produced by specifically public-health- or

[19] Expert Advisory Group on Data Access. Establishing Incentives and Changing Cultures to Support Data Access. The Wellcome Trust; 2014. http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/EAGDA/WTP056496.htm.

[20] The work of the Forum is underpinned by a joint statement of purpose – see: Walport M, Brest P. Sharing research data to improve public health. Lancet 2011;377(9765):537-539.

[21] See Big Data To Knowledge initiative: http://bd2k.nih.gov/index.html#sthash.RqZW0Izt.dpbs

epidemiology-related projects. The examination of best practices around data discoverability looked at other domains where data is heavily used and shared, notably in the areas of social science research, economics, behavioural science, and official statistics. These models were considered not only from the perspective of the data involved, but also in terms of their use of information technology, and in terms of the organizational culture in the domains where they are used. The project was global in scope – we did not confine ourselves to the UK or Europe, but tried to include those researchers working also in developing countries, North America, Australia, and elsewhere.

E. Objectives

The overarching goals of this project were threefold:

1. To examine the current discoverability of public health and epidemiology data sets, and determine any barriers to access.
2. To examine current models for data discoverability such as archives, data portals/catalogues, and other systems to facilitate data discoverability, and to determine which are relevant to public health and epidemiology data.
3. To identify possible models for funders which would enhance the discoverability of, and access to, public health and epidemiology data, and to determine their feasibility and resource requirements.

F. Stakeholders

Four categories of stakeholders were identified:

1. ***Researchers and users of existing data*** – all types of researchers working in the public sphere, including academic researchers and those working for the government to support policy.
2. ***Data producers*** – collectors of data suited to secondary use, including research projects and statistics collected by government agencies; includes survey data but also data coming from clinical systems and administrative registers, and so forth.
3. ***Data archives and data libraries, and other disseminators of data*** – any organization or project which holds and disseminates data to researchers for secondary use, including long-term research projects.
4. ***Funders*** – all types of funders providing money for research projects and related activities, but with a focus on those which are either charitable or using public funds.

These categories are not mutually exclusive – the producers of data are often researchers themselves, and many data producers also act as disseminators of data. However, for the purposes of conducting a survey and focus groups, and for guiding other discussions, it is generally possible to place any given individual clearly into one of these categories. It was felt that coverage of all four areas was important, and could be assessed from the self-description supplied by survey respondents, and through discussions with others consulted during the project.

The list of identified stakeholders for the data sets analyzed in this project is given in the annex for Work Package 1, along with a listing of the data sets themselves.

G. Methodology

This section provides brief descriptions of each of the work packages undertaken as part of the project. Where noted, annexes with the relevant detail concerning specific activities are also provided.

WP 1 – Data Holdings

This work package provided the context for the entire project, first by identifying significant stakeholder groups and contact individuals for each of them, and then by performing an assessment of the data sets produced and disseminated by the relevant groups of stakeholders – data producers and archives/data libraries.

Through consultations with Forum partners and web research, the project team identified 49 significant data sets, which were reviewed against an agreed set of criteria. The spreadsheet can be found in Annex A. Following this activity, 13 of the 49 data sets were assessed regarding the state of their discoverability documentation – the results of this activity are also provided in the annex. As the work was on-going, it was discovered that the Journal of Epidemiology has published a series of what it terms "Data Profiles" – this was an interesting type of article to find, and an explanatory description is found in Annex B. The contents of these Data Profiles were subsumed in the on-going work.

WP 2 – Online Survey

An online survey was conducted – a copy of which is appended at Annex C. Respondents were self-selecting with invitations sent out to the stakeholder communities identified by the project team, including several mailing lists for public health and epidemiology research, the data archival community, and individuals known to the members of the project team. All questions were optional, and many questions were open-ended, with the idea of getting the most accurate input regarding questions where we could not necessarily anticipate all possible answers. A total of 253 responses were received, with a global spread of respondents acting in different stakeholder capacities.

The online survey was conducted using REDCap, a popular software package which is provided free of charge to not-for-profit use. They require the following citation for projects using their software:

Study data were collected and managed using REDCap electronic data capture tools.[22] REDCap (Research Electronic Data Capture) is a secure, web-based application designed to support data capture for research studies, providing 1) an intuitive interface for validated data entry; 2) audit trails for tracking data manipulation and export procedures; 3) automated export procedures for seamless data downloads to common statistical packages; and 4) procedures for importing data from external sources.

The full summary of the survey results can be found in Annex D.

WP 3 – Small Group Interview Process

Two small-group interviews were conducted with researchers at various levels of experience, from the London School of Hygiene and Tropical Medicine and Imperial College London. The interviews lasted for a period of three hours each. A short list of topics was presented, with the discussion following on from the initial topic in each case. The major models for data discovery identified in other project work packages were then presented, using examples running live on the Internet, and the members of the group were asked for their impressions and ideas.

This activity was supplemented by remote interviews with researchers working overseas, to validate the findings, with an emphasis on research conducted in developing countries. Four follow-up interviews were conducted with researchers and those directly engaged in supporting research in Africa, Mexico, and Australia.

The topic list for the small group interviews is found in Annex E.

WP 4 – Review of Existing Approaches

An investigation was conducted into the useful models for data discovery being used within the public health and epidemiology research field, and in other domains including the social sciences, environmental sciences, behavioural research, economics, and official statistics. Through telephone conversations, e-mail exchanges, and an examination of different systems as they are presented on the web, an assessment of current practice was completed, identifying three dominant models which are potentially of utility for public health and epidemiological data. Major examples of each model are given along with the findings.

[22] Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG, Research electronic data capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-81. http://www.sciencedirect.com/science/article/pii/S1532046408001226

2. Findings

Overview

To summarize the overall findings of the project team, approaches to data discoverability and access within the public health and epidemiology research domain are not as advanced as they could be, when compared with other research domains. The landscape revealed by the project analysis is a very fragmented one, with a marked lack of established 'best practice' – individual research projects are typically (but not always) responsible for providing the data they collect to other researchers for re-use, but in some cases data are deposited in data archives.

In some cases, personal enquiries must be made to the principal investigators of projects which have collected data, and the information available about the data is very limited. Access can be challenging as well, since some projects do not manage their data effectively, providing outdated versions of the data to researchers who request it for secondary use. At the other end of the spectrum, a few examples of centralized portals using consistent and standards-based metadata exist. Most data sets fall somewhere between these two extremes.

For documentation, we see a similar fragmentation. In some cases, secondary data users have very little documentation – no more than a list of the variables and labels within the data set. Typically, there is some documentation available in print form, at a varying level of detail across different data sets. In the best cases, documentation is provided at a fine-grained level of detail, and is made available both in print form and in machine-actionable form, according to existing standards and software tools.

