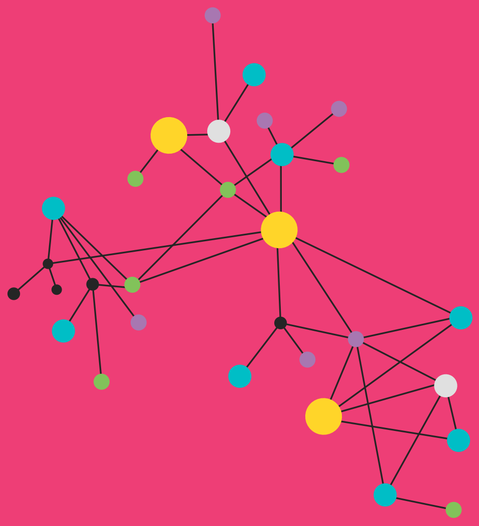
Public Health Research Data Forum

Enhancing Discoverability of Public Health and Epidemiology Research Data: Summary

July 2014



wellcome trust

Public Health Research Data Forum

Executive Summary

Introduction

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 – notably in the areas of social science research, economics, behavioural science, and official statistics. It sought to investigate the perspectives of four major groups of stakeholders: researchers and secondary users of data; data producers; data archives, libraries, and other data disseminators; and funding agencies.

Project team

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 Dr Spiros Denaxas, Clinical Epidemiology, Farr Institute at UCL Partners, University College London

Veerle Van Den Eyden, 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 summary and full report are available to download from **wellcome.ac.uk/PHRDF**

Executive Summary

Goals and methodology

The goals of this project were threefold:

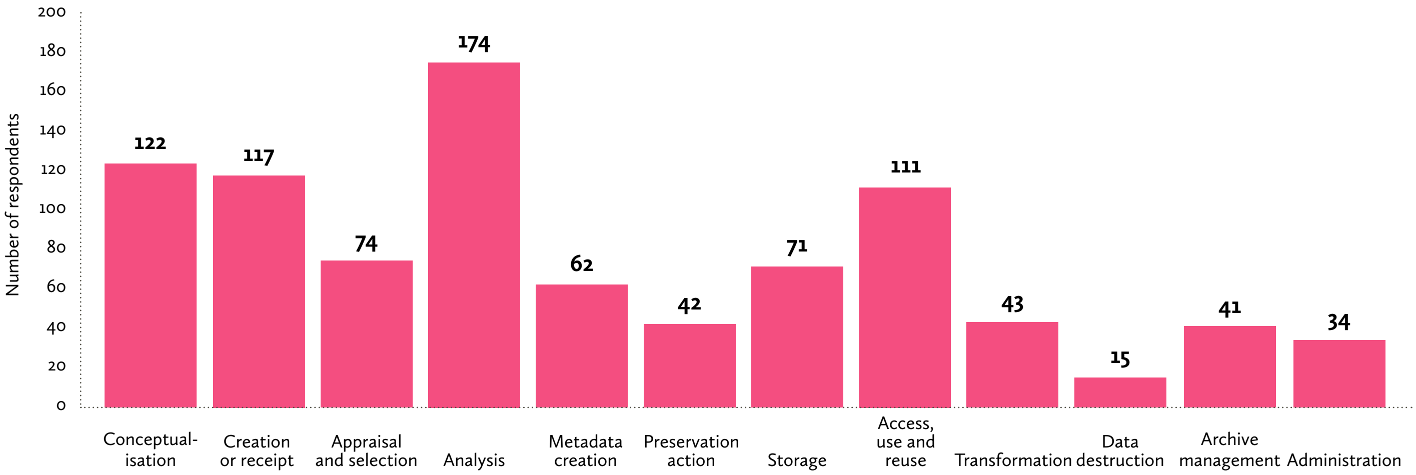
1. To examine the current discoverability of public health and epidemiology data sets, and determine any barriers to access
2. To examine current models for data discoverability such as archives, data portals/catalogues, and other systems to facilitate data discoverability, and to determine which are 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

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.

More than 250 responses were received to the online survey – with respondents from across the globe with expertise spanning the research data lifecycle (see **Figure 1** below).

