9-18-2014

Developing Professional Skills in STEM Students: Data Information Literacy

Lisa Zilinski
Purdue University, ldz@andrew.cmu.edu

Megan R. Sapp Nelson
msn@purdue.edu

Amy S. Van Epps
Purdue University, vanepa@purdue.edu

Issues in Science and Technology Librarianship Summer 2014
DOI:10.5062/F42V2D2Z

Refereed article

Developing Professional Skills in STEM Students: Data Information Literacy

Lisa D. Zilinski
Assistant Professor of Library Science
Data Specialist
lzilins@purdue.edu

Megan Sapp Nelson
Associate Professor of Library Science
Engineering Librarian
mrsapp@purdue.edu

Amy S. Van Epps
Associate Professor of Library Science
Engineering Librarian, Coordinator of Instruction
vanepa@purdue.edu

Purdue University Libraries
West Lafayette, Indiana

Abstract

Undergraduate STEM students are increasingly expected to have some data use skills upon graduation, whether they pursue post-graduate education or move into industry. This project was an initial foray into the application of data information literacy competencies to training undergraduate students to identify markers of data and information quality. The data consumer training appeared within two courses to help students evaluate data objects, including databases and datasets available on the Internet. The application of the Data Credibility Checklist provides a foundation for developing data reuse competencies. Based upon the initial presentation of the content, it became obvious that students need very basic introductions to data concepts, including definitions for database and dataset, and the process of data object discovery.

Introduction

In the quest to define data information literacy, graduate students have been the primary subject group. These skills, particularly that of data consumption (the location of, access to, and assimilation of data), provide a valuable skill set for early career STEM professionals and graduate researchers that are also applicable in the coursework of undergraduate students.

This paper details an approach to identify data consumption competencies for undergraduate education. The authors describe the introduction of those competencies in STEM undergraduate classrooms as part of established disciplinary curricula.

Data Sharing

The science and engineering fields rely heavily on access to data. Researchers continue to produce record amounts of data. Other investigators and students "can mine the data to ask their own questions or to identify bases for comparison with data from other sources" (Borgman 2012, p. 1063). Combining and aggregating data from different sources can allow others to ask new questions, conduct a meta-analysis, facilitate replication of research, and improve collaboration among scientists (Borgman 2012; Wallis, Rolando, & Borgman 2013; National Research Council 2009).

Data sharing and reuse is also important in the apprenticeship process. Data reuse by novice scholars has two positive impacts: it facilitates participation in the research process and it enculturates the scholars in a community of practice (Kreisberg, Frank, Faniel, & Yakel 2013).

Professional Research Data Management Skills

Graduates of engineering and technology programs will be expected to use available information to make decisions in order to be successful in the workforce. According to the Center for Energy Workforce Development (2011), technical competencies relating to data consumption for an entry level engineer includes:

- Identify problems through data collection and analysis
- Apply logical processes to analyze information and draw conclusions
- Identify inconsistent or missing information
- Critically review, analyze, synthesize, compare and interpret information

These competencies are complementary to the National Society of Professional Engineers Code of Ethics for Engineers statement: "Engineers may express publicly technical opinions that are founded upon knowledge of the facts and competence in the subject matter" (2007, p. 1).

Despite professional expectations, research data protocols, such as the Guidelines for Responsible Data Management in Scientific Research, do not address data consumption competencies and skills (Coulehan & Wells 2006). Similarly, professional societies such as the American Chemical Society (ACS) do not address data consumption in their information skills and competencies documents (2012).

Data Reuse in the Literature

Given the necessity of using computational scientific tools in the broad spectrum of research and the central necessity for the reproducibility of science, the ability to reuse data in an ethical, educated manner is an emerging skill particularly crucial for post-graduate education. "To prepare the next generation of scholars, the knowledge and skills for managing data should become part of an education process that includes opportunities for students to contribute to the creation and preservation of research in their fields" (Ogburn 2010, p. 244). Additionally, Calzada Prado and Marzal noted that "...we feel that data literacy, like information literacy, should be acquired gradually at all levels of schooling and even throughout individuals' lifetimes" (2013, p. 124).

