Cornell University's

Research Data Management
Service Group

Summary of Activities and Accomplishments
July 2016 – June 2017

Contents

Cornell University

RESEARCH DATA MANAGEMENT SERVICE GROUP

Comprehensive Data Management Planning & Services

Find solutions to store and share data.
Learn about best practices.
Write successful data management plans.
Consult with experts.

✉ rdmsg-help@cornell.edu
🌐 data.research.cornell.edu

Introduction

After discussions in early 2008 addressing issues surrounding data discovery, re-use and curation, and a two year project funding by the Vice Provost's office to build campus-wide collaboration in support of data services, the Research Data Management Service Group (RDMSG) was founded in December of 2010.

The RDMSG is a cross-campus, collaborative organization that helps researchers find and implement the data management services they need, and provides assistance in the preparation of data management plans for research funding proposals. Current service providers include: the Center for Advanced Computing (CAC), Cornell Information Technologies (IT@Cornell), the Cornell Institute for Social and Economic Research (CISER), Cornell University Libraries (CUL), the Cornell Biotechnology Research Center (BRC), and Weill Cornell Medical College (WCMC) (see also appendix A). The group receives oversight by a Faculty Advisory Board (Appendix B), and activities and services are developed and coordinated by the RDMSG Management Council (Appendix C).

This report contains a summary of activities of the RDMSG, from July 2016 through June 2017. Two implementation teams, Outreach and Training (including efforts to maintain and update the RDMSG website) and Consultants were active in this time. This report includes summarizes their accomplishments, including the following highlights:

> The Outreach and Training team held a variety of workshops and information sessions, introducing faculty, staff and students to the RDMSG at 5 campus events, teaching in 4 departmental courses or seminars, and holding 16 specialized data management-related classes or workshops.
>
> RDMSG Consultants received 114 requests for information, resulting in 89 consultations where we helped faculty, staff and students with their data management needs. Of those consultations, 36 were inquiries directly related to proposal data management plans, and another 59 were requests for general data management assistance or information. The consultant pool has grown to include 16 individuals with a wide variety of data expertise.
>
> The RDMSG website (http://data.research.cornell.edu) had a design overhaul this year, with both updated (data management planning) and new content (support for life science researchers).

RDMSG Outreach and Training Team

Team members: Chris Fournier (Mar 2015 – October 2016), Erica Johns (Sept 2014 – present, chair January 2016 - present), Wendy Kozlowski (Dec 2011 - present), Sarah Wright (Apr 2011 - present).

Charge: The RDMSG Outreach and Training (O&T) team develops and delivers training materials for Cornell researchers on issues of data management, with particular reference to meeting the requirements of research funders, and to services available through RDMSG providers. The Outreach and Training team also has primary responsibility for marketing RDMSG services to the Cornell community, and strives to fulfill its charge by recommending best practices while also meeting the practical needs of Cornell's researchers.

Scope: The team's target audience is researchers at Cornell, including faculty, students and staff. The RDMSG Outreach and Training team is not charged with training RDMSG participants (consultants, service providers) or with handling communication tasks of the RDMSG such as ongoing communication with (or recruitment of new) RDMSG service providers.

Overview of Major Activities

The Outreach and Training team participated in 25 specialized events where people with a particular need or of a certain user group would be likely to attend. Table 1 details these activities.

Date	Event	RDMSG personnel	Attendance	Description
06 Sep 2016	Inside Scoop	Erica Johns, Dianne Dietrich	18	Grad-student introduction to RDMSG
08 Sep 2016	Introduction to Data Management	Wendy Kozlowski, Erica Johns	8	Workshop
20 Sep 2016	Excel for Research Data: Formulas and Functions	Wendy Kozlowski	20	Workshop
29 Sep 2016	Data Documentation	Wendy Kozlowski, Erica Johns	4	Workshop
04 Oct 2016	Excel for Research Data: Pivot Tables	Wendy Kozlowski	13	Workshop
21 Oct 2016	Data Curation Network Research Engagement Event	Wendy Kozlowski, Erica Johns	19	Information gathering event
11 Jan 2017	Digital Archiving for Humanities Immersion Program	Wendy Kozlowski, Michelle Paolillo	14	Workshop
12 Jan 2017	Excel for Humanities Immersion Program	Wendy Kozlowski	15	Workshop
19 Jan 2017	Data Organization for Science Immersion Program	Wendy Kozlowski	15	Workshop
20 Jan 2017	Collaboration: OSF/ELN for Science Immersion Program	Wendy Kozlowski, Erica Johns	15	Workshop

