11-3-2014

Thinking Critically About Data Consumption: Creating the Data Credibility Checklist

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

Megan R. Sapp Nelson
Purdue University, mrsapp@illinois.edu

Thinking Critically About Data Consumption: Creating the Data Credibility Checklist

Lisa D. Zilinski
Purdue University
West Lafayette, IN, USA
lzilins@purdue.edu

Megan S. Nelson
Purdue University
West Lafayette, IN, USA
mrsapp@purdue.edu

ABSTRACT
As 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. Accordingly, undergraduate STEM curricula is increasingly focused on research-based group projects that develop professional skills, building the professional portfolio needed for early career scientists, technologists, and engineers. This project works to develop new tools to implement basic data skills within the undergraduate disciplinary curricula. The first step in this process was to identify the competencies that are likely to be needed by those seeking data for reuse. Using competencies identified in the literature and via brainstorming, formatted similarly to the traditional model used by librarians to teach assessment of publication quality, a data credibility checklist was developed as the first tool in an emerging toolkit.

Keywords
Data reuse, data information literacy, data consumers, data management.

INTRODUCTION

The literature on data management education competencies is relatively new, emerging just in the past five years. In these initial forays into data competencies, the primary groups investigated have been faculty and graduate students from a variety of disciplines (Calzada Prado & Marzal, 2013; Carlson, Fosmire, Miller, & Nelson, 2011; Qin & D'ignazio, 2010). The reuse of data is highlighted in the literature as both a goal as enforced by granting agencies such as the National Science Foundation and National Institutes of Health (Calzada Prado & Marzal, 2013; NIH 2003; NSF, 2011). However, it seems evident to the authors that some data reuse skills are useful and even necessary in the undergraduate curricula. Stephenson and Schifter Caravello note in their pilot project that integrating data literacy skills "into a sociology course (rather than in a stand-alone information literacy course) would produce more opportunities for students to apply what they learn and for the instructors to assess learning in the context of doing sociology coursework" (2007, p. 525). With the rise of undergraduate research opportunities, and the increasing availability of data to be downloaded and reused, it is evident that those students who learn basic skills with data will have an advantage over those who do not (Kreisberg, Frank, Faniel, & Yakel, 2013; MacMillan, 2010).

Skills in working with data can be embedded in relevant ways throughout the undergraduate curriculum. Calzada, Prado and Marzal note 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). However, it has yet to be determined which competencies are appropriate and useful for an undergraduate level researcher. This poster represents an initial attempt to identify those skills and create a tool to assist students in practicing the competencies. The Data Credibility Checklist was developed and applied as part of two sessions on teaching data to undergraduate students in engineering and technology.

DEVELOPING DATA REUSE COMPETENCIES

Using the data competencies proposed in Carlson et al. (2011) and the tools developed during the Data Information Literacy grant, we identified discovery and acquisition of data, data conversion and interoperability and data quality and documentation as important for successful data reuse (http://wiki.lib.purdue.edu/display/ste/materials).

The authors discussed data management in the context of reuse and sharing and brainstormed the following attributes of data objects that were positive and leant credibility to the data content or were negative and detracted from the credibility of the data, as the case may be.

ASIST 2014,November 1-4, 2014, Seattle, WA, USA.

The competencies we came up with mirror existing information literacy tools that teach evaluation of information resources, such as the one hosted by the Purdue University Libraries (http://bit.ly/1rVmU6K). For instance, many universities ask undergraduates to consider the factors that indicate that 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.

Positive attributes of data objects that were brainstormed included:

- Human readable/Machine readable
- Clear format, including easily identifiable units of measure and labeling
- Explicit date of creation or publication
- Documentation, content map or guide of some sort that outlines labels/relationships
- Quality control process identified
- Option for download, including the option to download in an open source format
- Clear process to download

Negative attributes of data objects that were brainstormed included:

- Only human readable or machine readable
- Unclear authority including creator, creation date, and process by which data object was created
- No content guides to identify context
- No or poorly implement quality control, no process identified by which data object was checked for errors
- Unclear labels/ relationships with no documentation to explain
- Mixed formats or units of measure with no explanation
- No or an unclear process by which to download the data object
- Proprietary software formats as only option for download or no options for download

Once we identified the attributes of a data object that a student would look for to evaluate its suitability for reuse, we developed a logical model to explain the concepts and how they are considered and applied to critical evaluation of databases. The result of this attempt to apply logical consistency to this topic was the Data Credibility Checklist.