The reasons for this fragmentation are varied: in many cases, researchers who collect data are focused on their primary task, which is their own research. They are not motivated, nor required by the terms of their funding, to focus on making their data discoverable or re-usable. In many cases, these tasks are not resourced within the project budget. There is not a culture of support for data discoverability and re-use within the public health and epidemiology research community.

When other domains such as behavioural and social sciences, official statistics, economics, and environmental sciences are examined, we find three models which recur across them, and which could plausibly be used within the public health and epidemiology research domain. One approach involves the creation and operation of centralized search portals, based on standard metadata and known best practice for data management and dissemination. Another approach is more decentralized, involving the utilization of data journals, providing highly visible descriptions of themed groups of data sets, along with links to the data where possible (or instructions for how to request access). A third model is even less centralized, relying on modern developments in web technology: the "Web of Linked Data". In this solution, data sets are described in a fashion which allows for automated searches across the web to be performed, returning information about all the data sets which exist (that is, which have been exposed on the web).

In each case, the use of standard and detailed metadata and documentation either is critical or contributes significantly to making data visible, accessible, and useful. Each approach provides different strengths and weaknesses, and each requires different levels of resource and effort.

The findings of each work package are presented in more detail below.

A. Work Package 1 – Review of Data Sets

Of the significant public health and epidemiology data sets identified and analyzed, all provided links to and/or descriptions of the research papers based on them. Most were documented in PDF or Word format, although some examples had detailed standard metadata, mostly according to the DDI standard, but including MARC 21 and Dublin Core. The data sets made available through data archives such as the UK Data Service and the ICPSR data archive at University of Michigan had very detailed information and good facilities for data discovery. It is notable that both archives use standard metadata in DDI and other formats. (Approximately a quarter of the data sets analyzed in detail were documented in DDI.) Six criteria were used for assessing data discoverability (**Box 2**).

Box 2 – Six criteria for assessing data discoverability

1. ***Study protocols*** – assessing how much information is available about data collection (survey, other documentation of the protocols and methods employed).
2. ***Data documentation*** – form of data documentation, including level of detail and online access, etc., plus use of standard classifications to make data more comparable; includes assessment of standard metadata models such as DDI.
3. ***Data access*** – how the data may be accessed; whether in online form or through application or safe centre; also, what formats the data are available in (SPSS, CSV, Stata, etc.).
4. ***Online data visualization/analysis tool*** – an assessment of whether there was an online analysis tool available for the data, to help researchers explore data to determine its appropriateness for their purpose.
5. ***Online links to/descriptions of publications*** – whether or not citations or links were provided to research publications based on the data.
6. ***Use of social media/other forms of communication*** – assessment of whether social media were used to make the data more visible.

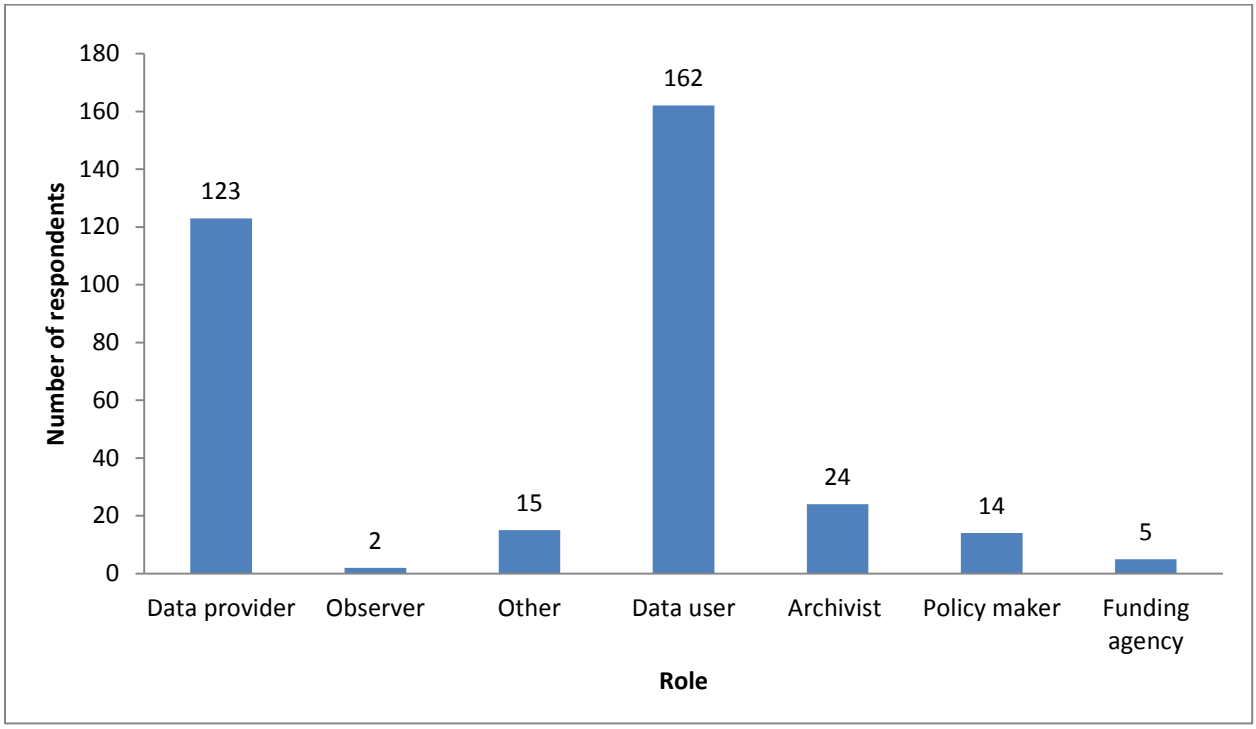
The most common formats for delivering the data were statistical package files: SPSS, SAS, and Stata. Other formats included ASCII text files and CSPro. In some cases, data were made available in multiple formats, based on the preference of the user (this is typical for data archives who specialize in the dissemination of data).

In approximately a third of cases, online tools existed not only for discovery of data sets, but also for online analysis of the data, allowing researchers to explore the data thoroughly before requesting access.

B. Work Package 2 – Online Survey

Among the 253 responses received on the survey, most were from people working for a University or in an affiliated research organization or archive. Respondents based in Europe made up 29% of the sample, but coverage from other regions was fairly even, with large numbers of respondents in Africa and Asia. Every continent was represented. Most respondents were either data producers or data users. When this was taken into account, however, it was somewhat surprising to see that respondents were actively involved across the stages of the data lifecycle in a fairly even manner. The respondents' self-described role in public health is shown in **Figure 1** below.