23 Jan 2017	RDMSG Information Session	Wendy Kozlowski, Sarah Wright	5	General intro to DMPs and RDMSG information session
30 Jan – 13 Mar 2017	NTRES 6600 (Research Data Management Seminar)	Erica Johns, Wendy Kozlowski	50	1 credit course taught by EJ and Cliff Kraft (with WK). 6 total sessions: intro, relational databases, SQLite, data documentation, data sharing, DMP writing
07 Feb 2017	Excel for Research Data: Formulas and Functions	Wendy Kozlowski, Sarah Wright	10	Workshop
16 Feb 2017	RDMSG Information Session	Wendy Kozlowski, Erica Johns	4	General intro to DMPs and RDMSG information session
27 Feb 2017	Excel For Research Data: Pivot Tables	Wendy Kozlowski	12	Actuarial Society Workshop
07 Mar 2017	Excel for Research Data: Pivot Tables	Wendy Kozlowski, Sarah Wright	15	Workshop
20 Mar 2017	BIOG 3020: Research Skills for the Life Sciences	Sarah Wright	11	Two class sessions on intro to data management for undergraduates doing research on campus
21 Mar 2017	Excel for Research Data: Macros	Wendy Kozlowski, Sarah Wright	10	Workshop
24 Mar 2017	Data Management	Wendy Kozlowski	~20	MSE5720 Data Management Class (Nicole Benedek) Guest Lecture
13 April 2017	OpenRefine	Erica Johns	4	Workshop
26 Apr 2017	Intro to Excel	Wendy Kozlowski	21	Society for Women in Business Workshop
02 May 2017	Data Management	Wendy Kozlowski	8	ANTHR 6440 Research Design class (Marina Welker) Guest Lecture
10 May 2017	Open Science Framework	Wendy Kozlowski, Erica Johns	7	Talk/Demo about OSF to Open Science grad student group
15 May 2017	OSP panel "How to get your NSF proposal funded"	Sarah Wright	~100	Casual discussion about DMPs at OSP roundtable event
29 June 2017	Intro to Excel	Wendy Kozlowski	4	Summer Immersion program for the humanities

Table 1: Summary of Outreach and Training Activities, 2016-2017.

Walk-in Consultations: In an effort to keep RDMSG in the public eye and to make our services more visible, the walk-in consultation service that we began in January of 2016 continued throughout this year. The RDMSG website (http://data.research.cornell.edu/calendar) listed the weekly schedule and locations, and any exceptions to the regular schedule. We advertised these via LCD signs located in Mann Library, Duffield Atrium, Physical Sciences, Food Science, and the College of Veterinary Medicine. Consultants were stationed in public, visible areas, and displayed an RDMSG sign while on duty.

Participating consultants were Tarek Chams, Christine Fournier, Erica Johns, Wendy Kozlowski, and Sarah Wright. Approximately 16 hours of walk-in time were held each month. Monday, Tuesday, and Friday's office hours were held in the Mann Library consultation area, and Wednesday and Thursday's walk-in hours were at Uris Library. Due to a reduction in available consultants in January 2017, the team reduced office hours to Monday-Thursday. Walk in consultation hours were closed beginning in mid-June, though consultations were still offered on an on-request basis throughout the summer.

While we had some uptake with these walk-in sessions, the majority of our help requests come in via email and make appointments to talk to the consultants. We plan to discontinue walk-in office hours as of Fall of 2017, but we will continue to market our always-available consultation services.

Marketing and Communication: Members of the outreach team worked with CUL-IT designers to develop a new look for the RDMSG website, launching on September 22, 2016. At the same time, the design was added to a new business card, distributed throughout the libraries and to all the consultants (see report cover page for image). Specifics of the new website design and content can be found in the website section of this report.

Our marketing and communication efforts included web-based communication, fliers in the library, use of campus LCD screens and word of mouth. Web-based communication included news items on the website and Twitter, and email announcements. LCD signs to promote consultation hours, communicating with library liaisons to promote RDMSG referrals within their departments, presentations at the CUL Research and Outreach forum to raise awareness, tables at graduate orientation events, participation in OSP roundtables to inform research administrators of data management issues and the services RDMSG can provide, and presentations and publications beyond Cornell.

In addition, there has been direct communication with faculty about the needs for Cornell data services that accommodate Big Data; members of the Outreach team are surveying (via interviews) the academic landscape to determine what other universities offer academics in the way of a IT package to support computationally heavy or "big data" research.

Plans for 2017-2018

- Work with consultants to make sure that department websites refer to the RDMSG website as a resource for researchers; initial plans are to work with Tarek Chams and CALS, and expand the model to other departments.
- Continue to communicate with CUL liaisons to encourage referrals to RDMSG.
- Participate in monitoring groups for funders within the purview of the Office of Science and Technology Policy memo on open access to research outputs including publications and research data.
- Expand our data-related workshop series to include other relevant topics. Possible new subjects for these hands-on workshops include Excel charts, the Open Science Framework, and SQL.

RDMSG Website

The RDMSG website includes information for researchers about funding requirements, service providers, the RDMSG team, best practices for data management and other information as appropriate. The target audience includes faculty, staff and student researchers at Cornell and RDMSG consultants. The Website Maintenance team is a sub-team under Outreach and Training. It consists of Dianne Dietrich, Sarah Wright and Wendy Kozlowski, who take primary responsibility for updates, communication with Drupal designers, and scheduling of content updates with service providers.

Figure 1. Front page of the re-designed RDMSG website, launched Sept 2016.

