Library Lab - Who is Using Data and Why?

Mercè Crosas and Gustavo Durand (Institute for Quantitative Social Science (IQSS)), Diane Sredl[1](Government Information and Data Services, Lamont Library)

Changes:
In “Summary” and “Introduction”, we’ve added reasons why the project overall is important to the Library as a solution for data management plans required by funding agencies, and in particular why the proposed tool is critical for completeness of the data management plan and seek continued funding.
In Step 2 of “Design and Implementation”, we’ve added information about privacy policies to protect users, and clarify that the data will only be reported for further analysis in an aggregated, deidentified form.

Summary:

We propose to build a tool to allow tracking user information when users download or analyze research data from the Dataverse Network. The Dataverse Network is becoming the data management plan (DMP) solution for many researchers across Harvard, and the Library can offer it (and in some cases is already doing so) as one of its DMP options. To seek continued funding, funding agencies are asking for information about data usage. The proposed tool will offer data owners and collection administrators a way to configure what they would like to learn about usage of their data. Once configured, users will be asked to provide information about themselves (e.g. are they students, professor, other? what discipline or department they work with?), and about how they plan to use the data. The aggregated information will also be available to the repository administrators (Data Librarians, when applies) to generate statistics on data usage, and help them improve and continue funding their data services.

Introduction

In recent years, researchers have been more open and required to share data[2]. Although it has been a slow growth, we’ve seen an increase in the number of researchers uploading data in public repositories and choosing to make them accessible to others. In addition, current requirements for a data management plan from funding agencies (NSF and NIH) and journals are building awareness of the importance of preserving data and making them accessible in the long term. This is a big step for data libraries and data archives. Research data that were often lost in individual computers now must be stored in well-supported repositories and all necessary steps must be taken to make them accessible years from now. The Dataverse Network developed at IQSS has already been used extensively as a data management plan solution for social science data (see Harvard Gazette: http://news.harvard.edu/gazette/story/2011/09/

[1] Proposal also supported by Katherine McNeill, Social Science Data Services Librarian at MIT.
[2] Projects like the Dataverse Network and Dryad have seen a steady increase of data uploads from researchers since they were created. NSF requires a Data Management plan since 2011. A number of domain specific data repositories have emerged within the last years.

data-may-not-compute/). In addition, IQSS is working with a variety of groups across Harvard to accommodate data from other domains - with the Wolbach Library and the Astronomy department to host an Astronomy Dataverse Network (http://dvn.theastrodata.org/); with Mass General Bioinformatics to host a life science Dataverse Network for biology and medical data; with Humanists (History faculty) to expand the IQSS Dataverse Network to support qualitative data; etc. This makes the Dataverse Network an attractive offer for the library as a data management plan option for moderate size data files (< 2GB), at a time when Libraries across the country confront the issue of helping researchers with a data management plan.

We all embrace this increase in data openness and accessibility. The Library benefits by providing over time more replication and/or new data for students and scientists, and researchers benefit by allowing others to use their data, getting more citations to their work and advancing their line of research [see http://www.plosone.org/article/fetchArticle.action?articleURI=info%3Adoi%2F10.1371%2Fjournal.pone.0000308]. However, we are still in the early stages of data sharing and well-defined data management plans, and we still need to show future researchers, funding agencies and journals the full benefits of sharing and preserving data. To seek continued funding, researchers need to provide information to funding agencies about who is using their data: Number of distinct users? Are they students, faculty or external to the University? What are they using the data for? For the Library it is important that the data management plan solution provides this type of information to the data owners, so the data owners can provide it to the funding agencies.

We propose to build a tool that will work with the Dataverse Network to allow Dataverse administrators to configure what they are interested in asking users when users are ready to download or analyze data. The tool will be designed with and made available to the Harvard and MIT data librarians as an option to enhance their services.

