

Chapter 10

Diving into Data: Developing Data Fluency for Librarians

Scott Martin and Jo Angela Oehrli

University of Michigan Library

As data services continue to emerge as a growth area for academic research libraries, many librarians are finding themselves in need of a skill upgrade in order to support research data at a level of excellence consistent with other library services. With the full support of its administrators, the University of Michigan (UM) Library has implemented a three-stage professional development program to develop readiness for data services within the organization. The first stage consists of a two-part workshop series presenting basic data concepts around data structures, storage, security, and sharing. These workshops build a shared understanding and vocabulary among librarians across the institution as a foundation for more specialized training. The second stage employs a workflow developed by UM librarians, the Deep Dive into Data, which provides a self-directed, iterative, non-linear method for exploring the data landscape around a particular research discipline. Deep Dive workshops apply this workflow to an example discipline through in-class exercises and discussion, providing librarians with an expert-mediated means to build their familiarity with the tool before applying it to their own work. The ongoing third stage addresses specific topics in research data which are broadly applicable across disciplines. Some of these workshops explore cross-disciplinary data techniques which librarians may encounter or recommend to researchers, such as text mining. Others connect librarians to services provided elsewhere on campus, such as data storage, by inviting colleagues from those providers to brief library staff on the basics of the service and how to make an effective referral.

Another type expands traditional public services skills, such as reference interviews, into the data services realm by providing an overview of current best practices and reports from individual librarians who have been piloting these services. Feedback gathered from participants at each stage has informed planning for subsequent offerings.

Getting Started

Like many other large academic libraries, the University of Michigan Library has been considering how to respond to the needs of researchers around managing, disseminating and curating their research data. The library had launched several initiatives in research data support, including participation in the DataCite service for Digital Object Identifier (DOI) assignment, support for ORCID identifiers for disambiguation of researcher names, and exploratory programs in Data Management Plan consultation and data archiving, all under the auspices of its new Research Data Services group. As these efforts unfolded they highlighted clear needs, both among library staff and the wider campus community, for increased education about data concepts. In response, a meeting was called in the summer of 2013 to bring together librarians from the Research unit, who were primarily tasked with liaison duties to specific schools and departments, with librarians from the Learning and Teaching unit, as well as from data-relevant programs of the Collections and Publishing units, to form a Data Education Working Group (DEWG) whose charge was to create library programming specifically to address the educational needs we had observed.

As a part of this initial meeting, the Director for Research Data Services outlined a vision for lifecycle-oriented data education. This education would begin with in-reach to librarians, to prepare them to engage productively with researchers as data services providers, and eventually

expand in scope to include outreach to our campus service population. Discussion during this and subsequent meetings refined the in-reach portion of the effort to a two-part approach. First, there would be training aimed at establishing a baseline level of general data literacy among librarians. This would set the stage for a second phase, in which librarians would receive more specific training in resources, techniques, and approaches rooted in the context of specific disciplines.

General Data Literacy

The UM Library's approach to general data literacy started with an examination of existing options, focusing largely on MANTRA research data management training <<http://datalib.edina.ac.uk/mantra/>> and its associated do-it-yourself training kit for librarians. While there was much value in these materials, the decision was made not to simply suggest them to our librarians as "background reading" prior to in-house workshops, for a number of reasons. First, the existing approaches lacked practical information: the materials discussing storage, for example, provided copious testimonials from researchers about the value of good backup practices, but less information about how to create an effective data storage plan. Second, the available materials were more heavily biased toward STEM examples and applications than we preferred, given our mandate to develop materials for the entire UM Library system. Finally, the existing programs, as structured, demanded more time commitment than could reasonably be expected from our librarians.

Rather than implement MANTRA, the UM librarians elected to create an in-house "Research Data Concepts for Librarians" training which would distill some of the more basic and practical information into a more accessible form. The objective was to create a common

institutional vocabulary of data concepts as a basis on which to build contextual data education later, and to provide librarians with some practical information on best practices for their own data management needs. A small subgroup, composed of volunteer staff with some experience in working with research data, divided the training into two 2-hour workshops. Each workshop was taught by two librarians and covered two thematic modules, using lecture, discussion, and in-class exercises.

The first workshop, “Working with Data,” focused on topics related to the creation, documentation, and storage of data. The session began with a module discussing basic concepts in data management and design (best practices in file naming conventions, open vs. proprietary file formats, relational database concepts, metadata), illustrating these concepts with a data design exercise for a hypothetical circulation-data project. The second module continued with an overview of best practices in data storage and security, such as physical vs. online options for storage, planning backup procedures, proper data disposal, and basic information about encryption, and spurred discussion with an interactive exploration of some of the available data-storage options for University of Michigan researchers.