The RDMSG website was re-designed in fall of 2016, with a new logo and re-organized, easy to find content. Website and Outreach Team members, with extensive contributions from James Vanee (BRC Consultant) launched a life science focused page and service spotlight on the RDMSG website. The Website and outreach teams completed two mini-sprints in October 2016 to redesign our data management planning page, streamlining the presentation and updating content to reflect current funder standards and recommendations.

Throughout the year, the Web Maintenance team remains focused on ensuring website content remains up to date, and that all links are functional and accurate. With the help of Manolo Perez, User Interface/Experience Specialist at CUL, the website team also makes small adjustments to website layout to enhance usability.

New and Updated Content Highlights

Readme Metadata Template. Based on feedback from both consultations and workshops over the last few years, we've noticed that metadata creation is an area that people seem have limited knowledge about, and in turn tend not to do in a formalized way. In an attempt to remedy this using a simple, non-machine readable but semi-standardized way, we added a new "best practices" page in November of 2016, that focused specifically on readme-style metadata (Figure 2). The page includes a description of how to create a readme document, and a downloadable template to use or modify. From time of upload to the end of June 2017, there were 233 downloads of that page, nearly once a day on average (29.1 downloads/month), from users around the world.
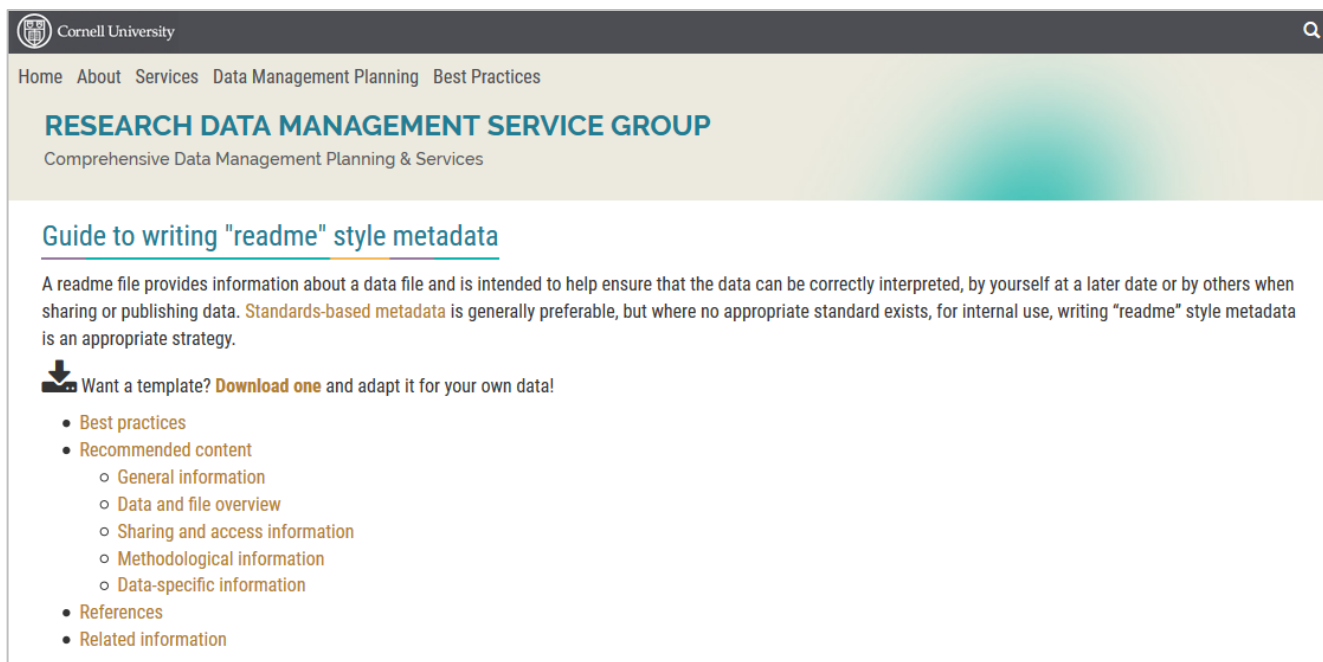

Figure 2: Readme metadata best practices page, with downloadable template for creating a text readme file.

Data Management Planning. When originally written, the website information about Data Management Planning was focused on requirements and formatting as directed by the National Science Foundation. In the years since, several other funder agencies have released their own guidelines. While there are some differences between these, it has become clear that there are some common areas that are expected to be addressed in any good Data Management Plan (DMP), regardless of funding source (Figure 3).

To better reflect those core DMP requirements, the RDMSG re-wrote our Data Management Planning guidance page, presenting six fundamental areas to include in a DMP, rather than specifically focusing on NSF requirements. In addition, we summarize and provide links DMP guidance provided by the top five research funders to Cornell: Department of Energy, Department of Defense, the National Institutes of Health, the National Science Foundation and the US Department of Agriculture, and link out to all known other funders who require Data Management Plans as part of the proposal process.

Figure 3: Re-designed Data Management Planning information page, with one section expanded to display detailed considerations and additional information on that specific topic.

