Data Management Toolkit

If you are reading this, chances are you already know that data management is a big topic. Recent changes in policy and research practice have increased the need for effective data-management services. Libraries are well positioned to serve many of these needs: they already have strong relationships with faculty and other stakeholders; they are able to advise on key concerns such as metadata, copyright, and preservation; and they have access to tools like institutional repositories that best showcase campus research. This toolkit provides a roadmap to better understand and start meeting data needs at your institution, including informational background and tools that can be downloaded for use.

I. Demystifying Data

In recent years, a number of forces have combined to transform the landscape for managing research data—some of the major concerns are listed below.

Why Manage and Share Data?

- **Funder Mandates:** Funders are requiring data-sharing and preservation policies. Federal agencies like the NSF, the NIH, and the NEH require that researchers make plans to share data as part of the application process.

- **Journal Policies:** Increasingly, journals are encouraging and even requiring data sharing as part the acceptance process.

- **Growth of Data:** A recent study by IBM showed that 90% of the data in the world has been generated in the last two years. Increased computing capacities will only lead to the rise of more data-driven projects. Given this atmosphere, organized, shared data is a valuable resource for researchers and the advance of science.

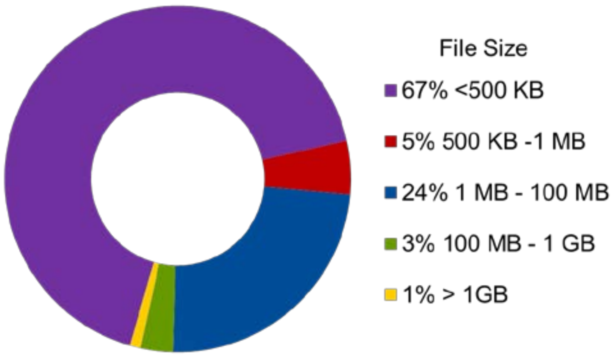
  Resources
  - Sherpa/Juliet database of research funders' open access policies
  - Funder Agency Data Guidelines compiled by University of Minnesota
  - Federal Funding Agencies and Data Mananagement Policies compiled by California Digital Library

What is Data?

Developing a data program is often intimidating—it can seem hard to get a handle on how big data files are, what they look like, and how they might fit in your repository. We did some research to examine what campus research data actually looks like and found that there is good news: most data files are just like most other files in your repository. Here are some of the questions we looked at:

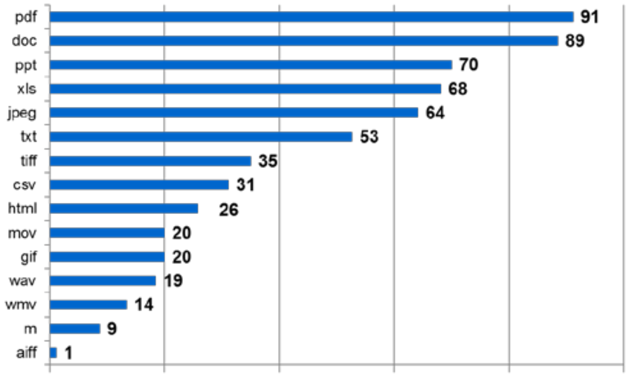
- **Data files are big; will my repository be able to handle the size demands of data?**

  - The majority of individual data files are quite small.
  - In a random sampling of data sets generated from leading data repositories (Figshare, Dataverse, Dryad), we found that 67% of data files were under 500 KB, and only 1 data set was actually over 1GB (at 1.3 GB).
  - Research that generates massive amounts of data in large file sizes will often already have storage solutions that are part of the research project. See Databib for a list of discipline-specific databases.



- **Data files are esoteric—–how can my repository support them?**

  In fact, the vast majority of data files are similar to most of the other materials already in your repository. Take a look, for example, at this survey conducted by Clemson University about research needs—most data files are not going to look that different from other files you already work with.



Percentages of respondents generating each file type. Respondents could choose multiple extensions.

CLEMSON
UNIVERSITY

II. Where to look for data on your campus

Every school will have a different landscape when it comes to data, and these needs will change over time. As you start your data program, you'll want to use a variety of strategies to look for data needs on your campus.