The second workshop, “Sharing and Preserving Data,” discussed data publishing, data citation, and data management planning. Its first module explained the differences between data repositories and data journals (using a “repository safari” exercise to explore the variety of repositories currently available in various disciplines) and discussed data citation practices and catalogs of published data sets, while the second module gave a brief overview of the Office of Science and Technology Policy (OSTP) memo and its potential impact on federally-funded research, outlined the basic elements of a data management plan, and previewed a UM Library project to assess the existing DMP practices at the College of Engineering.

The DEWG debuted this introductory training during the fall semester of 2013, offering two sessions of the first workshop on consecutive days in October, and two sessions of the second workshop in November. One session was offered in the morning and one in the afternoon, to provide maximum scheduling flexibility for staff who wished to attend. An additional session of each workshop was held in January 2014. Over 60 librarians attended at least one session of this training, with approximately half attending a session of both of the workshops.

The DEWG solicited feedback from attendees using an online survey, and received 23 responses. About 80% of respondents indicated that the level of the material presented was “just right”; none reported that it was “too advanced.” Twenty percent of respondents reported that the workshop material was relevant to questions they’d received from researchers; 71% reported that the material was relevant to questions that had arisen in their own work with data. Attendees were also polled about their perceived ability to provide data management services. When asked whether they were comfortable talking with researchers about data management issues, only 15% responded “yes,” while another 71% responded “somewhat,” and 14% responded “no.” A related question about preparedness to provide research data management support found no respondents answering “very prepared,” 67% indicated that they felt “somewhat prepared,” 14% that they felt “not prepared,” and 19% reported that the question was not applicable, presumably because their duties did not involve liaising with research communities. These responses indicated that while the initial offering was useful, more education would be needed before library staff felt ready to support a new set of research data services.

Following the basic instruction in data concepts, a larger subgroup of the Data Education Working Group was charged to plan contextual educational offerings focused on exploring data in particular disciplinary contexts. One obvious strategy would be to offer workshops which introduced participants to the data landscape of a particular discipline: ecology, psychology, etc. On reflection, however, there were some drawbacks to this approach. First, the sheer scope of the disciplines served by the University of Michigan Library prevented any sort of comprehensive coverage. Moreover, the librarians most likely to benefit from any particular discipline-based session were also likely to be the librarians the DEWG would want to recruit to teach such a session, and the divergence of data cultures between disciplines meant that most librarians were likely to derive substantially less practical benefit from attending sessions outside their disciplines.

Our solution to this dilemma was to develop and teach a methodology which individual subject librarians could use in a self-directed exploration of their own disciplines. This method, which we named the Deep Dive into Data, consisted of several categories of exploration, corresponding to common features of a particular data landscape: data requirements of stakeholders (such as journals and funding agencies); data repositories; metadata standards; subject-specific data literature; and disciplinary data culture. Librarians could begin their exploration in any of these categories, using knowledge about their discipline's data culture that they had already gleaned from other sources (conversations with researchers, conversations with professional colleagues, readings in research data topics, etc.), and proceed in a nonlinear fashion through the categories as they uncovered additional information. Exploration of a data repository, for example, might reveal a frequently-used metadata standard, which was originally published in a journal devoted to data issues in the discipline, whose most recent issue might

contain an article examining researcher attitudes toward data sharing. See Appendix A for the detailed Deep Dive workflow document, written in collaboration with a Council on Library and Information Resources (CLIR) Data Curation Fellow, who at that time was embedded with the UM Library.

In order to teach the Deep Dive methodology, we returned to the idea of subject-focused workshops but with a modified approach in teaching them. Rather than simply deliver factual information about the data landscape of a particular discipline, these workshops investigated sample disciplines interactively using the Deep Dive methodology as a means to illustrate how it could be applied. Attendees performed an in-class exercise for each category, under the direction of the subject liaison librarian for that discipline, and shared their results in group discussion. Teaching the methodology in this way provided two key benefits. First, since participants could not be expected to come into the workshop with prior knowledge of the discipline on which to build their investigations, the DEWG developed exercises which could uncover features of a data landscape from scratch. To investigate stakeholder requirements, for example, participants had to identify top-ranking journals in the discipline by using the Journal Citation Reports rankings and explore their data deposit requirements (if any). Disciplinary data repositories, likewise, were identified using online repository directories such as OpenDOAR, Databib, and re3data, which provided browsable subject categories. These approaches then enriched the methodology, as suggested tactics to augment librarians' existing knowledge of data in their disciplines. Second, leading librarians successfully through an exploration of a foreign discipline built their confidence with the methodology, which would facilitate its application to investigation of their "home" discipline(s).