Support for researchers in the life sciences. At the request of the Senior Associate Vice Provost for Research, we launched a new page on the website in January of 2017, with a focus on resources for researchers in the Life Sciences (Figure 4). There, we collected disciplinary data management information to facilitate exploration of solutions that have worked for colleagues, identification of learning opportunities, and finding new ways to build community around data management topics. There are currently four case studies described on the sight, exemplifying different needs and solutions researchers in the Life Sciences frequently encounter (example in Figure 5).
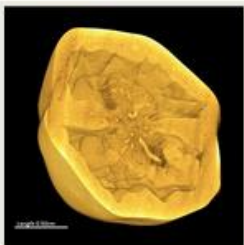
Figure 4: New RDMSG website content with focus on needs of Life Science researchers.

Figure 5: Example screenshot of a Life Science research case study. All case studies follow the format of a description of the issue or need, what specific challenges it poses for the researchers, and the solution implemented to address the need.

Website Statistics

Beginning in November of 2011, RDMSG website usage statistics were collected using Google Analytics. Due to multiple factors, including unexplained data loss from Google Analytics and a switch by CUL to a new analytics platform (Piwik), we are unable to report usage complete usage statistics that directly compare to our earlier data. Data gathered using the new analytics software (Figure 6) are not comparable to historical website stats, but clearly indicate that the website continues to receive steady use, with an expected drop around the holidays at the end of the calendar year.
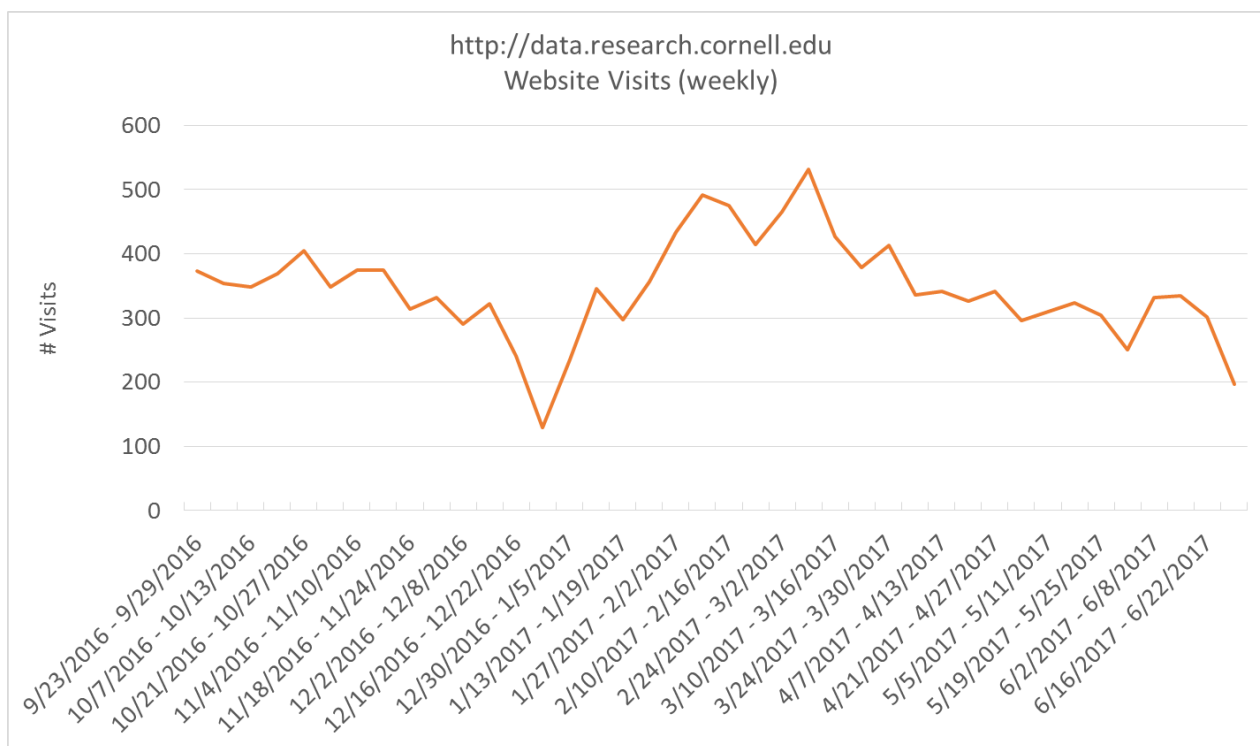
Figure 6: RDMSG website visits by week, starting beginning 9/23/2016, when new measurement algorithm implemented in Piwik. Note that these data are not comparable to past years due to change in collection software.

Plans for 2017—2018

- Data management encompasses a broad spectrum of activities, from proposal preparation through data collection, analysis, sharing and archiving. Support services for those activities at Cornell is not provided by a single entity, but rather is distributed across multiple geographically and organizationally separate groups including the library, information technology departments, and discipline-specific centers such as CISER and the Biotechnology Resource Center. While The RDMSG provides consultation upon request, researchers frequently want to do the information gathering themselves. Unfortunately, finding, comparing details, and making decisions about what services are appropriate for a given situation can be frustratingly time consuming and difficult.

