# Making Sense of the Data: Analyzing RFI Responses on Data Management Educational Resources

1/30/2015
Ariel Deardorff, NLM Associate Fellow
National Library of Medicine

**Project Sponsor:**
Valerie Florance, Director, NLM Extramural Programs

## Table of Contents

## Abstract

### Objective

The goal of this project was to analyze and organize the results from a Request for Information (RFI) released by the National Library of Medicine as part of the NIH Big Data to Knowledge Initiative (BD2K). The specific objective was to create a method of organizing the results, write a summary report, and create a series of recommendations for further action.

### Methods

The first stage of the project involved creating a method to organize email responses from the RFI. The responses were then analyzed and summarized, and recommendations were crafted on how best to share the results with the public.

### Results

The Request for Information resulted in 16 responses detailing over 205 online and in-person courses, tutorials, guides, and MOOCs from over 84 institutions and organizations.

### Conclusion

The RFI responses represented a wide variety of educational resource types and topics and are worth compiling into a searchable index of educational resources so that they can be shared with the public. This aligns well with the proposed educational resource discovery index (EruDIte) to be located at the new BD2K Biomedical Training Coordination Center.

# Introduction

### Extramural Programs

Extramural Programs (EP) is a division of the National Library of Medicine (NLM) that manages grants for research projects and training in biomedical informatics. As the grants division of NLM, EP is often involved in wider NIH grant and extramural research programs, including those associated with the Big Data to Knowledge Initiative (BD2K).

### NIH Big Data to Knowledge Initiative (BD2K)

BD2K was launched in 2012 as a trans-NIH initiative to "enable biomedical research as a digital research enterprise, to facilitate discovery and support new knowledge, and to maximize community engagement."[1] The BD2K initiative has four major goals:

1. To facilitate broad use of biomedical data assets by making them more discoverable, accessible, and citable.
2. To conduct research and develop the methods, software, and tools needed to analyze biomedical big data.
3. To enhance training in the development and use of methods and tools necessary for biomedical big data science.
4. To support a data ecosystem that accelerates discovery as part of a digital enterprise[2]

Since its beginnings in 2012 BD2K has made funding available for a data discovery index, centers of excellence, and the creation of educational resources related to big data and data science.

### Request for Information

On November 2, 2014 the National Library of Medicine, in association with the NIH and as part of the BD2K initiative, issued a Request for Information (RFI) on resources for teaching and learning biomedical big data management and data science. The goal of the RFI was to "identify the array of timely, high quality courses and online learning materials already available on data science and data management topics for biomedical big data."[3] In order to achieve those aims the RFI sought information on courses, workshops, guides, resources, and MOOCs (massive open online courses) on the topics of data management, statistics, and computer science. The results of the RFI would inform future BD2K training grants and programs.

### Project Goals and Objectives

The overall goal of this project was to analyze the results of the RFI and make recommendations about how to use the gathered information. The specific objectives were to create a method of organizing the results, produce statistics on the responses, and create recommendations for the best way to make the information available to the public.

[1] "Mission Statement." *NIH Big Data to Knowledge*. n.d. 22 Jan. 2015. <http://bd2k.nih.gov/about_bd2k.html#sthash.Wly6kDKJ.dpbs>

[2] Same as above

[3] "Request for Information (RFI) on the NIH Big Data to Knowledge (BD2K) Initiative Resources for Teaching and Learning Biomedical Big Data Management and Data Science." *National Library of Medicine.* 4 Nov. 2015. 22 Jan. 2015. <http://grants.nih.gov/grants/guide/notice-files/NOT-LM-15-001.html>

## Methods

### Devising a System for Organizing RFI Results

The project began in early November, roughly around the same time the RFI was released. The first phase of the project, therefore, involved deciding how the results would be organized and reviewed. As the responses were in email form they first needed to be transferred into a spreadsheet in the easiest and fastest way possible. A data entry form was designed and used to enter the responses into an Excel spreadsheet (the challenges of this phase are covered in the Discussion section).