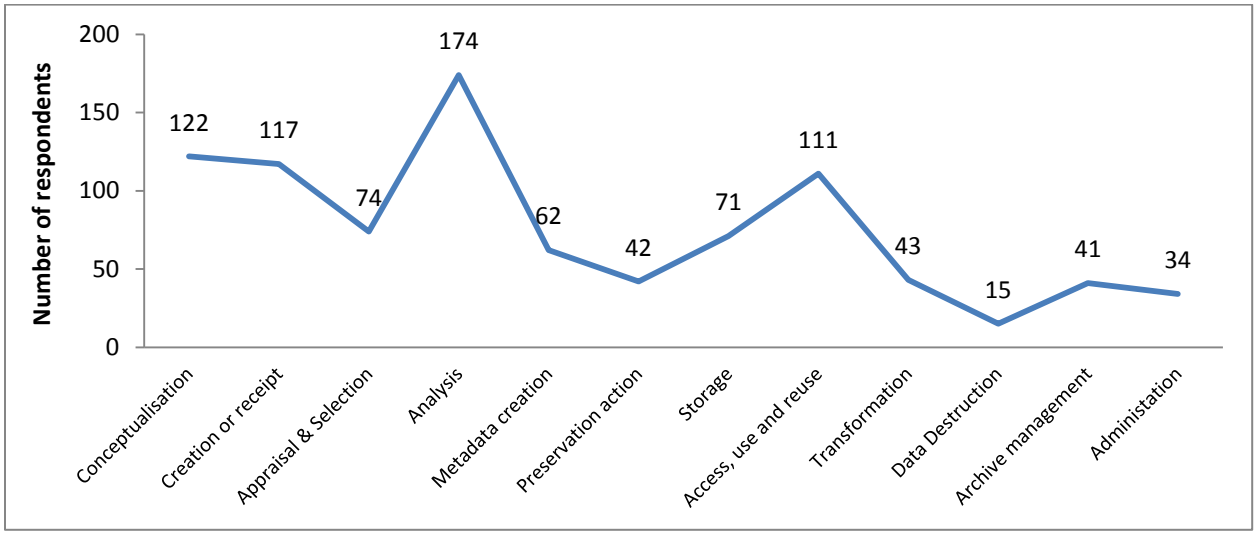
Figure 1 – Respondents' roles in public health



Base: 345 responses (multiple selection allowed) – 218 respondents

Involvement in the research data lifecycle can be seen in **Figure 2** below, but it should be borne in mind that most respondents were involved in collecting primary data for their own research, or were analyzing data. Given this, we see good representation among our respondents.

Figure 2 – Respondents' roles in stages of the research data lifecycle



Base: 1037 responses (multiple selection allowed) – 214 respondents

The five most common funders identified were the Medical Research Council (MRC) and the Economic and Social Research Council (ESRC) in the UK, the National Institutes of Health (NIH) and the Centers for Disease Control (CDC) in the US, the National Health and Medical Research Council in Australia, and the Bill and Melinda Gates Foundation.

The three most commonly used forms of data were survey data, health record data, and data from disease registries (**Table 1**). In free-text comments, it became clear that census and related data were very important, as were data coming from both clinical trials and administrative registers. Further, many of the data sets have a qualitative aspect to them, where interviews and similar types of collection are important.

Table 1 – Forms of data

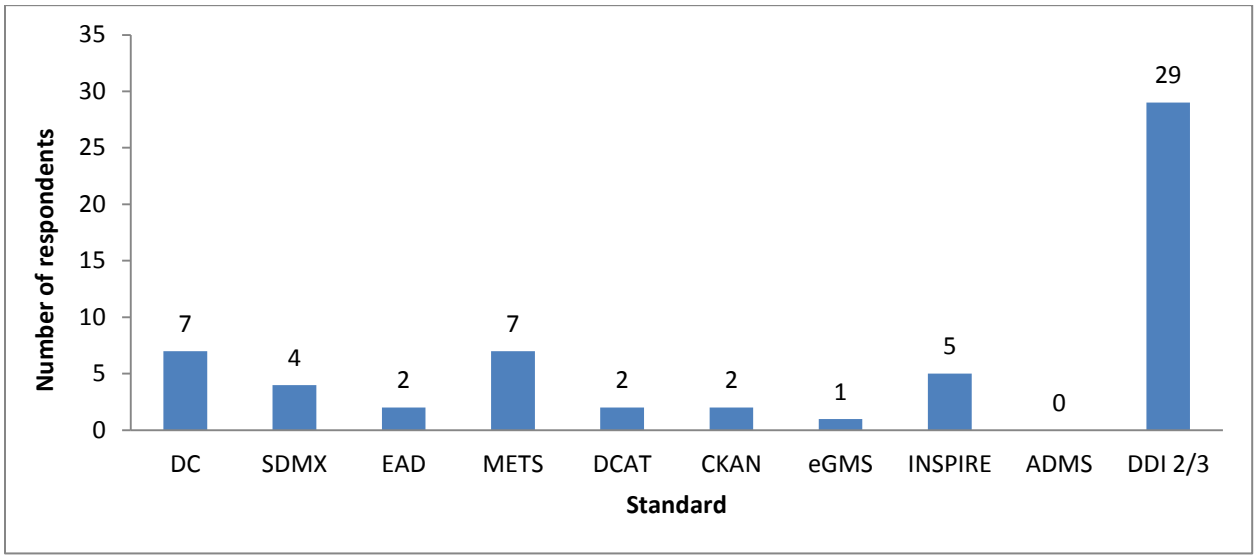
Data type	Number
Survey	157
Healthcare records	125
Disease registries	76
Ethnographic	24
Geospatial	46
Environmental	31
Genomic/Proteomic/Metabolomic	30
Imaging	19
Physiological measurement	47
Other	37

Base: 592 responses (multiple selection allowed) – 211 respondents

Most respondents emphasized that data should be discoverable on the web and that the data and metadata be machine-readable; somewhat less important was availability of the data in a non-proprietary form. The use of an underlying ontology – an agreed set of terms, meanings, and relationships among different parts of the data – was not seen as important. For making data searches, search by keyword and according to search terms were the preferred techniques. The use of related concepts was seen as less important. Notably, ClinicalTrials.gov was mentioned as a good example by some respondents.