Use your Liaison Librarians

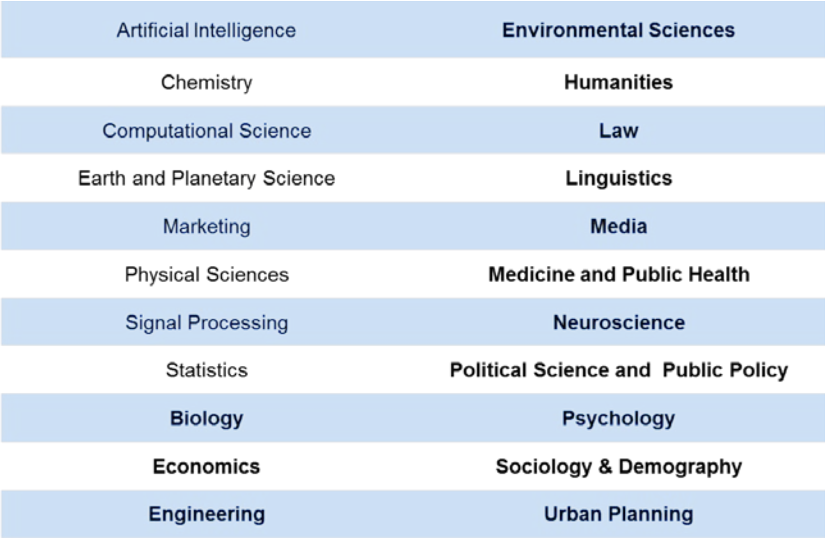
Finding data on your campus is very similar to finding other items that benefit from inclusion in the institutional repository. Your liaison and subject librarians are key resources to spread the word about your services. Start with liaisons who have particularly strong relationships with their departments.

Identify departments who have unmet data needs

The growth of computing technologies has increased data-driven research. While the most obvious place to start will be in the sciences, don't forget that there are many other disciplines and interdisciplinary groups that will have less-articulated standards and thus higher needs around data management. For example, the chart below shows how data-driven research has expanded to a variety of disciplines (shown in bold) at UC Berkeley over the last 20 years. If you are developing a list of departments to contact, this list is a great place to start!

Want to Get Started Now?

- Look at **Office of Research** publications for researcher profiles and lists of departments getting most funding.
- Talk to **IT** or Research Computing to see who has been contacting them about data needs.
- Work with **3 liaison librarians** with strong departmental relationships as pilots to find more about campus data needs.
- Publicize your data services in **author download reports** and other communications with faculty.



Artificial Intelligence	**Environmental Sciences**
Chemistry	**Humanities**
Computational Science	**Law**
Earth and Planetary Science	**Linguistics**
Marketing	**Media**
Physical Sciences	**Medicine and Public Health**
Signal Processing	**Neuroscience**
Statistics	**Political Science and Public Policy**
Biology	**Psychology**
Economics	**Sociology & Demography**
Engineering	**Urban Planning**

California, Winter 2013, vol. 124, No. 4

Contact other stakeholders

- **Office of Research**

The Office of Research can be your strongest ally in starting a data management program. Indeed, at some schools, data management is a combined service provided by the Library and Office of Research. The mission of the Office of Research is to showcase research done on campus, and in many cases they are also responsible for demonstrating compliance. They will know who is doing grant-funded research and may be looking for solutions. Look at their website and publications to find profiles of researchers working on grant-funded projects, lists of departments with high levels of government funding, and more.

- **Graduate Schools and ETD Administrators**

If you host Electronic Theses and Dissertations on your repository, you are probably already supporting many of the data sets associated with ETD's—formalizing your activity and talking to students about their research can be a great way to start your program.

- **IT Department / Office of Research Computing**

When researchers are working with large data sets and are thinking about questions around storage, they often go first to an IT department or the Office of Research Computing. Check in with departments that offer technology-support services to see if they have lists of students or faculty members with data-management needs.

- **Faculty Senate or other faculty groups on campus**

Attending faculty senate meetings and other faculty-groups meetings can also be a way to advocate for the important role the library plays in large, multi-layered research projects, and alert faculty to the services that you can provide.