### Data Cleaning

Given that the data in question was drawn from free text in emails it was not surprising that some cleanup was necessary. While the fields that required selecting from a list (such as resource type or format) were uniform, the names of institutions and people required some normalizing so that different versions of an institution's name could be counted as one. Additionally, some of the resources had been assigned to the wrong resource type and had to be reassigned. This was understandable given that many responses included little more than the title, URL, and institution of a resource and it was necessary to make educated guesses about the other categories. Eventually, the spreadsheet was cleaned and purged of duplicate entries and ready for analysis.

### Analysis and Recommendations

The analysis phase of the project involved writing descriptive statistics of the RFI results in order to better understand what had been submitted. These numbers then fed into a set of recommendations detailing what to do with the information that had been gathered. The recommendations were compiled along with a summary of the RFI results and were presented to the BD2K Training Committee as part of their weekly teleconference (the summary report for the training committee can be found in Appendix A).

## Results

### RFI Responses

The Request for Information on data management and data science educational materials was released on November 4, 2014 and closed on December 31, 2014.

In total the RFI received:

**16 Responses**

**Detailing over 205 Online and In-Person Courses, Tutorials, Guides, and MOOCs**

**From 84+ Institutions and Organizations**

**Sample Resource Titles included: Data Management 101, Computing for Biomedical Scientists, Statistical Learning, Tackling the Challenges of Big Data, and Introduction to Data Science**

While the learning materials described in the RFI response included everything from curriculum support materials to online journals and certificate programs, the most frequently mentioned resources were MOOCs and college courses, followed by online tutorials and in-person courses and workshops (see Figure 1).

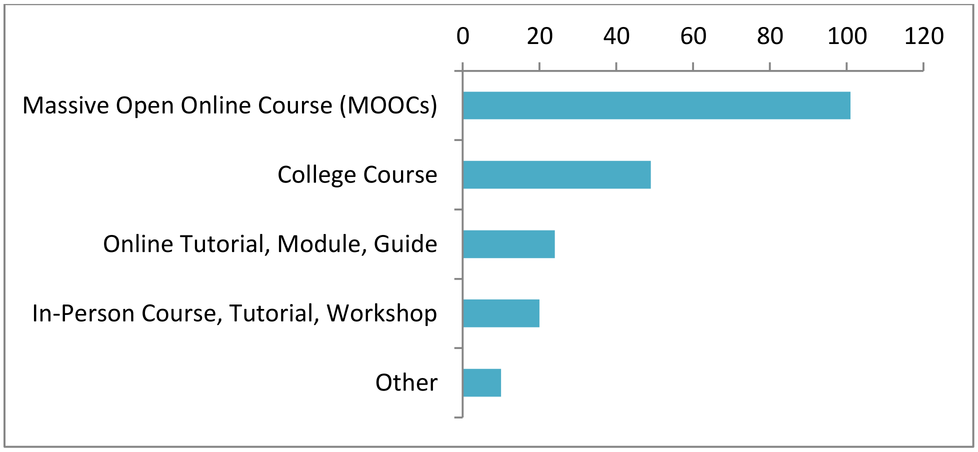


**Figure 1. RFI Response Types**[4]

[4] "Other" includes advanced degrees/certificates, curriculum materials, journal articles, online journals, and two white papers

Of the MOOCs that were submitted, the most popular platforms were Coursera, Udacity, EdX, and MIT's various platforms (see Figure 2).