Among survey respondents, the best-known repositories were ClinicalTrials.gov and social science data archives. The best-known controlled vocabularies and thesauri were the International Classification of Disease (ICD), MeSH, and the DSM 5. Analysis of the free-text responses showed that many vocabularies are very specific either to national context (as in Australia and the US) or to particular specializations (WHO ATC-DDD codes for measuring drug use). Overwhelmingly, DDI was the preferred standard for data documentation (notably, DDI is largely a product of the social science data archives.) **Figure 3** shows the findings. It is worth noting that those working in data archives tended to be the most familiar with DDI – many respondents acting in other roles did not respond at all to this question.

Figure 3 – Use of standards for data documentation



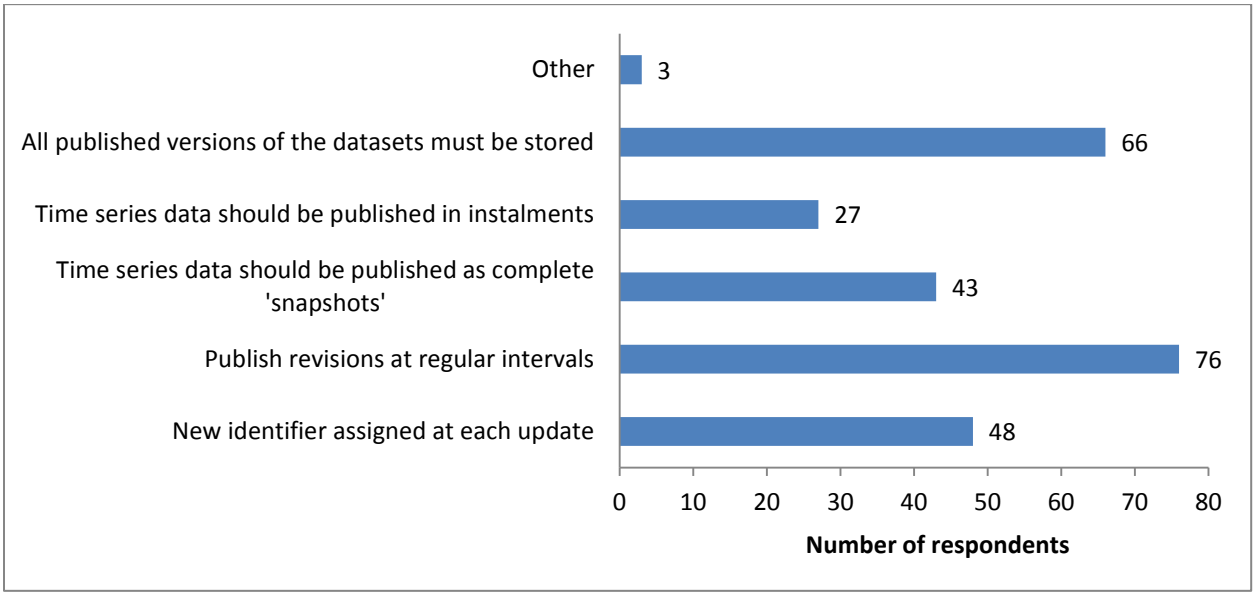
Base: 59 responses (multiple selection allowed) – 43 respondents

About half of respondents were aware of data journals, but most had discovered them through colleagues, journals, or conferences – internet searches had not made data journals visible. The benefits of data citation were those normally claimed, from making data easier to locate, linking with research publications and demonstrating the data's impact, to reducing plagiarism, promoting professional recognition, and providing good long-term infrastructure (such as data journals and archives). Citations were most wanted at the data set level and for collections of data sets – granular citation was seen as less necessary.

For longitudinal and regularly changing data sets, most respondents emphasized that every version of the data should be made available, and that updates be made regularly. Most also wanted to see time-series data published as snapshots – the complete data at a single point in time – and that each new version be identified separately. More detail is shown in **Figure 4** below.

In the free-text responses received to the query, respondents highlighted some of the key challenges and priorities for data discoverability (**Box 3**) – including issues around data documentation and publication.

Figure 4 – Views on how longitudinal and regularly changing data sets should be managed



Base: 263 individual responses (multiple selection allowed) – 140 respondents

Box 3 – Key findings from qualitative analysis of free-text responses

Areas of importance to data discoverability

1. Identification of commonalities and links between studies
2. Creation of standardised metadata and other associated study documentation
3. Ensuring effective use of technology
4. Implementation of relevant governance frameworks to help protect and respect participants' wishes

Remaining challenges

1. Data documentation

- Limited use of standards impacting approaches to research data management
- Insufficient recognition of the heterogeneity of research data impacting the development and generalizability of best practice guidelines
- Inadequate resource availability to meet the financial, time and technological requirements of generating metadata

2. Data publications

- Perceived limited significance of these articles given lack of formal academic recognition and limited acknowledgement of the potential benefits
- Principal investigators, in certain circumstances, are unable to assign writing tasks such as these to sufficiently trained members of staff due to limited resources

Areas for improvement and priority

- Collation and increased publicity of public health and epidemiology studies
- Ensuring the boundaries of consent are made clear to potential secondary researchers and that the necessary infrastructure to support this is implementable
- Establishing a set of best practices through increased standardization in research

C. Work Package 3 – Small Group Interviews

The small group interviews provided some interesting findings. The responses from better-established researchers were, not surprisingly, somewhat different from those from PhD students. Some themes emerged strongly, with the participants' experience and roles making little difference. One theme that emerged quickly is that data discoverability is not a major issue for the researchers consulted: they know their field, and – generally from reading research papers – are aware in broad terms of the data sets which are most important to their work. The issues were related to being able to access and use the data effectively, with a lack of documentation and poor data management as significant issues. Another issue identified was the time taken to receive permission to use confidential data.

For data searching, the centralized data portal was the preferred model for those interviewed. They viewed some of the DDI-driven sites such as the UK Data Service portal, the CESSDA portal, and the International Household Survey Network's catalogue, and all liked the consistent presentation of metadata at the variable level, which allowed them to inspect the data set's contents. Second in preference was the data-journal model (they were shown various data journals on the Ubiquity Press site). This was seen as more flexible than the centralized portals viewed, as the range of data was broader, covering such things as geographic data sets, and not just normal quantified data as seen in the centralized portals. The lack of consistent metadata available once the links to datasets were traversed was commented on, however. A system based on the highly decentralized Web of Linked Data model was seen as the least desirable, in part because the coverage was poor, but also because of the very inconsistent information provided for the different data sets. The application shown to interviewees was Data.gov.uk, an Open Data site operated by the British government, containing some public health-related data sets.