The literature contains multiple proposals for the educational competencies necessary for successful data information literacy (Calzada Prado & Marzal 2013; Carlson, et al. 2011; Qin & D'Ignazio 2010). The reuse of data is a fundamental skill that is necessary for the practice of STEM research and highlighted in each of these proposals. Additionally, professional and government organizations have called for education in data consumption and interpretation (Calzada Prado & Marzal 2013). The students recognize this need for education. Carlson et al. (2013) found that graduate students rate the skills supporting discovery and acquisition of data as more important than their supervisory faculty. This need has been recognized in the social science disciplines for a decade, and is now an integral part of sociology education for undergraduates (Stephenson & Caravello 2007; Wagenaar 2004). In the biological sciences, data discovery and acquisition have recently become core skills for undergraduates as well (MacMillan 2010).

Developing Data Reuse Competencies

The key to successful integration of data acquisition and reuse skills at the undergraduate level is building the competencies into existing courses and assignments (Carlson et al. 2011; Qin & D'Ignazio 2010). In this project, we sought to integrate these competencies into two existing upper-level undergraduate courses in Sciences and Nuclear Engineering.

Using the data competencies proposed in Carlson et al. (2011) and the tools developed during the Data Information Literacy grant, we mined the list of competencies related to discovery and acquisition of data, data conversion and interoperability, and data quality and documentation (http://wiki.lib.purdue.edu/display/ste/materials). We brainstormed additional related competencies that were particularly appropriate for an undergraduate audience.

- Ask a question and find a dataset that will have the data required
- Understand that there are fields within dataset
- Understand what the fields within a dataset mean
- Understand relationships between fields within a dataset
- Develop a question based on the data in a dataset
- Read and interpret charts, graphs, and other data visualizations

The competencies we came up with mirror existing information literacy tools that teach evaluation of information resources. For instance, many universities ask undergraduates to consider the factors that indicate a journal article is scholarly. Following a similar logic, we wanted students to identify what indicates that a data object is of good quality for re-use.

Once we identified the competencies listed above, we developed a logical model to explain the concepts applied to the critical evaluation of data objects. The result of this attempt to apply logical consistency to this topic was the data credibility checklist.

Data Credibility Checklist

We started from the idea that the students would be searching out new, unfamiliar data sources. We then sought to give the students tools to recognize a quality data object. Based on this vision, we brainstormed the following as important factors in recognizing quality in a data object:

1. Documentation - Is there a content map or guide of some sort? What is covered? What is not covered? Is there metadata included?
2. Authority - Who created the data? Who is managing it? Who paid for the data? What bias might be implicit? Is the data object currently maintained? Are there any references on how this data object has been used in the past? Are there clear release versions and updates information?
3. Format expectations -- Are there clear format expectations? What units are used? What fields are present? What naming conventions are used? Are the dates of creation or last update easily located?
4. Quality control -- Is quality control explicitly outlined? Who is in charge of checking for quality? What process do they use? How is missing data handled?
5. Human readable/machine readable -- Can a file be opened and a user understand the content? Is the file available for download in an open format? Is there a clear process to download?

From these factors, we developed the Data Credibility Checklist, an instructional tool for use with undergraduates being introduced to the concept of data consumption [Table 1].

Content map	**Authoritative**	**Format expectations**	**Quality control**	**Human readable/Machine readable**
What is covered ?	Who created the data?	What units are used?	Who is in charge of checking for quality?	Can you open a file and understand what is in it?
What is not covered?	Who is managing the data?	What process do they use?	Is the file available for download in an open source format?	
Is it relevant to my research question?	Who paid for the data?	How is missing data handled?	Is there a clear process for download?	
How is it relevant to my research question?	What bias might be implicit?			
Is there metadata included?	Is it currently maintained?			
	Has someone else used this data object for reuse in the past? How?			
	Are there clear release versions, updates with release dates?			