    To begin to meet this need for comprehensive, easy to use, self-help resources on our campus, we are in the process of creating an interactive, online tool to help researchers find and choose between data storage options at Cornell. We hope to launch the data storage finder tool in early 2018. The tool will be built in a way that it can be re-purposed for other data management services; once deployed for data storage, the RDMSG will consider setting up similar finder tools for other services, such as data repositories or metadata standards.

- The Website team will continue to re-visit the site layout and design for 2017-2018, making changes to improve usability as needed. In addition, we will continue to review, update and add additional content throughout the year.

10

RDMSG Consultation Activities

Consultant Team Members: Florio Arguillas (CISER, May 2012 – present), Adam Brazier (Astro/CAC, Dec 2010 – present), Tarek Chams (CALS-IT, Nov 2016 – present), Dianne Dietrich (CUL, Dec 2010 – present), Amy Dygert (CUL, Feb 2017 – present), Zachary Jacques (OSP-RAIS, June 2017 – present), Keith Jenkins (CUL, Oct 2011 – present), Erica Johns (CUL, Sept 2014 - present), Wendy Kozlowski (CUL, coordinator; Dec 2011 - present), Wyman Miles (ITSO, Oct 2011 – present), Michelle Paolillo (CUL, Mar 2013 – present), James Vanee (BRC, Mar 2016 – present), Drew Wright (WCM, Dec 2011 – present), Sarah Wright (CUL, Apr 2011 – present).

Previous Team Members include: Aaron Birkland (CAC, Dec 2011 – Oct 2015), Daina Bouquin (WCM, Jul 2013 – Sep 2015), Eric Chen (CAC, Dec 2010 – May 2011), Jeremy Cusker (CUL, Mar 2011 – Feb 2012), Michelle Edwards (CISER, Jan 2015 – Dec 2016), Chris Fournier (CUL, Mar 2015 – Oct 2016), Peter Hirtle (CUL, Mar 2011 – Oct 2015), Stefan Kramer (CISER, Dec 2010 – Jan 2012), Fiona Patrick (CUL, Feb 2011 – Sep 2012), Jaron Porciello (CUL, Mar 2013 – Aug 2015) Gail Steinhart (CUL, Dec 2010 – Mar 2015), Linda Woodard (CAC, Dec 2010 – Dec 2011)

Operating Principles:

RDMSG consultants strive to meet the Cornell research community's data management needs by:

- Providing timely and professional assistance in the creation and implementation of data management plans.
- Encouraging best practices in data management, including those that promote sharing, reuse, and preservation of research data, while respecting the concerns and practical constraints researchers face.
- Bringing diverse backgrounds and expertise to the RDMSG and recognizing when additional expertise may be required to meet a researcher's needs.
- Sharing information with other RDMSG consultants to provide the best possible service, while collectively treating all information in grant proposals as confidential.
- Referring researchers to the most appropriate services, whether at Cornell or elsewhere.

Summary of Activities, July 2016 – June 2017

- Consultants met monthly to share experiences and information regarding consultation issues, as well as developments at Cornell and elsewhere related to data management.
- We received 114 inquiries to RDMSG between July 2016 and June 2017 (Figure 7). Many inquiries continue to reach us via the RDMSG RT ticketing system (rdmsg-help@cornell.edu); in addition, inquiries directly addressed to specific consultants were added to RT for record keeping. In total, these inquiries led to 89 individual consultations with Cornell faculty, staff and students. The remaining 25 inquiries (22%) were administrative in nature.
  - This year, 36 (40%) of our consultations were about preparation of data management plans related to a proposal, down only slightly from 42.5% last year (Figure 8). Of those, the majority continue to be about National Science Foundation Data Management Plans (23 of 36 this year (63.8%), 20 of 23 last year (87%)). Other funding agency DMPs consulted on this year included NASA, USDA, USGS and NIH.