The discussion with the PhD students resulted in a very interesting picture emerging, regarding their access to data. Their experience was that the principal investigator who held the data set they wished to use would need to be contacted directly (phone or email), and that they would be provided with poor to marginally acceptable documentation (in one case, just a list of variable names in an email). On the basis of this documentation, they would request the needed variables. When the data was received – a process which could take months, as the principal investigators and the members of their research teams were typically busy conducting their own research – it was often an older version of the data, or contained a different set of cases than those the data producers had used in their own research.

It became clear that, unlike in some domains, these students did not have any professional assistance or infrastructure to help them find and use data. This contrasts sharply with domains such as the social sciences and economics research, where data archives, data libraries, and research institutions provide expertise and infrastructure to support the discovery, access, and use of the data they disseminate. This was less of an issue for more experienced researchers, or for the data manager, who worked frequently with data that they and their colleagues knew well.

The PhD students also discussed researchers' tendencies and motivations with regards to the capture of metadata and documentation. It was clear that they liked to work with well-documented data, but also that they understood why researchers rarely document sufficiently, as it is a low-priority task. One solution was suggested, based on the use of the statistical packages as tools for data management, an idea which has an increasing number of proponents in various research fields.

Four follow-up interviews with researchers conducted by phone and over the Internet focused on those working in developing countries (Africa, Central America, Oceania). In most cases, these interviews resulted in similar findings as those conducted in the UK. There were some issues which seemed to be

more common among this group of interviewees. The lack of infrastructure was mentioned, both at the level of reliable access to the Internet, and in terms of the available analysis tools: it was far more common to use simple spreadsheet tools such as Excel than some of the more sophisticated analysis packages such as Stata and SPSS. This was, however, dependent on the institutional setting in which the researchers worked – in some cases the infrastructure was very good, being provided by a well-funded organization.

Another issue which came up frequently was that of language – in research supported by large organizations such as the World Bank, there was typically good support for the translation of documentation into common local languages (French in parts of Africa, as well as English, etc.) But this was not true in all cases, where the language and level of detail could be problematic (only some of the existing documentation would be translated). Language was not, however, seen as a major barrier, as most available data sets have an English version, which seemed to be the de facto standard.

It was clearly the case that there was a greater reliance among this community on data collected through official statistical organizations (different UN organizations such as the WHO, the World Bank, national statistical offices, etc.) than is the case in the developed world. This is perhaps not surprising, considering that often the lack of infrastructure is solved by using that provided by international organizations. Examples of this include the International Household Survey Network (a group of UN organizations and the World Bank) and the INDEPTH Network.

D. Work Package 4 – Existing Approaches

Three models emerged in the review conducted by this work package, namely:

1. Centralized data portals
2. Data journals
3. Web of Linked Data approaches

Each of these will be considered in turn in Chapter 3, along with some examples considered during the work (as summarized in **Table 2** below) and consideration of the associated benefits and costs.

It should be noted that these approaches are not mutually exclusive – this will be significant when the possible options are considered, below. The findings from WP 4 have been integrated into the Options for the Future section, below.

Table 2 – Key models identified with examples examined

Model	Typical Examples
Centralized data portals	CESSDA portal (http://www.cessda.org) International Household Survey Network Survey Catalog (http://catalog.ihsn.org) MRC Gateway (https://www.datagateway.mrc.ac.uk/) Global Health Data Exchange (http://ghdx.healthdata.org/) INDEPTH Network (http://www.indepth-ishare.org/index.php/home)
Data journals	Ubiquity Press (http://www.ubiquitypress.com/journals) Earth System Science Data (http://www.earth-system-science-data.net/) Nature Scientific Data (http://www.nature.com/sdata/)
Web of Linked Data approaches	HealthData.gov (http://www.healthdata.gov) Data.Gov.UK (http://data.gov.uk/data/search) Digital Enterprise Research Institute (http://www.deri.ie/)

Interestingly, while this report was being prepared, the Medical Research Council in the UK published a strategic review of its investment in cohort studies[23]. Two of the recommendations from this report resonate with some of the themes we have identified regarding data sharing and centralized data portals – namely that

- Cohort leads should ensure that their studies are easily discoverable via directories. Processes are needed to ensure that all MRC funded cohorts comply with MRC data sharing policies. Studies need to be accessible and have transparent governance procedures in place that enable data sharing and where appropriate, access to samples.
- Adoption of core common data standards, sharing knowledge and improving meta-data quality should be encouraged and facilitated by cohort studies, the MRC and other funders.

[23] Medical Research Council. Maximising the value of UK population cohorts: MRC strategic review of the largest UK Population cohort studies. Medical Research Council; 2014.

3. Options for the Future

This section describes what the possible options are for implementing these models to improve the discoverability, accessibility, and usability of public health and epidemiology data.

A. Technology Approaches

In the findings from Work Package 4, we have three distinct models for the application of IT technology to the challenge of data discoverability. Detailed description of these is presented here, along with a discussion of possible future actions. As noted above, these approaches are not mutually exclusive, and could be used in combination. They are at different levels of maturity in terms of their use, and they have different degrees of familiarity within the public health and epidemiology research domain. Each has different strengths and weaknesses, and apportions the costs of implementation differently.

It should be noted that the problem is not a technology problem – all three of the approaches that emerged from the work of the project use technology that has been shown to work in production settings, albeit to different degrees. The problem is a cultural and organizational one: all of the technology approaches rely on standard metadata, and – if they are to be implemented in an optimal way – that metadata must be rich.

The culture of researchers is not to prioritize the creation of this type of metadata, nor is it to create and operate the type of infrastructure which is required by any of these models. The organizations which develop and operate solutions for data discoverability in domains such as social science do not exist to the same degree in the public health and epidemiology research community.

1. Centralized Data Portals

There were several examples of the centralized portal approach considered, both from within the public health and epidemiology research domain, and outside it. Those within the domain included the MRC Gateway, the INDEPTH Network, and the Global Health Data Exchange (GHDx), hosted at the University of Washington. From outside the public health and epidemiology domain, there were several good examples. The broadest of these in scope is the prototype now being created by a European Commission project called “Data without Boundaries” (DwB), which spans both the CESSDA archives (the social science and economics data archives in Europe) and the research data holdings of the national statistical agencies in Europe. The DwB portal is intended to become an expanded version of the CESSDA portal, which today includes only the archives’ holdings. The International Household Survey Network’s catalogue was also analyzed, holding data from many statistical agencies in the developing world. Holdings at archives such as the UK Data Service, ICPSR at the University of Michigan, and the IPUMs data at the University of Minnesota Population Center were also included.

In the best examples from this model, the metadata used was based on the DDI standard, although other standards such as MARC 21 and Dublin Core were also common (although typically populated from the DDI metadata as needed). In most cases, a sub-set of that standard had been identified and documented, and was implemented according to an agreed best practice. Many software tools are available, and in many cases are free to the community.