The Data Education Working Group offered a number of Deep Dive into Data workshops during the winter semester of 2014. Originally, we planned for two workshops to serve as paradigms for disciplinary groups: an Ecology workshop to model a STEM Deep Dive, and a Psychology workshop to model a social sciences Deep Dive. The Psychology session drew 31 attendees and included subject liaisons from arts, humanities, and STEM disciplines as well as non-liaison library staff. The Ecology session, by comparison, drew 15 attendees, largely from the Science/Engineering/Data Services cluster within the library's Research unit. The DEWG subsequently offered a Deep Dive into Clinical Data workshop, as a means to explore some of the sensitive-data and intellectual property issues more commonly found in medical research environments, as well as a workshop presenting the Deep Dive workflow in the abstract, without a paradigmatic subject to serve as an example (and necessitating, therefore, a lecture-based presentation rather than one integrating exercises and discussion). These drew 8 and 25 attendees, respectively, with the attendees at the Clinical Data workshop coming exclusively from the Health Sciences and Science/Engineering/Data Services groups. These attendance numbers suggest that while interest in data issues among librarians is broad, the contextual framing of presentation can significantly affect attendance: sessions advertised as STEM-related drew substantially fewer attendees, especially from non-STEM disciplines, than materials based on the same content but advertised using non-STEM-specific language.

Advanced Data Workshops

The third stage of the UM Library's educational effort emerged during planning sessions for the Deep Dive workshops, as the need for more in-depth education on specific data-related topics that transcended the boundaries of any individual discipline became evident. The DEWG

conceptualized this third stage as Advanced Data Workshops: an open-ended series designed to focus on one tool, strategy, task, or type of data and explore it in detail through presentations, active learning strategies, and guest lectures. The DEWG generated topics for these sessions via the feedback surveys from the previous workshops, conversations with colleagues (within the workshops and in other informal settings), and proposals from DEWG members. Once the list of topics had been assembled, the DEWG met to describe each possible workshop more fully, to brainstorm with possible UM librarians or campus instructors who could teach each workshop, and to assign a coordinator from the group who would be responsible for the logistics of the workshop (scheduling rooms, publicity, working with instructors to develop workshop content, etc.). In some cases, librarians from the DEWG both coordinated and taught the workshop.

The titles of the proposed Advanced Data Workshops for Fall 2014 were

- Data Storage
- Qualitative Data
- Data Reference Interview
- Strategies 101: Teaching Data Concepts
- DMPTool
- Data Interviews With Faculty
- Text Mining
- Grants Data Sources
- International Data
- Data Citation
- Legal and Ethical Concepts In Data

To facilitate so many workshops, we soon realized that it would be difficult to coordinate instructor availability and room availability with the campus academic calendar and with sufficient preparation time. The plan was to schedule all of the workshops so that only one of the workshops would be held in any given week, necessitating the use of a shared calendar to avoid clustering the workshops. While it might have been possible to create workshops that were curricular and progressive – each workshop building on knowledge gained from prior workshops

– the logistics of scheduling the workshops and coordinating library instructor availability seemed too difficult to coordinate the instruction on a more curricular level.

The Data Education Working Group was able to ask both UM librarians and another campus instructor to teach the workshops, signifying a great commitment by the library to collaborate and provide staffing for lifelong learning and professional development to grow the library’s capacity to meet data services needs across campus. The library reached out to a data storage specialist in the IT department of the College of Literature, Sciences and the Arts to present the Data Storage workshop. One librarian graciously agreed to teach more than one workshop. Members of the DEWG often coordinated more than one workshop throughout the semester and planned their assigned workshops simultaneously. Descriptions of individual workshops and librarians who served as instructors can be found in Table 10.1. Workshops used various strategies to convey each concept, tool, or strategy including presentation, group discussion, role playing, scenario building, inquiry-based learning, and think/pair/share activities.