- While the absolute number has gone up (53 of 89 this year, 31 of 67 last year), percentage of consultations that are about general data management issues continues to be a larger proportion of our total consultations (59.6% this year, 57.4% last year) than DMP-related inquiries (Figure 8).
- Average time spent on consultations remained close to an hour (~58 minutes), though consultation times ranged from just 5 minutes to as long as 4-5 hours in a couple of cases.
- 36 of the 89 consultations (40%) were in person and the rest usually done over email and very rarely by phone alone.
- The vast majority of the consultations (where recorded) were from Cornell's College of Agriculture and Life Science and College of Engineering (Figure 9).
- Non-DMP-related consultations included conversations with institutions outside of Cornell regarding organization and logistics for data management support and outreach, metadata creation inquiries, database design and support questions, follow-ups on execution of DMPs, requests for data analysis support in Excel and python.
- In addition to the above numbers:
    - There were at least 11 electronic lab notebook (ELN) related consultations, trainings and outreach sessions where the focus was use ELNs and the LabArchives product and/or the Open Science Framework product. Wendy Kozlowski continues to represent RDMSG to support research lab use of ELNs.
    - Wendy Kozlowski and Erica Johns curated a total of 8 datasets for deposit into eCommons. Curation services for those datasets included format and content review, dataset documentation and discovery metadata creation assistance, and either submission on the student or faculty's behalf, or training on the eCommons system for user self-submission.
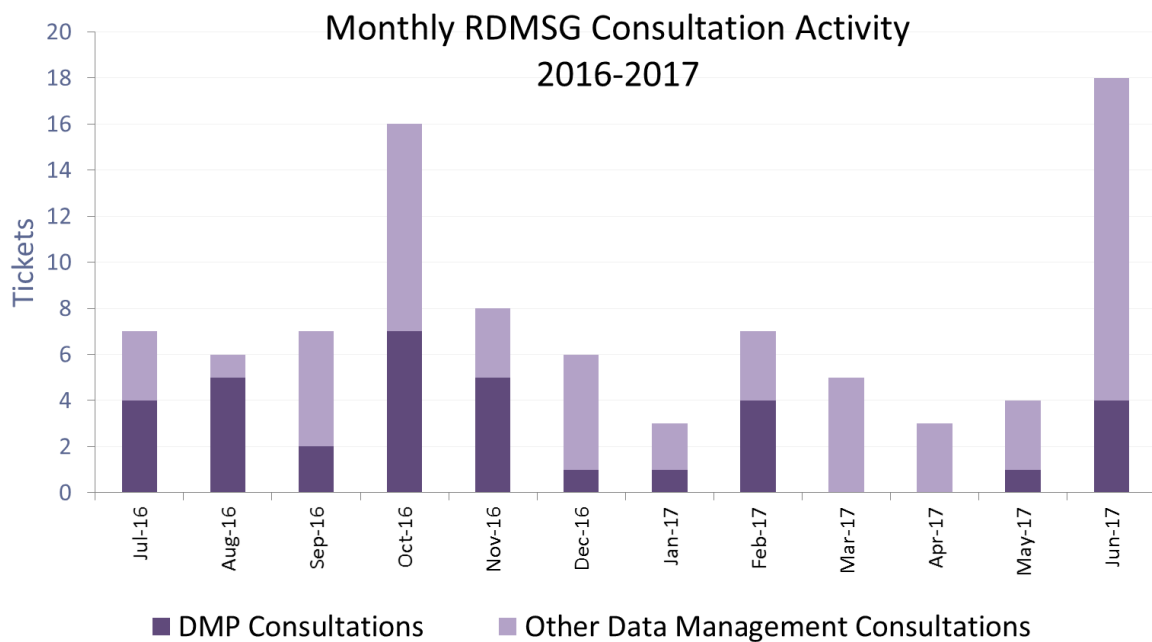
Figure 7: RDMSG Consultation activity, 2016 – 2017, showing clear peaks of activity in October and June (consistent with previous peak activity months).
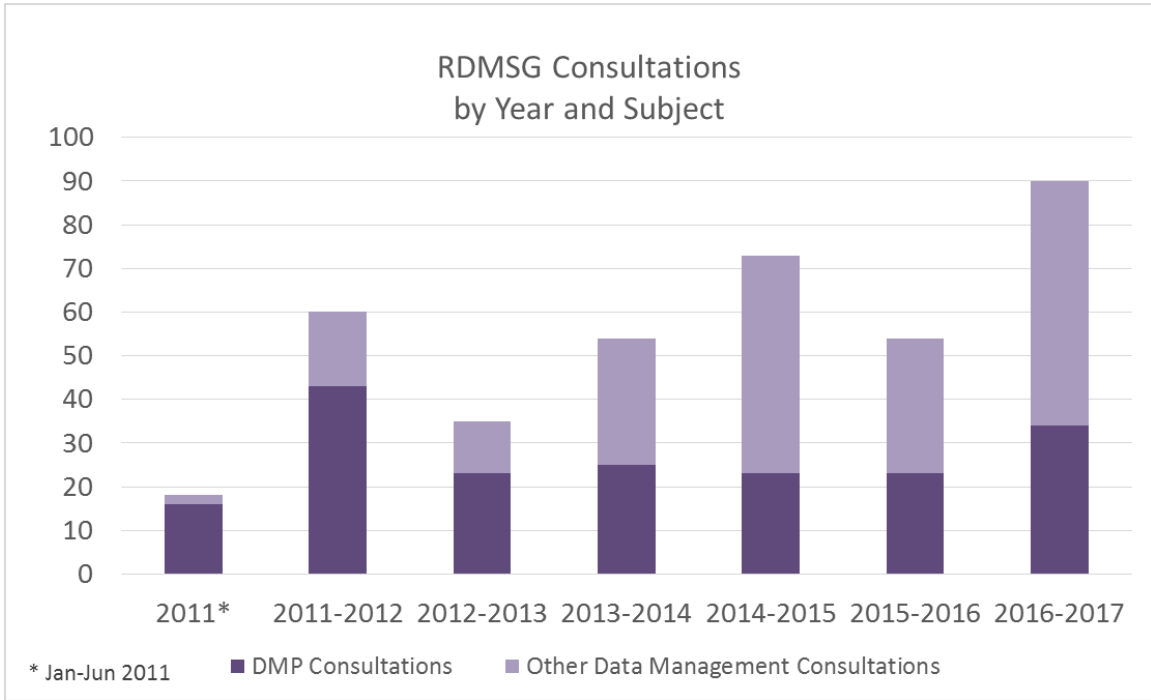
Figure 8: RDMSG Consultations by year and subject (DMP-related vs other data management topics).
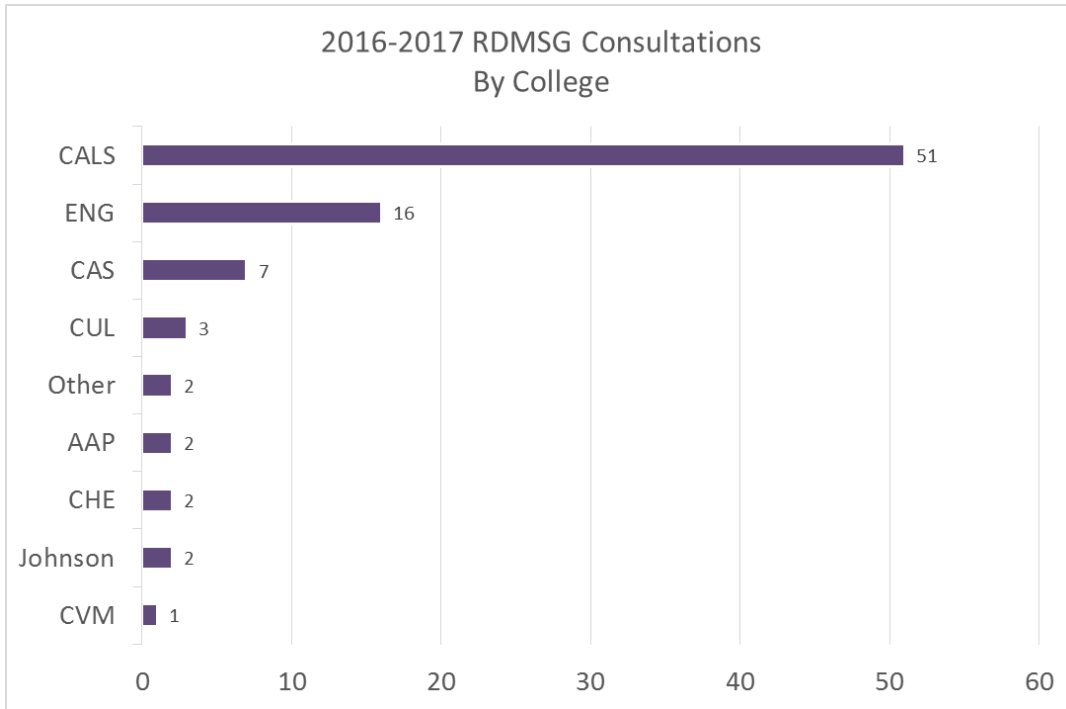
Figure 9: RDMSG Consultations by college. Data reflects only consultations where college was noted, not the full set of 2016-2017 consultations. "Other" category includes consultations with individuals outside of Cornell University.