Both within the social science domain and within official statistics (the IHSN, DwB) we see a culture of best practice, and the resourcing of infrastructure for promoting data discoverability and re-use. Within Europe, this is very pronounced – DwB is cooperation on a very large scale, spanning both domains. In addition, there is an established class of data professionals who are not themselves researchers, but whose job it is to assist researchers in locating and using data.

In these examples, detailed metadata are published in an agreed DDI form, and standard harvesting protocols are used so that the holdings of each organization can be programmatically indexed and updated by the data catalogue system.

In some cases, such as the INDEPTH Network and the GHDx portal, we see an attempt to provide the same type of service, but without benefit of the rich, standard metadata holdings that we find with the CESSDA archives or the IHSN, etc. The results are correspondingly less impressive. In the GHDx case, we see that the creation of the portal – and the collection of needed metadata – is an unfunded activity, performed when possible on the budgets of funded research projects.

Interestingly, the MRC Gateway is a good example of the use of standard DDI metadata across organizations who have agreed to that practice. It would seem to build on the best examples from the social science domain and that of official statistics.

The benefits of having rich, standard metadata produced according to an agreed best practice, and having dedicated infrastructure and staff available to support the location and re-use of data are many: the researchers can easily locate, access, and use the data they need. The costs here are large, however, both for the community and for each data-holding organization. Rich metadata is time-consuming to produce, and it is typically not created by the researchers themselves, but by dedicated staff. Infrastructure which operates across organization boundaries is difficult to operate, as many agreements are needed, and the costs of operation must be shared (CESSDA recently incorporated so that it could manage the shared infrastructure among the European data archives).

To give some idea of the costs around establishing large search portal infrastructure, the CESSDA ERIC, covering the social science data archives in Europe, requested an annual budget of 1 million Swedish Krona for 2013 and 2014, which would be supplemented by internal funding applied to maintaining local infrastructure within each national archive. It is difficult to estimate the actual cost of establishing a CESSDA-type infrastructure because the funding streams of national archives differ across Europe. However, a general sense of how much developing and operating such an infrastructure costs can be obtained by looking at CESSDA. Note that the four-year DwB project – also European in scope – had a funding level of approximately 9 million Euros, across many different work packages – perhaps a quarter of this money went into work packages related to data discovery. Note, however, that research and development is often more expensive than operations once a portal has been created.

Possible Future Actions

The basic possibilities can be understood in the context of the approaches outlined above. The first is to undertake the creation of large-scale centralized portals focused on public health and epidemiology. These could exist at several levels: national portals, regional portals (Europe, Africa, North America, etc.) or even global portals. The requirements for this are several:

1. *Organizational foundation for the portal creation and maintenance* – There needs to be an organization or a set of organizations which agree to fund and manage the development of the portal, and there must be staff to conduct on-going operation and maintenance. If we look to the CESSDA

example, the start-up costs were significant, but on-going operations and maintenance costs have been much less.

2. *Established tools and protocols for exposing data holdings to the portal* – Data producers and archives would need to have as a low a barrier to entry as possible. Metadata creation is expensive – if the costs of connecting to a central portal are also expensive from an IT perspective, then such an approach would likely fail. Existing standards – notably DDI – could provide the basis for this, especially since many useful DDI-based tools already exist.

3. *Best practices and incentives for data documentation* – Data producers would need to have clear guidelines about what was expected, not only tools and protocols. Further, they would need to be incentivized and funded for what is perceived today as additional work. If tools which were integrated into common statistical packages could be developed, then that might lessen the resistance to such a change.

4. *Archival and research infrastructure* – Something we see in the social sciences is a solid structure for archiving data and supporting researchers locating and using it. Building such a network of archives from the ground up is not feasible, but collaboration with existing archives could be. Many social science archives already hold some public health data sets, and archives exist also within the domain. Collaborations with these organizations could provide the needed support. Notably, the DDI community is very open to working with domains besides the social sciences, from which it emerged, as we have seen in its recent collaborations with statistical agencies, and with some members of the public health and epidemiology research community.

2. Data Journals

The data journal model is one which is becoming well-established, but it is relatively recent compared to the centralized portal approach[24]. Typically, a data journal is an online publication containing peer-reviewed articles written by the data producers about their data. It gives a description of how the data was produced and what its coverage is, along with information about where the data is and how it can be accessed. Importantly, data papers are different to ordinary research articles in that they only report the data itself, how it was produced and how it is preserved, as opposed to any analysis of the results or lengthy discussion of the data.

Data papers are peer-reviewed to ensure they conform to community norms, e.g. the data is uploaded to a suitable repository, is correctly labelled, is available in a non-proprietary format, and is accurately described by the paper. Furthermore, data journals ask authors to outline the background to the data, including the methods used, and to provide suggestions for re-use of the data. The intention of the data journal is to work within the current accepted system of academic credit through paper citations. This incentivizes authors to release the data and in turn ensures data can be cited according to standard academic practice (i.e., within an article's reference list).

There are several good examples of data journals in existence today. One good example is the Nature Scientific Data site (http://www.nature.com/sdata/), which has articles termed "data descriptors". There are many data descriptor articles available on their site, and the data policies, editorial board, and guidance for authors and peer reviewers is provided. Another similar example can be found at Ubiquity Press (http://www.ubiquitypress.com). Here, they have what they term "metajournals" covering the

[24] For an interesting discussion of these issues, see: Arend D et al. e!DAL - a framework to store, share and publish research data. BMC Bioinformatics 2014;15:214. http://www.biomedcentral.com/1471-2105/15/214

publication of data articles, software, and bioresources. The data publications are organized topically into data journals for different disciplines, so the coverage is wider than the Nature Scientific Data site, but the editorial board, author guidelines, and similar information are provided.

Data citation is obviously a counterpart to this model, because the data articles link to the data they describe. These data must be available in citable form – that is, they should be available on an on-going basis, and they should be accessible according to a known protocol. Again, we have an example popular within the social sciences domain – DataCite – which is effective in guaranteeing this.

The proponents of data journals claim many benefits (and we see from our survey results that many people concur), among them the enhancement of the data producer's professional reputation. Critics of this model claim that data publications are perceived in many disciplines as "second-class" publications, and are not given the same weight as more traditional research publications.

From the perspective of data discoverability, they have obvious benefits: they create a web description of the data, written by those who understand it best. Further, they have been subject to peer review, according to a publication model that is very familiar to researchers, journal publication. Access to the correct version of the data is also guaranteed.

