Enhancing Discoverability of Public Health and Epidemiology Research Data

NADDI
Madison, Wisconsin
April 9, 2015
Arofan Gregory
Open Data Foundation

Overview

- General points
- Activities
  - Review of significant data sets
  - Online survey
  - Focus groups
  - Technology review
  - Data journal from project
- Models
  - Centralized portal search
  - Data journals
  - Linked Data on the Web (LDOW)
- Options
- Recommendations
- Next Steps

General Points Regarding Public Health Research

- Researchers want a “data Google”
  - One place on Internet for getting all relevant hits
  - Detailed information about data sets
- Not a technology challenge!
  - What is needed is good variable-level information
  - All technology approaches use the same basic set of information
- Not on the “bleeding edge”
  - Public Health Research can benefit from the work in other domains
  - There are practical, proven solutions

The Data Discoverability Project

- Conducted by Wellcome Trust on behalf of the Public Health Research Data Forum
  - More than 20 international funders of health research
- Project lead was Tito Castillo
- Other participants from Farr Institute (UCL), London School of Hygiene and Tropical Medicine, UK Data Archive, Ubiquity Press, Open Data Foundation
- Short duration (6 months) – moderate budget

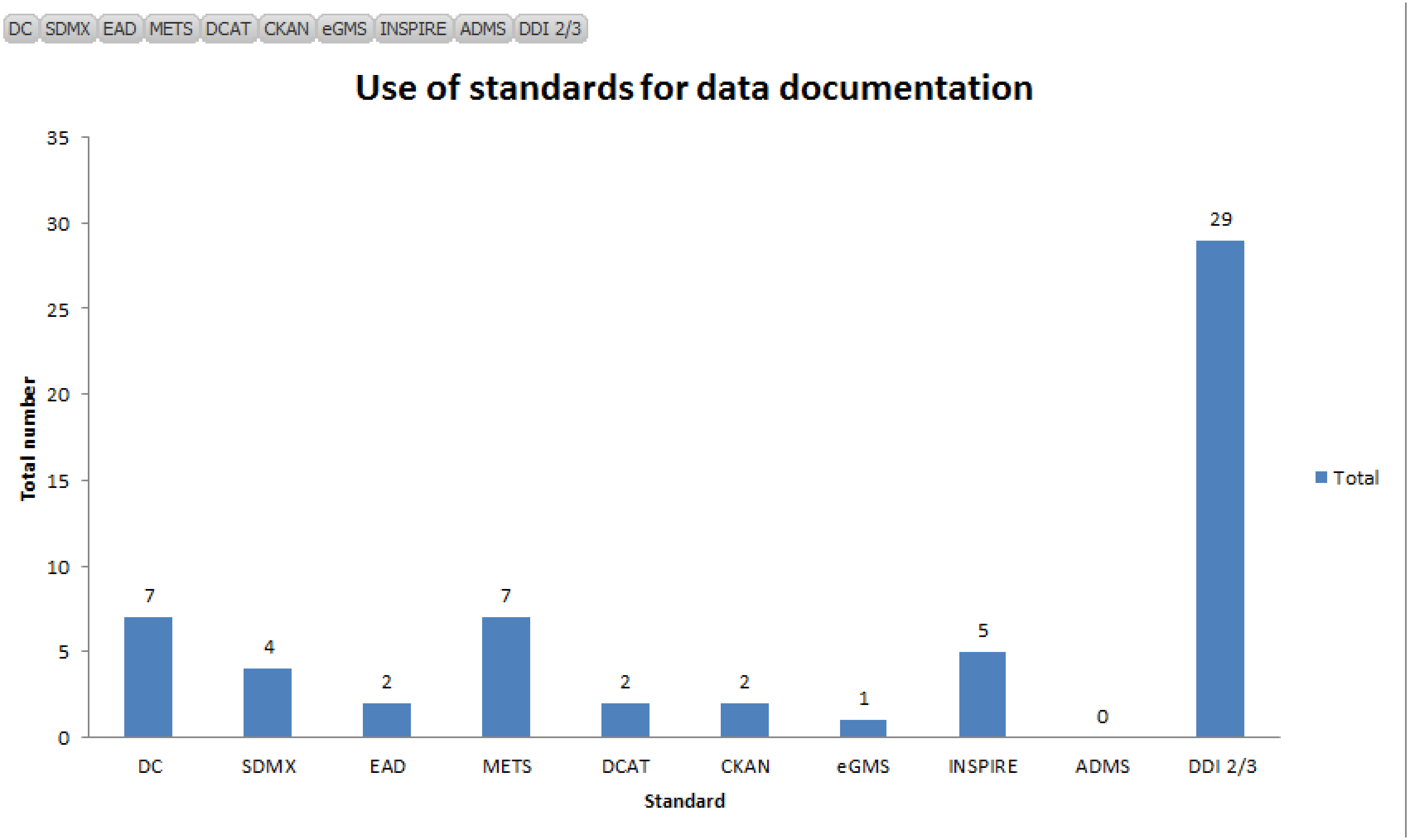
Activities: Review of Significant Data Sets