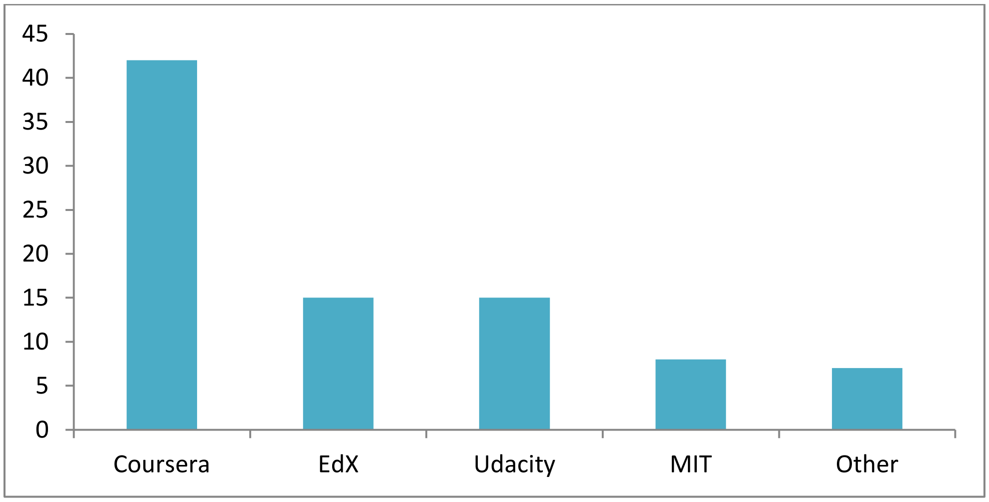


**Figure 2. MOOC Platforms**

While the topics were not always explicitly outlined in the response emails, a cursory glance at the resource's website revealed that the majority were on the topics of statistics and computer science, followed by data management and data science (see figure 3). This shows a breakdown between educational resources focused specifically on data management and data science and those that teach the tools or skills (including statistics and computer science) to work in those areas.

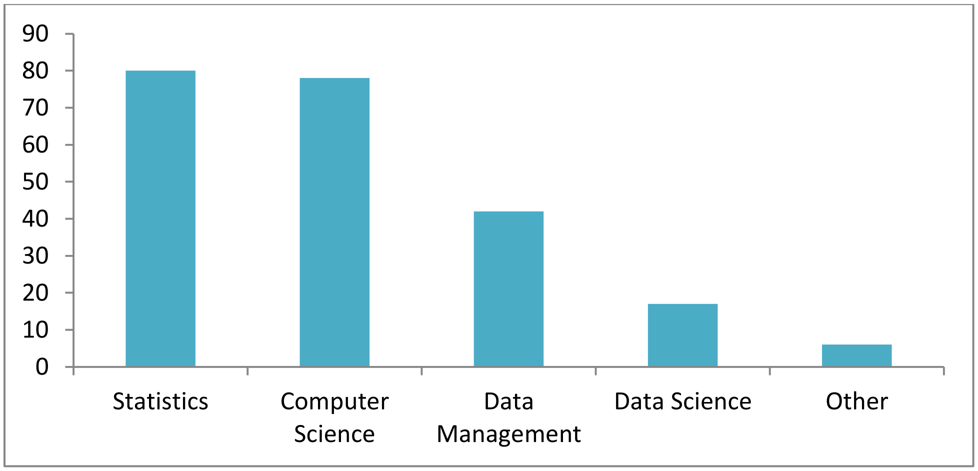


**Figure 3. Resource Topics**

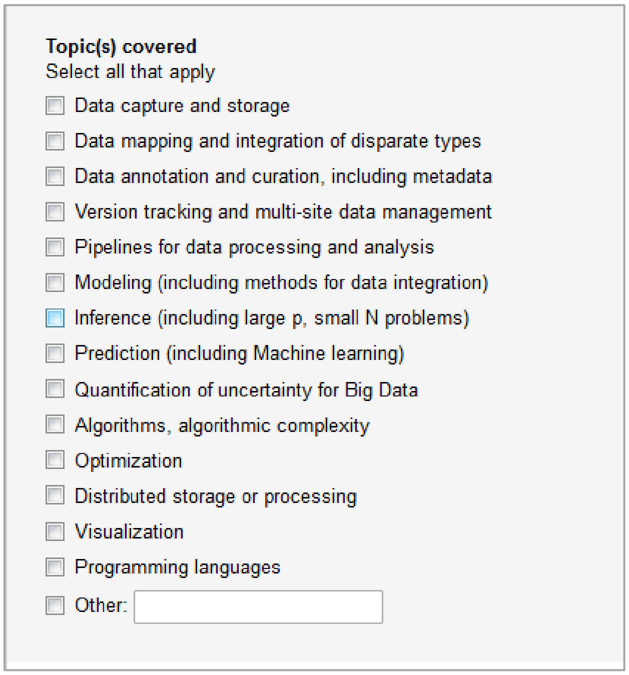
The learning resources described represented a wide variety of institutions and organizations. In addition to MIT, Johns Hopkins and Stanford University, there were also learning materials from Facebook, University of Oxford, and the University of Queensland Australia, representing the diversity of players involved in data management and data science.