DATA CREDIBILITY CHECKLIST

As we began designing a newly created data session for a technical communications course for nuclear engineering students, we started from the idea that the students would be searching out new and unfamiliar data sources. We wanted to provide the students with tools that would help them learn how to recognize a quality database. The first set of competencies we discussed included:

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

From here, we wanted to outline specific tasks that could be completed in-class that would guide the students in defining the need(s) for specific data from a dataset:

- Ask a question and find a dataset that will have the data required
- Develop a question based on the data in a dataset
- Know what the databases contain

Based on our vision of introducing data information literacy concepts to the students in the session, we brainstormed the following five factors as important in recognizing quality in a database:

1. Content map - What is covered? What is not covered? Is it (or how is it) relevant to my research question?
2. Authority - Who created the data? Who is managing it? Who paid for the data? What bias might be implicit? Is the dataset currently maintained? Are there any references on how this dataset has been used in the past? Are there clear release versions and updates information?
3. 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 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 the users understand what is happening? Is the file available for download in an

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 fields are present?	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?	What naming conventions are used?	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?	Are the dates of creation or update easy to find?		
Is there metadata included?	Is it currently maintained?			
	Has someone else used this dataset for reuse in the past? How?			
	Are there clear release versions, updates with release dates?			

Table 1. Data Credibility Checklist.

open source format? Is there a clear process to download? (Zilinski, Nelson, & Van Epps, in press).

Using these factors as a basis for developing an instructional tool for use with undergraduate students being introduced to data consumption as concept, we created the Data Credibility Checklist (Table 1). The nuclear engineering data session then provided us with the opportunity to pilot some of the criteria developed for the Data Credibility Checklist in a real world setting.

CONCLUSION

This emerging area of data information literacy is a valuable contribution by libraries to the professional skills that early career STEM professionals take with them to the workplace. As professional and academic 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. This curriculum can also assist those pursuing graduate education as they embark on new research areas and developing research projects of their own.

The intervention 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. Case studies in engineering and technology have shown that support from the library in integrating data consumer skills instruction can be a valuable contribution to the development of career skills for future STEM professionals (Zilinski et al, in press).

DISCUSSION

The tool that was created, the Data Credibility Checklist, can be used in a variety of courses in a number of different disciplines. Most recently, this checklist and corresponding class assignments and exercises have been used in engineering and technology courses (Zilinski et al., in press) with varying degrees of success. It will be important to find additional ways to test the checklist and incorporate it into the STEM curriculum through for-credit courses, one-shot library instruction sessions, workshops, and research programs.

Programs such as the Summer Undergraduate Research Fellowship (SURF) can offer opportunities to incorporate the checklist into a hands-on research experience for undergraduate engineering, science, and technology students. We can also see the Data Credibility Checklist used in research methods courses and undergraduate research fellowships in non-STEM disciplines, where students are exposed to datasets in the field.

REFERENCES

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. & Nelson, M. S. (2011). Determining Data Information Literacy Needs: A Study of Students and Research Faculty. *portal: Libraries and the Academy*, *11*(2), 629-657.

Kreisberg, A, Frank, R., Faniel, I., & Yakel, E. (2013). The Role of Data Reuse in the Apprenticeship Process. In *Proceedings of the American Society for Information Science and Technology, 50,* 1-10. Montreal, Canada. doi: 10.1002/meet.14505001051

MacMillan, D. (2010). Sequencing Genetics Information: Integrating Data into Information Literacy for Undergraduate Biology Students. *Issues in Science Technology Librarianship, 61*. doi: 10.5062/F44F1NNK

NIH Data Sharing Policy and Implementation Guidance. (2003). National Institutes of Health. Retrieved from https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

NSF Proposal Processing and Review. (2011). National Science Foundation. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_3.jsp

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

Stephenson, E. & Schifter Caravello, P. (2007). Incorporating data literacy into undergraduate information literacy programs in the social sciences: A pilot project. *Reference Services Review, 53(4),* 525-540. doi: 10.1108/00907320710838354

Zilinski, L. D., Nelson, M. S., & Van Epps, A. S. (in press). Developing Professional Skills in STEM Students: Data Information Literacy. *Issues in Science and Technology Librarianship*