Case Studies

Two case studies follow in which Libraries faculty integrated data consumption competencies into the undergraduate curriculum through two existing upper-level undergraduate courses in Science and Nuclear Engineering.

NUC480 Essential Communication Skills for Nuclear Engineers/597W Nuclear Engineering Literature

For approximately 35 years (1974-2006), a one-credit, graduate level course titled Nuclear Engineering Literature (NUCL580) was taught at Purdue University, usually in the fall semester. This was a required course for all graduate students in the nuclear engineering program, taught by librarian. A one-credit technical communication (NUC581) class was introduced as another requirement for nuclear engineering students over the course of the 35-year history of this class. A new three-credit course (NUC480/597W) was created in Fall 2006 that combined the material in both one-credit courses, teaching communication skills for nuclear engineers, as well as discussing the literature of the field and its use in supporting statements in audience appropriate communications. The course is cross-listed as NUC480 (Essential Communication Skills for Nuclear Engineers) and NUC597W (Nuclear Engineering Literature) and allows upper-level undergraduates to focus on written communication skills. The graduate level is designed to help prepare graduate students for the research assessment paper that is a portion of the school Ph.D. qualifying exam. This new class is co-taught by a professor from nuclear engineering and an engineering librarian. Of the 16 weeks in the course, an engineering librarian teaches eight sessions, and covers topics such as patents, government documents and information, evaluation of information and web sites, copyright, proper attribution, and citations.

In Fall 2013, a data session was introduced into the class, partially due to the fact that the session on nuclear engineering datasets never felt complete or productive. With the growing interest in data information literacy, the authors had a perfect opportunity to introduce the concepts of data consumption and relate it to some of the entry-level engineering competencies set forth by the Center for Energy Workforce Development (2011).

The data session introduced the concepts of authority, quality, and accuracy and how they apply to the use of data repositories. The in-class exercise allowed the students to apply the criteria laid forth in the Data Credibility Checklist to two different datasets with the same name: Evaluated Nuclear Data File. The two datasets were published by different organizations: NNDC at http://www.nndc.bnl.gov/exfor/endf00.jsp and the International Atomic Energy Agency (IAEA) Nuclear Data Services at https://www-nds.iaea.org/exfor/endf.htm. The interactive session allowed the students to think critically about data and learned how to evaluate datasets based on a set of criteria. This class did not participate in pre- or post-class assessments; however, students commented on how helpful the session was and wished data information literacy concepts were taught in other classes.

SCI360 Great Issues in Science and Society

Great Issues in Science and Society (SCI360) is a course co-developed collaboratively by an Earth, Atmospheric, and Planetary Sciences professor and a Library Sciences professor. This course, one example of a required course for all College of Science majors, is intended to assist students in developing skills in professional communication (both written and oral) and lifelong learning. The majority of students completing this course are in their junior or senior year. The topic of the course evolved over several semesters to focus on "Natural Hazards- Data Driven Decision Making." The author, who co-developed the course, created a number of assignments walking students through a variety of data skills with a particular focus on data consumption. The types of data used by students in the class include Twitter data, data produced by multiple U.S. government agencies (notably Federal Emergency Management Agency, Department of Homeland Security, and National Weather Service), data produced by non-profit, non-governmental agencies responding to specific natural disasters, and

data collected by local and government agencies. The students select a geographic region within an area of impact for a natural disaster (i.e., a region within the impact zone of Superstorm Sandy). The students must then go through a multistep process of identifying research questions to answer and identify likely sources of data that may help to answer those research questions. Using the Data Credibility Checklist, the students locate datasets and identify the quality of the data source that they have found and note any concerns about credibility. The students then integrate the found datasets into an overall analysis which includes integration of data for visualization, and a final analysis that then is turned into a natural hazard policy recommendation for the impacted local government body.