- 49 significant data sets were identified
- 13 were randomly selected for an in-depth analysis of current practices around data discoverability
- Very wide range of findings – there is no “typical” data set in terms of discoverability
  - Some (archives) were very good
  - Some (projects) were very poor

Activities: Online Survey

- Goal: 100+ responses (200 ideal)
- We got 253!
- Single survey with optional responses
  - Some questions not relevant for some groups of respondents
- Open-ended “sample”
  - Used existing online groups to recruit respondents
  - Good response from researchers and data producers especially
- Achieved a global scope

DDI Rules!


DC
SDMX
EAD
METS
DCAT
CKAN
eGMS
INSPIRE
ADMS
DDI 2/3
Use of standards for data documentation
Total number
35
30
25
20
15
10
5
0
7
4
2
7
2
2
1
5
0
29
DC
SDMX
EAD
METS
DCAT
CKAN
eGMS
INSPIRE
ADMS
DDI 2/3
Standard
Total

Activities: Focus Groups

- Not as many participants as desired
- Very detailed interviews
- Virtual interviews with researchers in the developing world

Focus Groups: Lesson Learned

- Discoverability is not the primary challenge – usability is!
- The burden of supplying data to secondary users rests with PIs and research teams
- This is *not* their primary mission!
- Unlike other domains, there is no infrastructure for disseminating data and supporting reuse of data
- Infrastructure poses challenges in the developing world, but there are solutions through hosting

Technology Approaches Review (Examples)

- Several interesting models
  - Some within Public Health, some from outside
- Centralized Data Portals
  - International Household Survey Network (IHSN)
    - Official Statistics
    - 90+ agencies
    - Centralized portal approach
  - Data without Boundaries/CESSDA Portal
    - Social and economic research
    - European scope
    - Centralized portal approach
    - Multi-lingual “controlled vocabularies”
- Data journal model (Ubiquity Press)
  - Emphasis on data citation
- Linked Data on the Web
  - Open government and Public Health

Project Data Journal

- A data journal was created as part of the project
- Hosted by Ubiquity Press
- Data articles were written by those producing important data sets
- Data journal is ongoing – not merely a product of this project

Options

- Technology approaches all based on the same set of information
  - We need good detailed information about our data!
  - There are standards we can build on in other domains (DDI especially)
  - Information must be machine-actionable
- Options are not mutually exclusive
- Data journal approach is more powerful when combined with good standard metadata
- Challenge is one of culture and best practice

Options: Good Information Standards/Best Practices

- Best practices are needed regarding what information is captured regarding data production
  - Need tools for researchers
  - Need to educate researchers and data producers
- Other domains have built on/extended the DDI standard especially
  - Environmental sciences provide a good example
  - Social and economic communities have developed tools and practices which can be leveraged
- No need to start from scratch!
- Emphasis is on what funders require of data producers to facilitate cultural change

Options: Research Support Infrastructure

- Leverage existing organizations/approaches
  - MRC Gateway is a good example
- Centralized portal for domain-wide search
  - Controlled vocabularies are a challenge
  - Multi-lingual support is needed
  - PIs/research teams must be supplemented by dedicated data professionals
- Data journals and data citation is good
  - Can be enhanced by having better, standard metadata
  - Variable-level information is needed for data description
- Linked Data on the Web model is intriguing
  - Could be useful as a long-term strategy
  - Not easy to predict or control
  - Much activity in pharmaceutical research and some other areas

Recommendations (1)

- Focus on the creation of a centralized domain portal for public health and epidemiology research, taking the following steps:
  - 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.
  - 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.
  - Establish guidelines and best practices for the use of technical standards and protocols for exposing data holdings to the domain portal.

Recommendations (2)

- 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.

- 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.

- 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.

Recommendations (3)

- 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).

- 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.

Next Steps

- Wellcome Trust is currently considering a second phase of the project
- Would involve developing tools, guidelines, and training materials for researchers, data managers, and other in public health research
- Emphasis on making the recommendations more real, and showing what is needed in a practical way

Conclusions

- The Wellcome Trust project identified many of the same issues, requirements, and solutions as the other projects in the social, behavioral, and economic research domain
  - Public health research is less mature as a domain
- Detailed metadata and data management is critical to both data discovery and reusability
- Centralized infrastructure approaches based on standards appear to be best practice

Thank You!