Design and Implementation

The proposed usage tracking tool - or guestbook for data download - will be implemented as an extension of the Dataverse Network software, and will be designed and implemented following open source guidelines to allow the code to be easily usable by other systems. The code will mostly use Java and Javascript. It will have three main user interfaces:

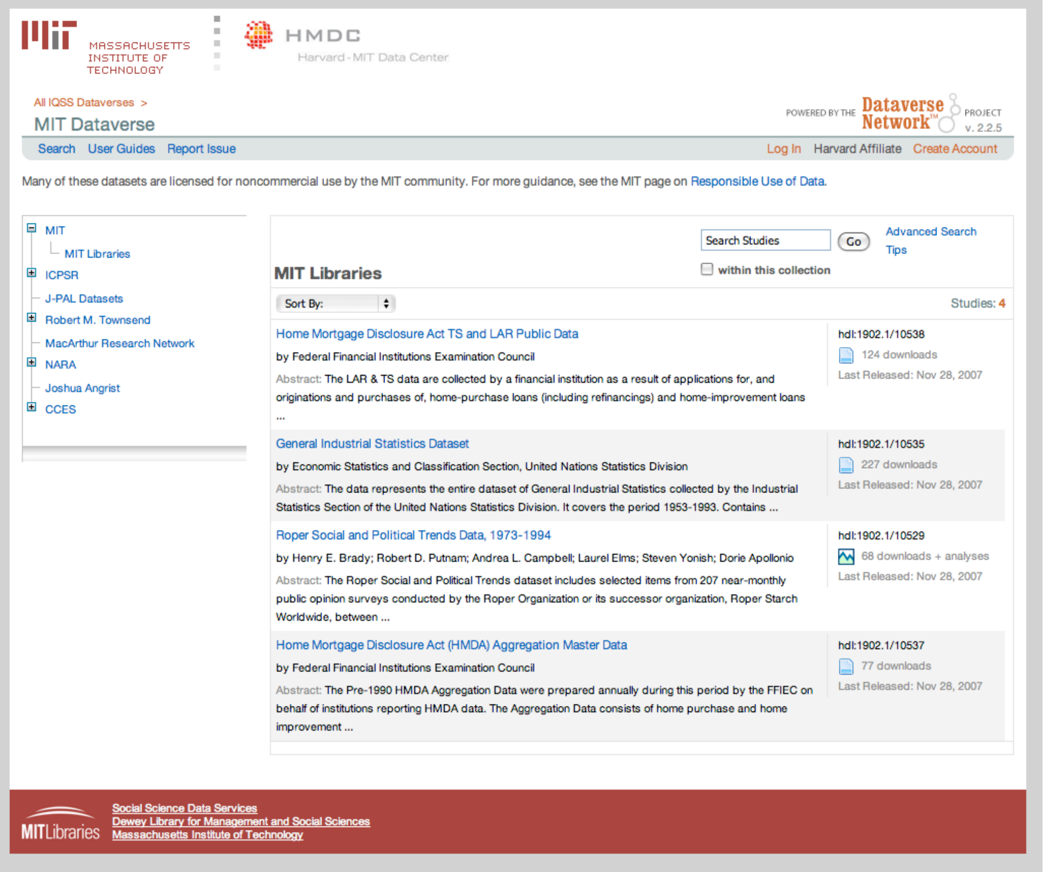
1. A web interface for administrators to be able to configure whether or not they want to ask for user information before downloading or analyzing the data. This interface will allow administrators to set up the fields or questions, whether the question is optional or required and the type of value expected.
2. A web interface to display the fields and questions to the users when they agree to terms of use, right before they download or view the data.
3. Another administrative web interface to view the information gathered from downloads and provide useful statistics reports.

Use Case/Workflow

Initial State:

Harvard and MIT libraries' data services offer access to research data through the Dataverse Network with Dataverses like the one in Figure 1.

Libraries' data services also offer their affiliated researchers or data owners the option to create their own Dataverse. Then the library can choose to add a link in the library Dataverse which points to the data uploaded by the researcher into another Dataverse.