Through the course of one semester, the students work intensively with the skills needed to consume and reuse data in the context of this one project. Students can freely explore and use whatever data they can identify with the goal of practicing data consumption skills repeatedly in the context of the assignment while simultaneously developing confidence in their own ability to think about, analyze, and communicate their own knowledge developed from the data. The deliverables throughout the semester were graded.

The checklist highlighted in this article was used in conjunction with a data inventory assignment. In this assignment, the student teams identify data sources appropriate for their project. The students were asked to brainstorm the following:

- What types of data are needed to fully answer our research questions?
- Who or what entities are likely to create the data that we need to answer our research questions?
- What are a few search terms that might help us to find the data that we need?
- What resources have we heard about in SCI360 so far that may have useful data?

The teams then create a spreadsheet that included a title for the dataset that they found, a URL, an author or authority for the data, the date of creation or publication for the dataset, a brief description or annotation of the contents of the dataset, and the team member who found and recorded the dataset. Each team member was responsible for using the Data Credibility Checklist to assess the quality of the resource prior to including the resource in the team spreadsheet.

This assignment was assessed as a part of the entire deliverable for the class. The curriculum was under a period of formative assessment when this assignment was piloted. The feedback given on the assignment was that the students had difficulty understanding what constituted a dataset. News articles and white papers were included in the spreadsheet. The author was able to identify datasets that those articles and white papers were based upon, but the students did not recognize that they needed to continue their search to locate the raw data. The data credibility checklist tool was considered to be useful, as was the exercise of performing the data inventory.

Support from the Library

In both case studies, an extended history of course integrated library instruction had already been established. For both courses, library faculty members originated the course in collaboration with disciplinary faculty members. In SCI360, the library faculty member was fully responsible for the development of all of the course assignments, which allowed for much tighter integration into the full course content. In the case of NUC480/597, the library faculty member had 1/6 of the course sessions to integrate course content. The development of integrated instruction opportunities such as these requires significant investments of time and resources to make the library faculty members available for the course.

Additionally, the library system makes time available for faculty members to explore new areas of research, particularly in research data management, and makes research data management specialists available to collaborate with subject liaisons for the development of curricula and other opportunities to integrate emerging skills into disciplinary curriculum.

Discussion

This pilot year of integrating data consumption skills has opened up many opportunities for future research. As has been acknowledged elsewhere in the literature, undergraduate students have gaps in their underlying computer use skill sets that had implications for developing data consumption skills (Nicholas, Rowlands, Clark, & Williams 2011). For instance, in SCI360, it became evident that many students did not understand the concept of a database, and what distinguished a database from other types of information. Therefore, a step must be taken to provide a basic level of understanding of what a database is, how it is similar or different from other sources of tabular data, and what it enables a student to do.

An additional area of exploration needs to look at the development of scaffolded activities across a course, where data consumption skills are gradually introduced and built upon through the semester. The initial introduction of data consumption skills in SCI360 happened about halfway through the semester. It became evident that the basic information referred to above would have been valuable from the first week of the semester, and could have been built upon in a way that gave students a broader base of understanding when it came to the data consumption assignment.

In the NUC480/597W data session, students were informed and comfortable with using a database. The challenge for them was in not knowing that these resources existed and how to find relevant data objects. It would be beneficial to integrate methods for locating data objects either early in the course or find additional opportunities to further incorporate information literacy into the sophomore and junior years of the undergraduate nuclear engineering curriculum.

Conclusion

This emerging area of instruction is a valuable contribution by libraries to the professional skills that early career STEM professionals take with them to the workplace. As the STEM areas increasingly rely on pre-existing data, either to validate or extend the scientific body of knowledge, students who have baseline knowledge of how to find, evaluate, and access data will have an advantage. The data credibility checklist and corresponding exercises also assist those pursuing post-graduate education as they embark on new research areas and developing research projects of their own. The exercise developed as an initial foray into this type of instruction provided an excellent opportunity to gather insight into improvements and changes that may need to be made in the future.