## Discussion

### Data Management

It is interesting that a project based on gathering information about data management and data science resources had some rather tricky data management questions. The first phase of the project in particular was challenging because of the way the RFI was issued. Originally the designers of the RFI had created a Google form that would allow respondents to simply fill in information about the resources they were submitting, which would then be entered into a Google spreadsheet. Crucially, the form allowed respondents to select from a list of topics in order to give more detail about the resource they were submitting (see Figure 4). Unfortunately the NIH Policy Office thought that a form required too much work on the part of the submitter and asked the issuing team to change the method of response to email. This presented a challenge and required some additional steps to get the data in the proper format.

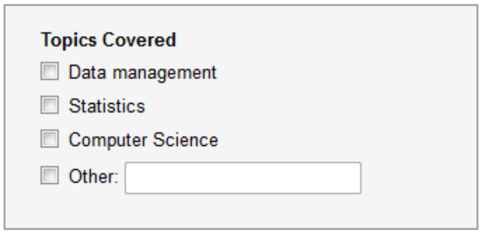
As a Google form had already been created with the appropriate parameters it seemed worthwhile to try and use it for the data entry side even though it couldn't be used by the respondents. The Google form was therefore edited so that it had only the information fields that respondents would likely include and fields that the data entry person could easily ascertain and select (including a much simpler list of topics, see Figure 5). The idea behind this was that it would be easier for the person doing data entry to click on different options (for example type of course or format) rather than having to retype it directly into a spreadsheet. It was hoped that this would also lead to better formatted entries as at least some of them would come automatically from the form.  Once the Google form was ready it was used to enter information from the incoming emails into a spreadsheet.



**Topic(s) covered**
Select all that apply

□ Data capture and storage
□ Data mapping and integration of disparate types
□ Data annotation and curation, including metadata
□ Version tracking and multi-site data management
□ Pipelines for data processing and analysis
□ Modeling (including methods for data integration)
□ Inference (including large p, small N problems)
□ Prediction (including Machine learning)
□ Quantification of uncertainty for Big Data
□ Algorithms, algorithmic complexity
□ Optimization
□ Distributed storage or processing
□ Visualization
□ Programming languages
□ Other: ____

**Figure 4. Topics on the Initial Google Form**



**Topics Covered**

□ Data management

□ Statistics

□ Computer Science

□ Other: ________

**Figure 5. Topics on the revised Google Form**

While this method of entering the data worked somewhat well it did require a large amount of data entry and the results were not as uniform as they could have been. Additionally it was somewhat challenging to track the various email responses to ensure that everything had been copied over to the spreadsheet. Given the available constraints, however, it was thought to be the best way of organizing and processing the data.

## Recommendations

### 1. Build a Data Management Educational Resource Index

The learning materials uncovered by the RFI represented valuable information that should be shared with the public. At the same time, it was clear that the learning resources mentioned failed to represent the entire catalog of available data management tools. It was therefore decided that the best way to share this information with the public would be through an interactive database or index of resources, which would allow students and practitioners to search for educational opportunities, and educators to post information about their courses and resources.

In order to be a useful tool the index would need to include at least the following information:

- Title of Resource
- Topic of Resource
- Format of Resource (MOOC, workshop, etc.)
- Target Audience (Librarians, Grad Students, Scientists, etc.)
- Sponsoring Institution
- Cost
- URL

In addition, the database might also include resource reviews or ratings. Ideally the database would allow users to search for courses on a wide variety of topic areas and for a wide variety of audiences.

The database described above closely mirrors the educational resource discovery index (EruDIte) that is a part of the current BD2K Funding Opportunity Announcement (FOA) for a Biomedical Training Coordination Center.[5] It is therefore recommended that the materials gathered as part of the RFI become part of the proposed index when it is built. The results will no doubt need to be reformatted and not all of them will necessarily need to be included (several referred to college courses that were only open to enrolled members of the institution, which was not really the type of resource sought) but they would be an interesting starting point and test group to build the database around. In the meantime, as an index will no doubt take some time to construct, the information should be released in the form of a spreadsheet so that the results are publically accessible.