Figure 1

Respondents' roles in stages of the research data lifecycle



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

BOX 1
SIX CRITERIA FOR ASSESSING DATA DISCOVERABILITY

In examining data resources, the team developed a process for assessing data discoverability based on six key criteria

Study protocols – assessing how much information is available about data collection (survey, other documentation of the protocols and methods employed)

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 Data Documentation Initiative

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.)

Online data visualisation/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 purposes

Online links to/descriptions of publications – whether or not citations or links were provided to research publications based on the data

Use of social media/other forms of communication – assessment of whether social media were used to make the data more visible

Key findings

The 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 for infrastructure related to data discovery and re-use.

In the free-text responses, survey respondents highlighted some of the key challenges and priorities for data discoverability (**Box 2**) – including issues around data documentation and publication.

BOX 2
KEY FINDINGS FROM QUALITATIVE ANALYSIS OF FREE-TEXT RESPONSES

Areas of importance to data discoverability

- Identification of commonalities and links between studies
- Creation of standardized metadata and other associated study documentation
- Ensuring effective use of technology
- Implementation of relevant governance frameworks to help protect and respect participants' wishes

Remaining challenges

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

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

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 exists and is 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 is 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.

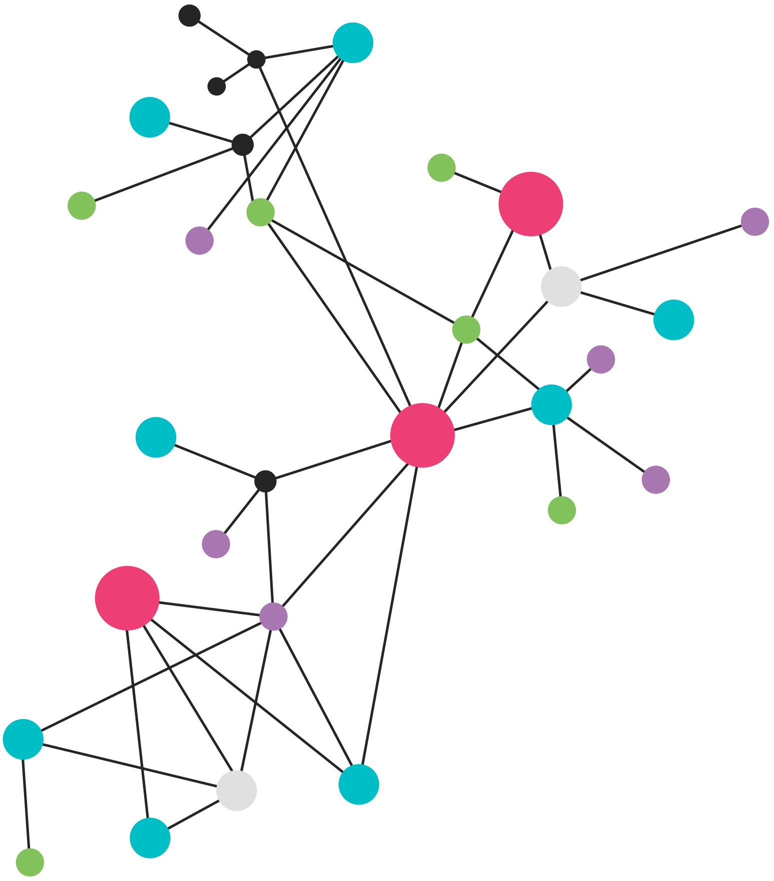
The Centralized Portal Model was the preferred approach among researchers. This is also a model which requires a high degree of coordinated infrastructure across organizational boundaries, both for the cataloguing of data sets and for the reliable archiving of data. The production of standard, rich metadata 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.



BOX 3
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[1] and UK Data Service[2] 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[3] protocol used by CESSDA (and others), and the DwB WP[4] 12 Prototype. Other networks (such as the MRC Gateway[5] and the INDEPTH Network[6]) 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[7] 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] cessda.net
[2] ukdataservice.ac.uk
[3] openarchives.org/pmh
[4] dwbproject.org
[5] www.datagateway.mrc.ac.uk
[6] indepth-ishare.org
[7] datacite.org

The Wellcome Trust

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 8888
F +44 (0)20 7611 8545
E contact@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-6042.3/07-2014/MD