References

Borgman, C. 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* 63(6): 1059-1078. doi: 10.1002/asi.22634.

Calzada Prado, J., & Marzal, M. Á. 2013. Incorporating data literacy into information literacy programs: core competencies and contents. *Libri: International Journal of Libraries & Information Services* 63(2): 123-134. doi: 10.1515/libri-2013-0010

Carlson, J., Fosmire, M., Miller, C.C., & Nelson, M.S. 2011. Determining data information literacy needs: a study of students and research faculty. *Portal: Libraries & the Academy* [Internet]. 11(2):629- 657. Available from: http://muse.jhu.edu/journals/portal_libraries_and_the_academy/v011/11.2.carlson.html

Carlson, J., Johnston, L., Westra, B.O. and Nichols, M. 2013. Developing an Understanding of Data Management Education: A Report from the Data Information Literacy Project. [Internet]. Available from: http://docs.lib.purdue.edu/lib_fspres/11/

Chemical Information Skills. [Internet]. [Updated March 2012]. Washington DC: American Chemical Society. Available from: http://www.acs.org/content/dam/acsorg/about/governance/committees/training/acsapproved/degreeprogram/chemical- information-skills.pdf

Code of Ethics for Engineers. [Internet]. [Updated July 2007]. Alexandria (VA): National Society of Professional Engineers. Available from: http://www.nspe.org/sites/default/files/resources/pdfs/Ethics/CodeofEthics/Code-2007-July.pdf

Coulehan, M.B. & Wells, J.F. 2006. Guidelines for Responsible Data Management in Scientific Research. Office of Research Integrity, U.S. Department of Health and Human Services [Internet]. Available from: https://ori.hhs.gov/images/ddblock/data.pdf

Engineering Competencies. [Internet]. [Updated 2011]. Washington, DC: Center for Energy Workforce Development. Available from: http://cewd.org/Documents/EngCompModel.pdf

Kreisberg, A, Frank, R., Faniel, I., & Yakel, E. 2013. The Role of Data Reuse in the Apprenticeship Process. *Proceedings of the American Society for Information Science and Technology* [Internet]. 50:1-10. Available from: http://www.asis.org/asist2013/proceedings/submissions/papers/49paper.pdf

MacMillan, D. 2010. Sequencing genetics information: integrating data into information literacy for undergraduate biology students. *Issues in Science Technology Librarianship*[Internet]. 61. Available from: http://www.istl.org/10-spring/refereed3.html

National Research Council. 2009. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. Washington, DC: The National Academies Press. Available from: http://www.nap.edu/catalog.php?record_id=12615#toc

Nicholas, D., Rowlands, I., Clark, D., & Williams, P. 2011. Google Generation II: Web Behaviour Experiments with the BBC. *Aslib Proceedings: New Information Perspectives* 63(1): 28-45.

Ogburn, J.L. 2010. The imperative for data curation. *Portal: Libraries and the Academy* 10(2):241- 246.

Qin, J., & D'Ignazio, J. 2010. Lessons Learned from a Two-Year Experience in Science Data Literacy Education. *International Association of Scientific and Technological University Libraries, 31st Annual Conference* [Internet]. Available from: http://docs.lib.purdue.edu/cgi/viewcontent.cgi? article=1009&context=iatul2010

Stephenson, E., & Caravello, P.S. 2007. Incorporating data literacy into undergraduate information literacy programs in the social sciences: A pilot project. *Reference Services Review 35*(4): 525-540.

Wagenaar, T.C. 2004. Is There a core in sociology? Results from a survey. *Teaching Sociology* 32(1):1-18. doi: 10.1177/0092055X0403200101

Wallis, J., Rolando, E., & Borgman, C. 2013. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PloS One* 8(7). doi: 10.1371/journal.pone.0067332