MASSACHUSETTS INSTITUTE OF TECHNOLOGY
HMDC
Harvard-MIT Data Center
POWERED BY THE Dataverse Network™ PROJECT v. 2.2.5

All IQSS Dataverses >

MIT Dataverse

Search User Guides Report Issue — Log In Harvard Affiliate Create Account

Many of these datasets are licensed for noncommercial use by the MIT community. For more guidance, see the MIT page on Responsible Use of Data.

- MIT
  - MIT Libraries
- ICPSR
- J-PAL Datasets
- Robert M. Townsend
- MacArthur Research Network
- NARA
- Joshua Angrist
- CCES

MIT Libraries

Search Studies [Go] Advanced Search Tips

□ within this collection

Sort By: Studies: 4

Home Mortgage Disclosure Act TS and LAR Public Data
by Federal Financial Institutions Examination Council
Abstract: The LAR & TS data are collected by a financial institution as a result of applications for, and originations and purchases of, home-purchase loans (including refinancings) and home-improvement loans ...
hdl:1902.1/10538 — 124 downloads — Last Released: Nov 28, 2007

General Industrial Statistics Dataset
by Economic Statistics and Classification Section, United Nations Statistics Division
Abstract: The data represents the entire dataset of General Industrial Statistics collected by the Industrial Statistics Section of the United Nations Statistics Division. It covers the period 1953-1993. Contains ...
hdl:1902.1/10535 — 227 downloads — Last Released: Nov 28, 2007

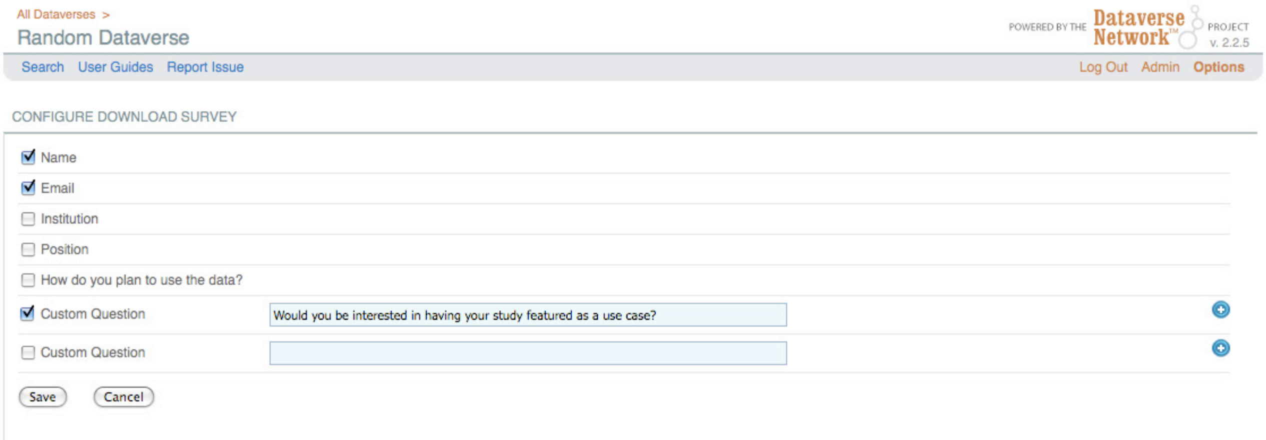
Roper Social and Political Trends Data, 1973-1994
by Henry E. Brady; Robert D. Putnam; Andrea L. Campbell; Laurel Elms; Steven Yonish; Dorie Apollonio
Abstract: The Roper Social and Political Trends dataset includes selected items from 207 near-monthly public opinion surveys conducted by the Roper Organization or its successor organization, Roper Starch Worldwide, between ...
hdl:1902.1/10529 — 68 downloads + analyses — Last Released: Nov 28, 2007