### 2. Revisit the Idea of a Form to Manage RFI Responses

Requests for Information can have a wide variety of responses depending on the information they are seeking. Whereas one RFI might elicit paragraphs of text, others might receive links to resources. Because this particular RFI sought lists of educational materials it would have been easier to manage if respondents had been allowed to enter their responses directly into a form or spreadsheet such as the one originally designed for that purpose. This also would have improved the integrity of the data by allowing respondents to select from lists of topics and formats. While a form would have been difficult for respondents interested in submitting several resources at once, this could have been addressed with an option to send submissions via email as well. The Associate therefore recommends that similar RFIs in the future should revisit the idea of using a Google form and try to convince the NIH Policy Office that it will in fact lead to better data and results and will not cause undue hardship on respondents.

[5] "NIH Big Data to Knowledge (BD2K) Biomedical Data Science Training Coordination Center (U24)" *National Institutes of Health.* 18 Dec, 2014. 26 Jan. 2015. <http://grants.nih.gov/grants/guide/notice-files/NOT-LM-15-001.html>

## Acknowledgements

This project would not have been possible without the help of project sponsor Valerie Florance and project assistant Ebony Hughes. Special thanks also to Kathel Dunn, Wanda Whitney, and Maureen Madden for their feedback and support.

## Appendix A: RFI Results Summary

### Responses to the NIH Big Data to Knowledge (BD2K) Request for Information (RFI) on Resources for Teaching and Learning Biomedical Big Data Management and Data Science

### I. RFI Overview

The goal of this Request for Information (RFI) was to **identify the array of timely, high quality courses and online learning materials already available on data science and data management topics for biomedical big data**. In order to achieve those aims the RFI sought information on courses, workshops, guides, resources, and MOOCs (Massive Open Online Courses) on the topics of data management, statistics, and computer science. The RFI was released November 4, 2014 and closed December 31, 2014.

---

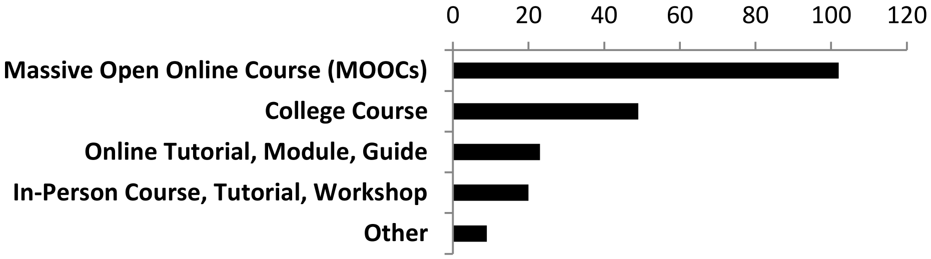
### II. Response Summary

In response to the RFI we received:

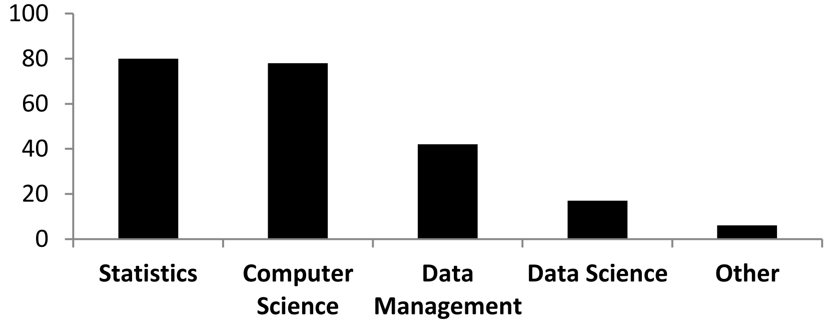
**16 Responses**

**Detailing over 205 Online and In-Person Courses, Tutorials, Guides, and MOOCs**

**From 84+ Institutions and Organizations**



**Resource Formats**



**Resource Topics**

**Sample Resource Titles: Data Management 101, Computing for Biomedical Scientists, Statistical Learning**

### III. Recommendations

Based on the large number of resources recorded **we recommend that the results of the RFI be submitted to the proposed BD2K Educational Resource Discovery Index (ERuDIte) at the future Biomedical Training Coordination Center**. This index would allow users to search for resources as well as submit new courses and resources to ensure that education related to data management and data science is accessible to all.