- Continue consultation efforts in support of agency data management plan preparation and other data-related needs.
- Disseminate information about data sharing requirements by funding agencies and publishers.
- Meet regularly as a group to share consultation experiences, and explore and educate ourselves about new tools and services at Cornell, Weill Medical or elsewhere.

Summary and Conclusions

As far back as 1995, agencies such as the National Science Foundation (NSF) have had statements in their grant and award proposal preparation guides (http://www.nsf.gov/bfa/dias/policy/dmp.jsp) stating the expectation of sharing research results. Since then, several other major funding agencies, including the NIH (2003), the NSF (2011) and many others require data management plans with proposals, and stipulate that sharing of research data (with exceptions for confidentiality, privacy, IP and commercial concerns) is a requirement of receiving funding. In addition, many journals (Science, SpringerNature, AGU Publications, PlosOne and others) now also require sharing of data as a stipulation of article publishing.

While the RDMSG continues to consult on preparation of data management plans, there is also a clear need to support active data management, and data sharing. These activities are often not "one-time" interactions, but rather an iterative process by which the researchers learns what they need and how to best make that happen. The RDMSG strives to offer multiple ways to facilitate that learning, including comprehensive, up-to-date, self-service information on the website, and accessible and relevant consultative expertise. The RDMSG offers researchers support, but those interactions in turn allow us to gather information about needs and practices of Cornell's researchers, ensuring that campus service providers know what is needed to meet researcher requirements across our campuses.
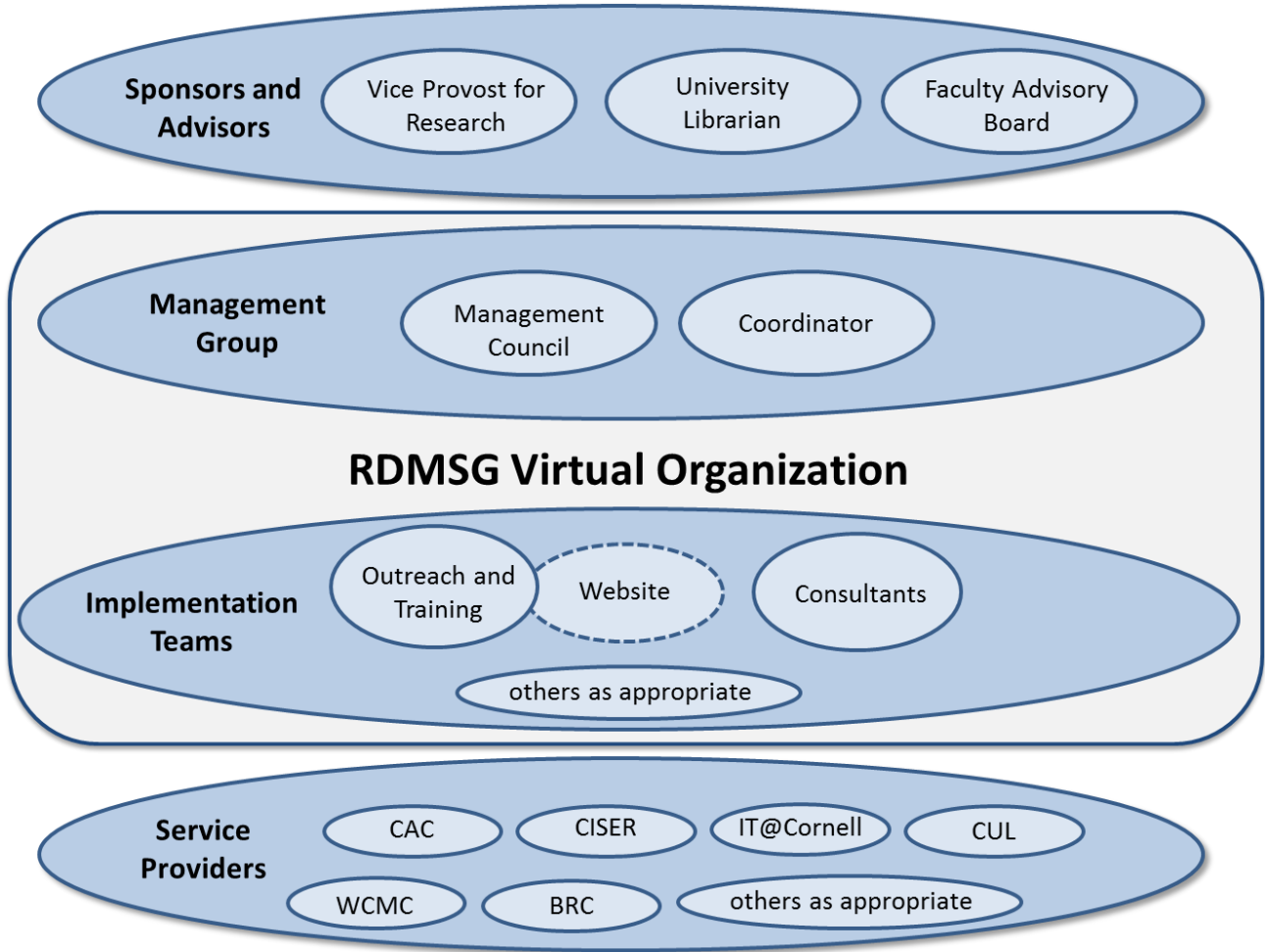
Finally, continued interest and attendance at workshops and training sessions by staff and students, indicate that they too acknowledge the importance of good and practical data management skills. Such practical data management skills are not necessarily inherent in traditional research pedagogy, and the RDMSG continues to provide supplementary and alternative sources of information by which they can acquire those skills.

During 2017-2018, the RDMSG will focus efforts on:
- Information dissemination about funder and publisher requirements for data management planning, sharing and archiving.
- Outreach and communication about RDMSG services, especially to those who don't yet know about our services.
- Adding and updating relevant content to the RDMSG website.
- Listening to the needs of faculty, staff and students, and communicating those needs to service groups on campus who may be able to assist / provide appropriate services.

Appendix A:
RDMSG Organizational Structure

Sponsors and Advisors
- Vice Provost for Research
- University Librarian
- Faculty Advisory Board

RDMSG Virtual Organization

Management Group
- Management Council
- Coordinator

Implementation Teams
- Outreach and Training
- Website
- Consultants
- others as appropriate

Service Providers
- CAC
- CISER
- IT@Cornell
- CUL
- WCMC
- BRC
- others as appropriate

Appendix B:
RDMSG Faculty Advisory Board Members

John Abowd (ILRLE)
Myles Gideon (ORIA)
William Arms (CS)
Richard Burkhauser (PAM)
Claire Cardie (CS)
James Cordes (ASTRO)
Art DeGaetano (EAS)
Mary-Margaret Klempa (OSP)
Lawrence Gibbons (PHYS)
Natalie Mahowald (EAS)
Michael Webster (Lab of O)
Wendy Kozlowski (ex officio, RDMSG Coordinator)

Appendix C:
RDMSG Management Council Members

William Block (CISER)
James Cordes (ASTRO)
Dean Krafft (CUL-IT)
David Lifka (CU-IT CIO, CAC)
Oya Rieger (CUL)
Vinay Varughese (WCMC-IT)
David Vernon (CIT)
Wendy Kozlowski (ex officio, RDMSG Coordinator)