<insert Table 10.1 here>

Table 10.1. Descriptions of Individual Workshops and Librarians Who Served as Instructors

Of the eleven workshops that the DEWG planned to offer in fall 2014, seven were ultimately offered. The four workshops that could not be completed were the Qualitative Data workshop, the Data Citation workshop, the DMPTool workshop, and the Legal and Ethical Concepts in Data workshop. Instructors were willing to teach all of these workshops, but the DMPTool workshop was delayed until the university deployed the newest version of the tool (and shelved it altogether following the University’s decision not to support it institutionally),

and scheduling difficulties ultimately prevented the others being offered during fall semester. The original plan was to offer repeat sessions of all workshops in the following Winter semester, but attendance patterns for the workshops suggested that the DEWG should repeat only the three workshops that fostered the most interest (Data Reference Interview, Strategies 101: Teaching Data Concepts, and Data Storage). This approach freed up staff effort which could then be dedicated to revisiting the planned workshops which were not offered in the Fall, as well as planning and offering new workshops; these efforts are ongoing.

Assessment

Several strategies were used to assess our overall data education efforts. Attendance records were compiled from the individual sessions into a de-duplicated master list of attendees, which gave some demographic information about the audience. Feedback was solicited following each of the individual workshops, using a standard survey form in Qualtrics delivered to attendees via e-mail. The DEWG also used the master list of attendees to distribute a broader survey about incorporation of the data education material into librarians' regular duties.

As of the end of 2014, a total of 128 library staff had attended at least one of the data education workshops (Table 10.2). This represents approximately 26.5% of the library's total staff count (as of mid-November 2014). A handful of the attendees listed, however, had left the library by the time data analysis began, so this number is a slight overrepresentation of the coverage of current staff. Unsurprisingly, a large percentage of attendees came from the library's Research unit, which includes most of the subject liaison librarians; these librarians are most likely to receive requests or perceive a need for research data services among the researchers they serve, and as such are likely to be more motivated to advance their skills in that area. We

also achieved our highest unit-level coverage of staff among the Research librarians, with 80% of the unit staff attending one or more workshops. By head count, the next most-represented area of the library was the Collections unit; the next highest-coverage unit was the Health Sciences Library. The DEWG's greatest concern from this analysis was the relatively sparse coverage of the library Information Technology unit, an area in which we would expect substantial interest in research data issues. Offerings to date may have been deemed either too basic or too public service-focused to address the specific needs of the more infrastructure-oriented staff in that unit, although it's not possible to rule out problems of communication or staff concerns about work-scheduling priorities. Meetings will be arranged with managers in the unit to explore these issues.

<insert Table 10.2 here>

Table 10.2. Workshop Attendance by Library Unit

DEWG members asked participants in the Deep Dive and Advanced Data Training workshops for their feedback on the workshop's content, ratings of their comfort with research data, and suggestions for additional workshops (see Appendix B for a sample survey). In general, participants rated the content level of the workshops "just right" compared to their existing knowledge, with scattered responses of "too basic" and no responses of "too advanced." When asked whether the information was relevant to data questions received from researchers, respondents answered "yes" four times more often than "no"; when asked whether the information was relevant to data questions arising from their own work with data, "yes" responses were six times more frequent than "no" responses.

When attendees were asked about their comfort with providing data services, however, the results were less encouraging. Over half (54%) of the respondents answered “somewhat” when asked if they were comfortable talking about data issues with researchers, with 17.4% responding “no.” When asked how prepared they felt to provide research data management support, only 3% responded “very”; over a quarter (27.6%) responded “not prepared,” with the remainder answering “somewhat.” Moreover, the responses to these questions did not show any temporal trend to indicate that comfort with data issues increased as the educational program proceeded.

As registration for these workshops is voluntary, the continuing indications of discomfort with providing research data services may reflect a self-selecting population of librarians who attend because they are uncomfortable with data concepts, and who may therefore stop attending once they start to feel comfortable. Contrary to this hypothesis, the attendance data clearly shows that individual librarians have a fairly strong tendency to attend multiple sessions (128 attending staffers represent 371 individual attendances), and therefore presumably have submitted multiple feedback responses. A more likely scenario is that while they find the workshops to date valuable, attendees still perceive a gap between their current skills and the skills they need to confidently provide data services to their researchers. Reviewing the free-text answers to the question “Which topics should be addressed by future workshops?” suggests that data management plans remain a prime area of need for further education. While the plan was for DMP education from the beginning of the process, institutional uncertainty about support for DMPTool delayed and ultimately scuttled this offering; the library has, however, explored DMP consulting in the meantime via individual pilot projects, and those experiences will be translated into educational offerings in the months to come.