Data Management Plans

In addition to helping support data sharing and preservation, you may also be helping campus researchers write data management plans for their grant applications. For most grants, this is a two-page document that covers many of the topics included in this toolkit. We've included some samples below. The DMP Tool is another great resource for researchers at this stage in the process.

Sample Plans

Sample Earth Sciences DMP

Sample Image File DMP

Sample NSF DMP

Discipline specific data repositories

Databib data repository directory

Data Repository list
OAD/Simmons College

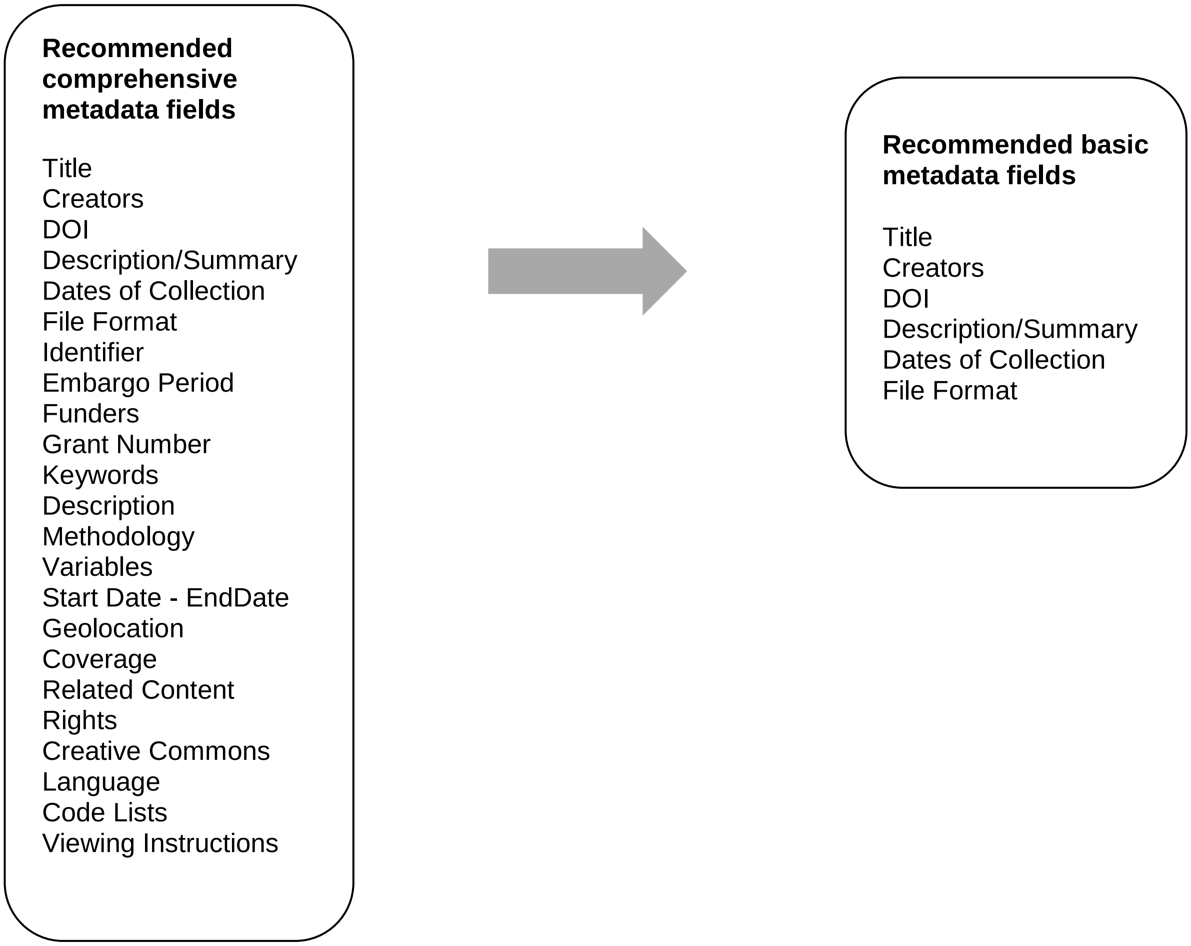
III. So you found some data—now what?