The biggest perceived problem with this approach is that the documentation and metadata accompanying the data are often inconsistent. It should be noted, however, that this is a correctable problem – it took the social science community two decades of effort to create the infrastructure for building centralized search portals. For data journals, the establishment of best practices around data documentation and metadata is something which could be achieved in time.

The costs associated with data journals are smaller than those of centralized data portals – the journal publishers need to exist, but this is not a cost for the data producers. If data publications are taken seriously, then researchers are incentivized to create them as a normal part of their work. An infrastructure for data archiving is needed, however, and some agreement must be made for the resolution of citations. These are costs which fall partly on the data producer, and partly on the journal publisher.

Possible Future Actions

The second possibility is to utilize the data journal model. This approach could involve a collaboration with the publishers of data journals, and the promotion of their use within the domain.

Again, there are several requirements:

1. *Established journal or journals* – This would probably not be difficult to do, but the funding models for the journal publishers would need to be considered. The journal(s) would need to be a stable presence on the web or through existing repositories such as PubMed.

2. *Creating best practice for data documentation* – As for the centralized portal approach, having good metadata for researchers to use when discovering, accessing, and using data is very desirable. If this model is to be the core of a solution to data discoverability, then we would want to optimize for the researcher's requirements. As we saw in the small group interviews, consistent, rich documentation is what is wanted.

3. *Tools and protocols for data documentation and data citation* – If we are to realize our vision of having good metadata and documentation, the tools must exist for producing it, just as for the

centralized portal approach. In addition, support for data citation standards is also needed, as this is central to the data journal model.

4. *Archival infrastructure* – If data are to be available for citation from the journals, it is important that there be places which can reliably provide the data to researchers for linking. This is very similar to the requirement for the centralized portal approach.

5. *Incentives for data publication* – While the creation of data publications is a trend that may continue in its own right, pro-active work to promote this model would be very useful in increasing the coverage of important public health and epidemiology data sets. This means selling the concept to the researchers who have the knowledge to write data articles.

3. The Web of Linked Data

This approach is one which comes from the technology world, but has been promoted heavily by the proponents of Open Data, advocating for greater government transparency. Although many members of the public health and epidemiology research community are familiar with the idea of Linked Open Data, actual experience with it is slight.[25]

This model is ontology-based: an ontology is an agreed set of terms and definitions for various types of information within the domain, providing relationships between them, and describing their properties. Within any domain, an ontology is established for describing some set of information, and anyone who wishes can publish descriptions of their data (and even the data itself) according to the standard ontology. Tools for searching and working with the information published are based on standards from the W3C, which is the organization responsible for all web standards. There is no centralized, dedicated infrastructure in this model – the infrastructure is the Internet, and the web itself.

Clients – the software packages and websites researchers would use to interact with the Web of Linked Data – could be created by anyone. Data resources and metadata could be linked with anything publishable on the web, including, of course, research publications (or articles in data journals).

Some standards used for data description – notably DDI and SDMX – have published versions of their models so that they can be used as domain ontologies for data description. One good example of their use comes from DERI, an institute associated with the University of Galway in Ireland. There, they have created a program which takes all of the publicly available data from Eurostat (the European Commission's statistical arm) and publishes it into the Web of Linked Data.

It is difficult to assess the costs of this model, and it is difficult to understand how it could be managed within a domain. The Web of Linked Data is a phenomenon which continues to evolve, but it is not as mature – from a public health and epidemiology perspective – as either of the other two models. It is difficult to understand what impact it could have as the only approach to data discoverability, especially since it is not a familiar model within the domain.

It is worth noting, however, that the cost of infrastructure is minimal: we already have the web, and there exist DDI-based ontologies which could be used for data description. The costs fall on whoever undertakes implementation of the technology standards for a given data set: this could be the data

[25] For information about how these technologies might be used, see: Marshall MS et al. Emerging practices for mapping and linking life sciences data using RDF. Web Semantics: Science, Services and Agents on the World Wide Web 2012;14:2-13; and Sinaci AA, Laleci Erturkmen GB., A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains, Journal of Biomedical Informatics 2013;46:789-794.

producer, a data publisher (e.g. a data journal), or even an external third party who has access to the data (as in the case with DERI and the Eurostat data – note that DERI is a technology institute).

That said, there is growing interest among the pharmaceutical industry in the technologies around the Web of Linked Data, so it may come to be of greater importance in health-related fields in time. An interesting discussion on this topic can be found in the Journal of Cheminformatics (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3121711/), titled "Linked open drug data for pharmaceutical research and development". It is also notable that the W3C has a working group in this area as part of its stable of Linked Open Data standards – Linked Open Drug Data.

Possible Future Actions

The third approach would be to use the Web of Linked Data model. The requirements here are less clear than for our first two approaches. In order to promote this model, it would be a good idea to establish a significant presence within the public health and epidemiology domain on the web, by describing some important data sets with an appropriate ontology, and then to promote this idea among all data producers and disseminators in the domain.

The requirements here are somewhat different than before:

1. *Identify an appropriate ontology* – As mentioned, DDI has one which could be used, but others might also exist and be worthy of evaluation. Looking at the existing Open Data sites could provide other options (e.g. HealthData.gov, Data.gov.uk), as would an evaluation of the standards emerging from the W3C.

2. *Collaborate with high-profile data producers within the domain* – Data produces clearly do not see this as a priority today. They would need funding and assistance in learning the new technologies and standards, and to be incentivized to engage in this approach. Partnering with members of the linked data community and running joint projects might be a good approach.

3. *Establish tools for creating rich documentation and metadata* – As for the other approaches, tools would need to be created if data producers are expected to engage.

4. *Create guidelines and incentives for describing data sets* – Again, data producers would need to be incentivized, and would need clear guidance on what is expected.

5. *Create a client application for end users* – The functionality for researchers to query across datasets does not currently exist in a usable form, although some generic tools which do exist could be used as a basis. Again, engaging with the Linked Open Data community – the technologists who are promoting these standards in different domains – would help here.

B. Metadata and Data Management Practices

It should be noted that for all three models, there is a need for rich metadata, regardless of how it is expressed or implemented with IT. The requirements here are very similar for the three approaches:

- Tools are needed for the capture of metadata, regardless of whether this is done by researchers, archivists, or others.
- The culture of metadata capture and documentation needs to be established. Clear guidelines and defined best practice are needed.

- Incentives must be provided for data producers to document their data at the desired level, as today this is something which is neglected in the majority of cases. What is needed is rich, standard metadata which is expressed in machine-actionable forms, rather than as PDF or Word documents.