Home Mortgage Disclosure Act (HMDA) Aggregation Master Data
by Federal Financial Institutions Examination Council
Abstract: The Pre-1990 HMDA Aggregation Data were prepared annually during this period by the FFIEC on behalf of institutions reporting HMDA data. The Aggregation Data consists of home purchase and home improvement ...
hdl:1902.1/10537 — 77 downloads — Last Released: Nov 28, 2007

MITLibraries — Social Science Data Services / Dewey Library for Management and Social Sciences / Massachusetts Institute of Technology

Figure 1. MIT Dataverse

Step 1. Data owners, in addition to selecting whether they want to make data public or restricted, will configure what fields and questions (if any) they want to ask to the user upon downloading or viewing the data.



All Dataverses >

Random Dataverse

Search User Guides Report Issue — Log Out Admin Options

CONFIGURE DOWNLOAD SURVEY

☑ Name

☑ Email

□ Institution

□ Position

□ How do you plan to use the data?

☑ Custom Question: Would you be interested in having your study featured as a use case?

□ Custom Question:

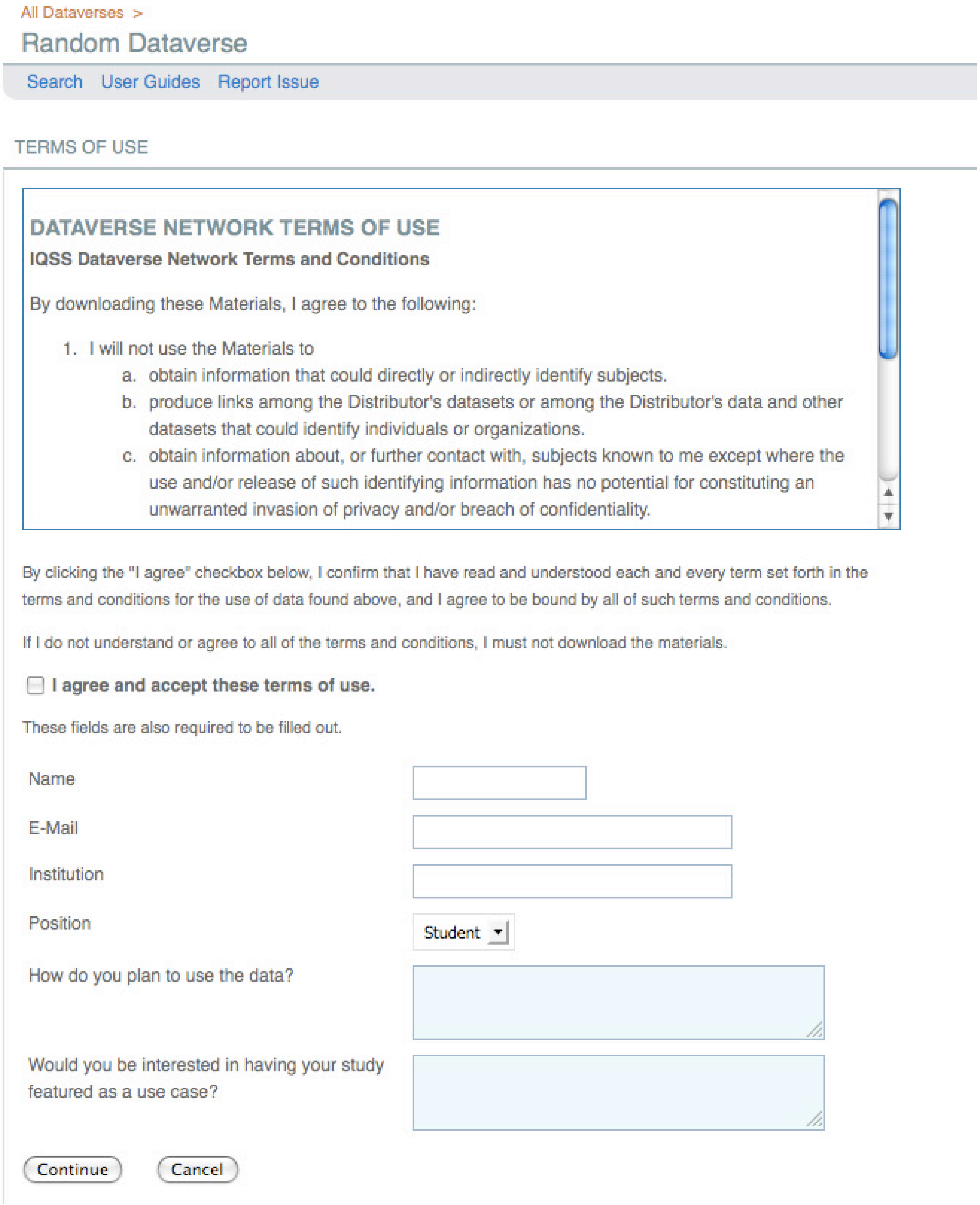
[Save] [Cancel]