Now that you know of some researchers working with data, you can start offering them concrete services. Below we have included strategies and resources that you can adapt to start meeting data needs at your campus.

Metadata

You will encounter a wide variety of awareness about the need for metadata, and helping researchers capture the right kind of metadata for data sets is central to the advising process. Researchers should know that metadata captures the logic behind data collection. A strong metadata record allows the data to be discovered by others and also includes instructions that will enable effective reuse.

Standards continue to evolve, but bepress has developed a series of recommended metadata fields based on Dublin Core standards, which can be customized depending on need. Our Consulting Services staff can work with you, using the data interview, to ensure that the proper metadata is being captured for maximum impact and discoverability. And if metadata standards change, Digital Commons offers a lot of flexibility to make adjustments. You'll notice that the basic metadata for data is similar to other metadata records, just in the way that many data files are similar in size and format to others in your repository.



Recommended comprehensive metadata fields

Title
Creators
DOI
Description/Summary
Dates of Collection
File Format
Identifier
Embargo Period
Funders
Grant Number
Keywords
Description
Methodology
Variables
Start Date - EndDate
Geolocation
Coverage
Related Content
Rights
Creative Commons
Language
Code Lists
Viewing Instructions

Recommended basic metadata fields

Title
Creators
DOI
Description/Summary
Dates of Collection
File Format

Conduct a Data Interview

In order to best showcase data in your IR, you'll need to talk to researchers about their data. A structured interview can jumpstart this conversation by helping the researcher and librarian communicate more effectively and better understand the key issues around metadata capture and presentation of the data.

This data-interview template is a sample framework. As you conduct interviews, you'll be able to adapt these questions to your community's research needs.

Data Interview*

Metadata

- Describe your project.
- What is the title of the file?
- Who are the creators of the data?
- Does it have a DOI number or a unique identifier?
- Who are the funders?
- What is the grant number?
- What methodology did you use (tools, experimental protocol, lab notes)?
- Do you have any special instructions for viewing the data?
- What file format is your data in?
- What are the start and end dates of collection?
- Where was your data gathered?
- Is the data covered under a Creative Commons License?
- Who owns the rights to the data?
- When was/will the data be released? Are there embargo periods?
- Does the data require any specific access controls?

Presentation

- What other content should be linked to this data?
- Describe the relationship between the related content.
- What is the minimum bundle of data that someone would want to search or need to cite?
- Are there any additional files that you'd like to attach to this data?