It is interesting to note that DDI and some other standards for describing data are used to drive data management, with discovery and documentation as only a part of the process. DDI was designed to support good data management, which was in some cases not being practised by the data producers we encountered during the project. Although the scope of this project did not extend to data management, it is useful to note that much of the metadata and documentation needed for data discoverability is also very useful in managing data effectively, and can be re-used for both functions.

4. Recommendations

These recommendations are based on the clear preferences expressed by users in the survey and small group interviews, and on plausibility and relative costs. The end goal is to provide useful solutions for researchers, but we cannot ignore the needs of other stakeholders such as data producers, funders, and archives.

Of the three approaches identified by the project, the preference of researchers consulted was for a centralized data portal. This is also the most mature model, with large-scale implementations such as the CESSDA portal in production use for many years in the social sciences domain.

The second approach, perceived as useful but not as well-liked, was the data journal model. If some aspects of this could be guaranteed – such as the existence of consistent and rich metadata for data sets which were the subject of data articles – then this approach would become more useful.

The Web of Linked Data approach was not seen to be as useful by researchers, mainly because there is no centralized place where they know to go when looking for data. It is also less easy to see how this approach could be managed, since it does not naturally provide any single place which can be governed and managed by the funders of public health and epidemiology research. Further, this seems to be the least mature approach, and uses technology which is not generally familiar to data producers and other members of the research community.

All three approaches lend themselves to combined use. One could imagine a centralized portal which supported data citation for the data sets it indexed – this would make the portal a place where data journals could point when citing the data sets, and would guarantee a consistency and richness of the metadata and documentation for the cited data. One can also imagine that data journals might provide descriptions of the data not just as data articles, but also as linked-data descriptions of the data sets, according to a standard ontology. This would expose the data sets to the Web of Linked Data for discovery using that set of technologies and tools, as well as supporting other approaches.

It is perhaps not feasible to fund the combined use of all three approaches now, but it might serve as a goal for the longer term. Given the priorities as described above, the recommendation would be to focus first on the creation of a centralized data portal, following the successful models mentioned. The clear choice for a standard metadata model is DDI, but a suitable profile for using it would need to be identified, and tools and guidelines established. Archival infrastructure is needed, which suggests collaborating with the existing archives, whether they come from within the domain or from related domains such as social science. Engaging with members of the DDI community, many of whom are archives, would be potentially useful. Further, incentives for both documenting and archiving data would need to be provided – the funders are in an excellent position to mandate these activities, assuming they are willing to recognize that they require additional resources on the data producers' part. In order to coordinate these activities, the Data Forum or similar initiative could be used as a basis for the creation for an organisation to establish and host the central data portal, allowing for its funding, development, governance, and operation. This need not be a legal entity – CESSDA was purely a collaboration until very recently – but would need to be an initiative involving the significant players.

Having taken these steps, it would be possible then to engage with the publishers of data journals, and to make sure that the centralized data portal was also useful as a resource for them – that is, that it would support data citation at the level of data sets and collections of data sets.

A third step would be to take the profile of DDI already established, and to see if the linked-data ontology provided by the DDI Alliance, based on the same model, could be recommended for use by those implementing linked-data technologies. Because this was the lowest-priority approach for

researchers, and because it is the least mature and least-known approach generally, we recommend a wait-and-see attitude at this time. Having a recommended ontology, however, could be useful to those who wish to employ these technologies in future. This is especially true as there would be no large cost in using an existing ontology, based on the DDI metadata model which has already been selected, although implemented using different technology standards.

To summarize:

(1) Focus on the creation of a centralized domain portal for public health and epidemiology research, taking the following steps:

   (A) Develop a search portal, with an interface similar to the examples described (such as CESSDA's and the UK Data Service portals). This would involve also a mechanism for harvesting metadata exposed by data producers and archives.

   (B) Identify technical standards and protocols based on the DDI standard and an analysis of the various harvesting protocols such as the OAI-PMH protocol used by CESSDA (and others), and the DwB WP 12 Prototype. Other networks (such as the MRC Gateway and the INDEPTH Network) should also be considered.

   (C) Establish guidelines and best practices for the use of technical standards and protocols for exposing data holdings to the domain portal.

   (D) Establish best practices and guidelines for archiving data holdings, based on any of the archival best practices found in the public health and epidemiology domain, the behavioural and social sciences, and the economics domain. Engage with existing archival infrastructure where possible, rather than trying to create wholly new archives. Researchers looking for secondary data to use should be supported following best practice as found in these archives.

   (E) Develop tools and guidelines for researchers where required to encourage good practices around data management and documentation. Tools should be DDI-based, so that data can easily be exposed to the centralized portal and archived.

   (F) Create incentives for research projects to follow established best practice for data management, documentation, archiving, and data sharing. Funders must recognize that these activities do require additional resources on the part of research projects which produce data.

(2) Encourage the use of data journals and further publication of data articles in the public health and epidemiology research domain. This will include making sure that archival practices established for the centralized portal include dissemination of data sets which are citable, to allow for easy linking into the same data sets catalogued in the portal. A standard such as DataCite might be considered here. Also, standards and best practices for data documentation should be established (the DDI documentation used by the centralized portal could be re-used for this purpose, or a direct link to the portal could be used from the data article).

(3) Evaluate the potential of the Web of Linked Data regarding public health and epidemiology research data. The data journals, the archives, and the centralized portal might wish to leverage this technology approach in the medium term, so agreed ontologies (based on the DDI ontologies and other data-related ones) should be established and promoted.

We see this approach as best meeting the researcher's needs, and also providing an optimal solution to the problems of data discovery, access, and re-use.

Annexes

There are several annexes to this report providing further detail on various aspects of the project. These can be found in the separate Annexes document:

Annex A: Data Documentation and Access: Characterizing Current Practice

Annex B: Documenting Cohorts and Data Resources

Annex C: Survey Questionnaire

Annex D: Results of Online Survey

Annex E: Topic List for Small Group Interviews

Annex F: Project Data Journal – Progress to Date

We are a global charitable foundation dedicated to achieving extraordinary improvements in human and animal health. We support the brightest minds in biomedical research and the medical humanities. Our breadth of support includes public engagement, education and the application of research to improve health. We are independent of both political and commercial interests.

Wellcome Trust
Gibbs Building
215 Euston Road
London NW1 2BE, UK
T +44 (0)20 7611 7221
F +44 (0)20 7611 8254
E education@wellcome.ac.uk
wellcome.ac.uk

The Wellcome Trust is a charity registered in England and Wales, no. 210183. Its sole trustee is The Wellcome Trust Limited, a company registered in England and Wales, no. 2711000 (whose registered office is at 215 Euston Road, London NW1 2BE, UK). SP-6052/6-2014/JC