Figure 2. User Interface to configure fields and questions for users using data.

Step 2. Students and researchers find a study of interest and choose to download the data.

They will be prompted with the terms of use defined for that data plus a set of questions configured by the dataverse administrator, who could be the data owner or librarian. The Dataverse will defined and make easily available a privacy policy that will state that identifiable information will not be shared or misused (to be compliant with http://www.ala.org/ala/aboutala/offices/oif/ifissues/issuesrelatedlinks/alaprivacypolicies.cfm)

For analysis of data usage, we will only use aggregates of data, which has been properly deidentified.



TERMS OF USE

DATAVERSE NETWORK TERMS OF USE

IQSS Dataverse Network Terms and Conditions

By downloading these Materials, I agree to the following:

1. I will not use the Materials to
    a. obtain information that could directly or indirectly identify subjects.
    b. produce links among the Distributor's datasets or among the Distributor's data and other datasets that could identify individuals or organizations.
    c. obtain information about, or further contact with, subjects known to me except where the use and/or release of such identifying information has no potential for constituting an unwarranted invasion of privacy and/or breach of confidentiality.

By clicking the "I agree" checkbox below, I confirm that I have read and understood each and every term set forth in the terms and conditions for the use of data found above, and I agree to be bound by all of such terms and conditions.

If I do not understand or agree to all of the terms and conditions, I must not download the materials.

☐ **I agree and accept these terms of use.**

These fields are also required to be filled out.

Name

E-Mail

Institution

Position: Student

How do you plan to use the data?

Would you be interested in having your study featured as a use case?

Figure 3. Terms of Use page with “guestbook” sign in and questionnaire.

Step 3. Dataverse administrators will be able to review the information about data downloads and usage. This will allow them to generate reports and statistics which will give them an insight to who is interested in the data they are providing and how the data is being used. (Proposed UI for reports and statistics not provided)

Such information has an impact into funding for data owners and archivists. Funding agencies are extremely interested in learning whether or not the data are being used and how, and in evaluating the results of a good data management and accessibility plan. (As an example, the Harvard Election Archive project has been able to get funding by showing the number of downloads of the data posted in their Dataverse (http://projects.iq.harvard.edu/eda/data). The researchers from this project have been asked to provide additional information on what type of users - students, faculty, etc - use their data and why. Gathering and reporting that information will help them obtained additional funding.)

Budget and Timing

In order to design, implement, test and evaluate the feedback for this new tool, we ask for:

Project Time: Total: 6 months
1/2 month - Final requirements gathering, design and architecture
2 ½ months - Implementation
1 month -Testing
2 months - Feedback and Evaluation from users, researchers, data owners and data librarians. Iterations incorporating feedback.

Funding Requested
4 months FTE developer
Data Librarians will be involved in initial requirements gathering and review, and then the testing and feedback phase.