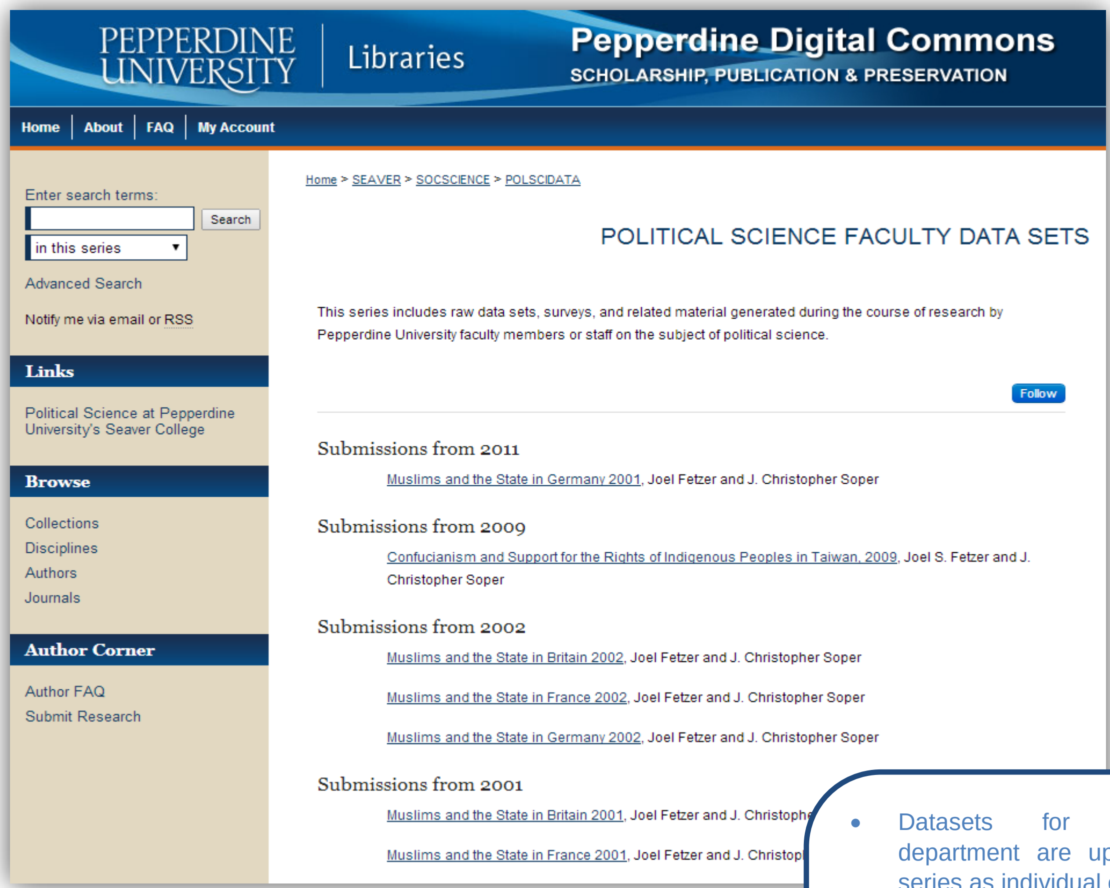
*Adapted from: Conducting the Data Interview" by Jake Carlson and Michael Witt http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1092&context=lib_research Purdue Data Curation Profiles directory http://docs.lib.purdue.edu/dcp/. University of Minnesota. UMass Amherst

IV. Using Digital Commons to Showcase Data Sets

Digital Commons offers a great deal of flexibility for organizing and featuring data sets. The **data interview** included in this toolkit can help you work with researchers to determine relevant information about their data. Once you've completed the data interview, send it to your bepress consultant and we'll work with you to determine the best way to capture and showcase it.

Showcasing data sets on Digital Commons

Example One: Departmental Data



Home > SEAVER > SOCSCIENCE > POLSCIDATA

POLITICAL SCIENCE FACULTY DATA SETS

This series includes raw data sets, surveys, and related material generated during the course of research by Pepperdine University faculty members or staff on the subject of political science.

Submissions from 2011

Muslims and the State in Germany 2001, Joel Fetzer and J. Christopher Soper

Submissions from 2009

Confucianism and Support for the Rights of Indigenous Peoples in Taiwan, 2009, Joel S. Fetzer and J. Christopher Soper