In mid-November 2014, an additional survey was sent to all of the current Library employees who had attended at least one data education session. In total, of 114 e-mails sent, there were 35 responses received, for a 30.7% response rate. The largest set of respondents was from the library's Research unit, with Collections, Learning and Teaching, and Health Sciences making up the bulk of the remainder. When asked about the workshops they had attended, about two-thirds of the respondents reported that they had attended one or more of the Deep Dive into Data workshops, but only six respondents indicated that they had used the method to explore their own disciplines, and three of those six had served as instructors for Deep Dive workshops. (As the Deep Dive materials are freely available to library staff, an affirmative response to the attendance question was not required to view the usage question.) This suggests that while they may find the method valuable, few librarians have felt enough urgency around exploration of subject data landscapes to prioritize this explanation above other projects and duties. Another contributing factor was that only the subject liaison librarians have formally-assigned research communities to support, but attendance at the Deep Dive workshops was not limited to that population; it would be natural to expect non-liaison librarians to feel less urgency about conducting a Deep Dive in the absence of a specific project which requires greater familiarity with a specific subject's data landscape. Respondents' answers to the question "How have you applied what you learned in these workshops to your work?" support these conclusions: of the 18 respondents who gave an answer to this question, half reported that they had applied the content in some fashion, and many of the remainder indicated that they expected to use the information in the future or had no service communities to which the content could be applied.

Respondents were asked whether they had conversed with researchers about data needs; depending on their response, they were asked either what needs had been expressed or observed,

or what the library could do to help facilitate those conversations. There were 29 responses to this portion of the survey. A third of the respondents answered “yes” to the first inquiry; the observed researcher needs that they reported included storage and/or preservation of data sets, metadata assistance, data management training, data sharing, and assistance in handling sensitive data. The respondents who answered “no” to the first question suggested conversation-facilitating activities ranging from “a checklist of things to cover with researchers when talking about data” to “conversations about how to best approach and market to faculty” to “I’ll let you know.”

Lessons Learned

At this stage in data education efforts, we have amassed enough data and experience to make some generalizations about the process:

- Based on both the survey feedback and informal conversations with colleagues, this staged approach works well as a library-wide strategy for data education: it provides an entry point for staff with less familiarity with data concepts, while quickly moving into areas that can effectively engage more experienced staff. We cannot, however, rule out the possibility that the attendees effectively self-select for workshops that match their perceived skill level and interests; the relatively low levels of attendance by library Information Technology staff provides some support for this hypothesis.
- Library interest in data education is both broad and deep: attendees were drawn from every unit of the UM Library, not just those tasked directly to research support, and represented approximately a quarter of the current staff.

- Proper time allocation was crucial to the success of the program. Planning and implementing workshops that use a variety of teaching methods can be time-intensive. In addition, it can be helpful to have multiple instructors which may add other difficult time components in coordinating schedules. Finally, providing professional development workshops during the fall semester when many other library activities may take place, including more information literacy sessions, can strain librarian and space scheduling.
- The DEWG recruited a variety of library colleagues and campus partners to present and/or teach during these sessions. This helped distribute the effort outside of the DEWG members, while highlighting the existence of a broad-based community of practice within the library.
- Use of an assortment of instructional methods in the workshops, ranging from presentation style to role playing, helped to sustain librarian interest. Workshop length was tailored to suit both content and methods: sessions using active learning strategies usually were 80-90 minutes in length, while presentation-style sessions usually ran 60 minutes (including time for questions).
- While the Deep Dive method is perceived as valuable, that value alone is not a sufficient motivator for librarians to independently put it to use; other duties are apparently perceived as having a higher priority at present. DEWG members are exploring ideas for a library-wide data event with a report-out component, which might spur additional use of the method.
- Additional education, even when positively received, does not always translate into increased readiness for service, particularly when a service need (in this case, data

management plan consulting) which staff perceive to be significant has not yet been addressed by the educational program.

Conclusion

Our educational efforts in this area are far from complete. The work to date has established a firm foundation for research data support; the next challenge is to build on it. DEWG members are currently planning several activities in this area. First, the Advanced Data Training series continues, with new workshops (Metadata for Research Data, Introduction to Visualization) and previous successes scheduled for winter semester of 2015. A formal announcement of library-wide research data services is expected later in 2015, and an important role for the Data Education Working Group in offering more targeted skill-building education in support of these services. A sustainable plan is being developed for making this education available both to new library staff and to existing staff whose changing duties may create new needs for education. Finally, DEWG members will need to look beyond the library to the on-campus research communities we support, and develop educational offerings targeted to their needs.

Acknowledgements

Thanks to RDS Director Jen Green, for kicking off this whole effort; CLIR Data Curation Fellows Katherine Akers, Natsuko Nicholls, and Fe Sferdean, for their dedicated contributions; Jen Brown, for assistance in analyzing the assessment data; and our DEWG colleagues: Jake Carlson, Marisa Conte, Leena Lalwani, and Alexa Pearce.