Submissions from 2002

Muslims and the State in Britain 2002, Joel Fetzer and J. Christopher Soper

Muslims and the State in France 2002, Joel Fetzer and J. Christopher Soper

Muslims and the State in Germany 2002, Joel Fetzer and J. Christopher Soper

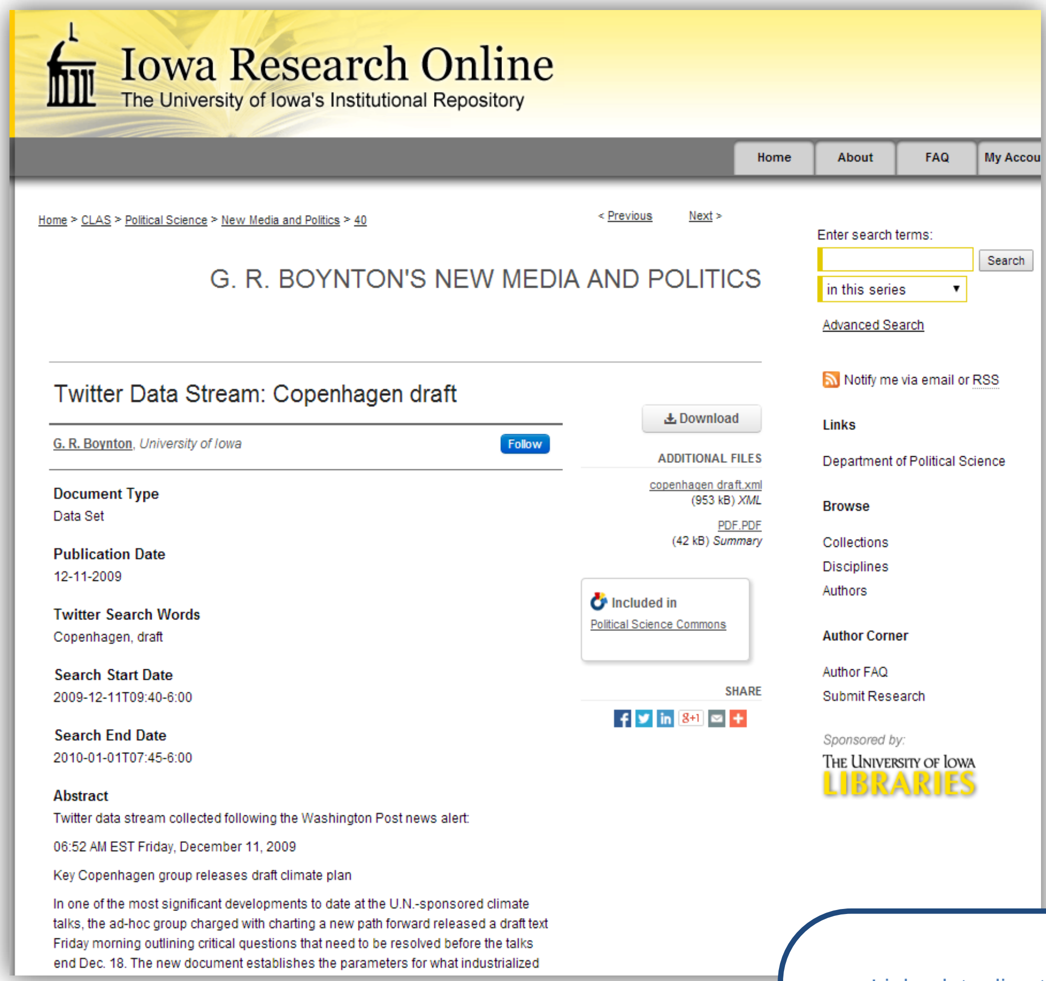
Submissions from 2001

Muslims and the State in Britain 2001, Joel Fetzer and J. Christophe

Muslims and the State in France 2001, Joel Fetzer and J. Christop

- Datasets for an entire department are uploaded to a series as individual objects.
- Metadata customized for that discipline.
- Data is directly citable.
- Can use embargoes or other means for access control.
- Consulting Services can help you use collection tool to have data appear in one or more relevant series.

Example Two: Data as Supplemental Files



Home > CLAS > Political Science > New Media and Politics > 40

< Previous Next >

G. R. BOYNTON'S NEW MEDIA AND POLITICS

Twitter Data Stream: Copenhagen draft

G. R. Boynton, *University of Iowa*

Document Type
Data Set

Publication Date
12-11-2009

Twitter Search Words
Copenhagen, draft

Search Start Date
2009-12-11T09:40-6:00

Search End Date
2010-01-01T07:45-6:00

Abstract
Twitter data stream collected following the Washington Post news alert:

06:52 AM EST Friday, December 11, 2009

Key Copenhagen group releases draft climate plan

In one of the most significant developments to date at the U.N.-sponsored climate talks, the ad-hoc group charged with charting a new path forward released a draft text Friday morning outlining critical questions that need to be resolved before the talks end Dec. 18. The new document establishes the parameters for what industrialized

ADDITIONAL FILES

copenhagen draft.xml
(953 kB) *XML*

PDF.PDF
(42 kB) *Summary*

Included in
Political Science Commons

Enter search terms:

Advanced Search

Notify me via email or RSS

Links

Department of Political Science

Browse

Collections
Disciplines
Authors

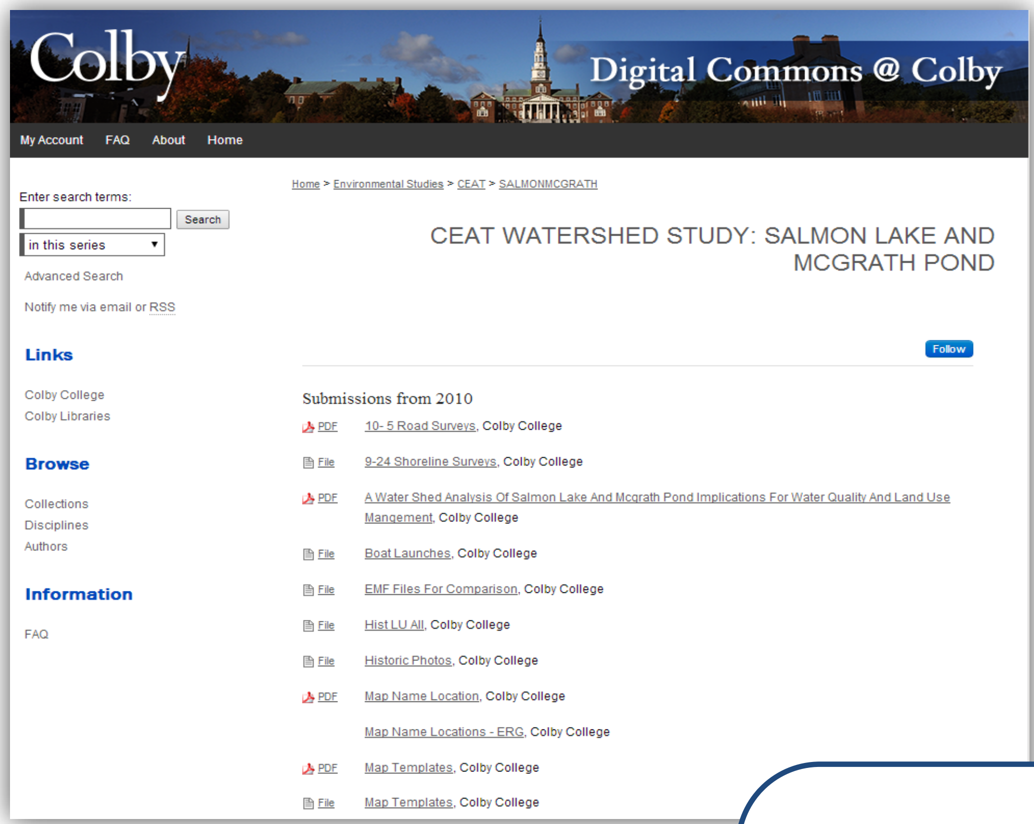
Author Corner

Author FAQ
Submit Research

Sponsored by:
The University of Iowa
LIBRARIES

- Links data directly to the related article that cites it.
- Data shares metadata with primary article.
- Supplemental files can be added to any type of record or publication type.
- Good choice when data set does not need separate citation.

Example Three: Project-Specific Data



Home > Environmental Studies > CEAT > SALMONMCGRATH

CEAT WATERSHED STUDY: SALMON LAKE AND MCGRATH POND

Submissions from 2010

PDF 10- 5 Road Surveys, Colby College

File 9-24 Shoreline Surveys, Colby College

PDF A Water Shed Analysis Of Salmon Lake And Mcgrath Pond Implications For Water Quality And Land Use Mangement, Colby College

File Boat Launches, Colby College

File EMF Files For Comparison, Colby College

File Hist LU All, Colby College

File Historic Photos, Colby College

PDF Map Name Location, Colby College

Map Name Locations - ERG, Colby College

PDF Map Templates, Colby College

File Map Templates, Colby College

- Allows data and other research products to be linked together.
- Metadata for project can be customized.
- Data is directly citable.
- Can use embargoes or other means for access control.
- Consulting Services can help you use collection tool to have data appear in one or more relevant series.

V. Questions?

Data is an exciting and evolving topic and we look forward to learning more about your experiences and needs for data in the future. Stay tuned to this site – we'll be offering more resources, links and tools that you can use to support data needs at your school.

And feel free to contact us at outreach@bepress.com with any questions that you may have as you get started supporting research data on your campus.