Data Visualization and Information Literacy

by Ryan Womack[1]

Abstract

Data visualization has grown in significance and complexity as the quantity of data and the technology supporting it have developed. Understanding and using data visualization is now a core skill that should be incorporated into information literacy goals by librarians and educators. Competency in data visualization is also closely related to data literacy and other quantitative literacies. Undergraduate students and other general learners should be exposed to the fundamentals of data visualization early in their education. This article proposes that evaluation, critique, and use of data visualization be the initial focus of education, and discusses some starting points for training in these three areas.

Keywords:Data Visualization, Information Literacy, Data Literacy.

Introduction

Data visualization has existed in certain forms for centuries. William Playfair (1786) is credited as the first to systematically use tools such as bar charts and line graphs in print to illustrate his arguments. Pioneers such as Charles Minard and Charles Marey (1878) continued to develop new techniques in the 19th century, with applications to social and natural sciences. Advances in statistical analysis entrained refinements in the visual expression and exploration of data, exemplified in the more recent work of William Cleveland (1985, 1993) and John Tukey (1977). But the contemporary imagination has been captured by the ever more rapidly developing forms of visualization generated by sophisticated and powerful computing techniques applied to large volumes of data. Today's data visualization, as expressed in interactive web graphics such as the enormous variety presented at Visual Complexity (2014), represents one of the most beautiful and intriguing flowers of "Big Data".

This new attention to data visualization raises the question of whether it has a place in general education, and most particularly in information literacy. Information literacy addresses the elements required for interpreting and understanding the informational content of the contemporary, technologically rich environment for communication, whether scholarly and otherwise. Information literacy standards and guidelines have been incorporated into the academic aims of institutions, to be promulgated and assessed by librarians and accrediting bodies. By its relative long-standing and well-developed framework, information literacy offers an appealing model to emulate for data visualization. The Association for College and Research Libraries' Information Literacy Competency Standards for Higher Education are a leading example of such a framework (ACRL, 2014). While information literacy began with a textual focus, the approach has been extended to encompass media literacy, numerical literacy, data literacy, and other literacies as they have emerged. Since data visualization is now emergent, and represents a major tool for the communication of complex results from large and often heterogeneous data sources, it is natural to consider data visualization as another type of literacy.

The more advanced reaches of data visualization encompass high-performance computing, advanced graphic design, sophisticated studies of the cognitive perception of visual imagery, and other expert research. For examples of the frontiers of research in this area, consider Linsen (2012) for medical imaging, Marchese and Banissi (2013) for humanities applications, and Huang (2014) for an overview of human-centered design in visualization. In fact, the field's rapid development has been recognized as a challenge by educators (Owen, 2013). This paper does not attempt to address or survey this vast range of material. Rather, it focuses on the data visualization skills and literacies that should form the foundational elements of the knowledge of a generally educated person today. These core elements should retain utility and validity even as the field changes.

Just as someone trained in information literacy can evaluate and use textual information with greater sophistication than the untrained, going beyond a simple and unreflective understanding of data visualization will improve the communication and analytical skills of students. After all, data visualization is just another way of presenting, interpreting, and

using information. It is time to bring data visualization into the literacy training offered by librarians and educators.

Literacies: Data, Statistical, Quantitative, and more

While data visualization has not yet entered the literature of library and information science [LIS] to a large degree, a body of work has developed on the various literacies appropriate to the numeric side of LIS that the International Association for Social Science Information Services and Technology (IASSIST) represents. Gray (2004, p. 24) emphasizes statistical literacy as an important value, and hints at the importance of critical analysis of data graphics, saying 'We live in the Information Age with rapid distribution of news and content, where content is often overlooked in favor of images, and at a time when more and more statistics and data products are being made available to a larger and less data-literate audience'. She goes on to argue for an expanded role of libraries in training others in statistical literacy concepts. Schield (2004) addresses the differences and overlaps between the concepts of data literacy, statistical literacy, and information literacy. He argues that both statistical literacy and data literacy need to be taught more widely. Stephenson and Caravello (2007) describe the challenges of implementing a classroom instruction program on data literacy.

Data information literacy [DIL] is a shift of emphasis that has emerged out of the increased attention to research data, whether the data is big or not. While many of the concepts of data information literacy are not new, and are certainly well-known to the social science data community, what is new is the emphasis on the educational mission of academic institutions to train new scholars in this set of skills. Wright et. al. (2012) identify 12 categories of training needs for data information literacy, one of which is data visualization. Certainly one place for literacy associated with data visualization is under this organizing rubric. Carlson et. al. (2013a) suggests that faculty do not necessarily feel they have all of the knowledge necessary to train their students in DIL, and welcome assistance from others in the education effort.

Data information literacy is closely tied to the educational and outreach efforts surrounding research data management, and thus parallels the content of data management training. Data management training courses have sprung up at universities around the world, such as the University of Minnesota (Jeffryes and Johnston, 2013) or the University of Edinburgh (Rice and Haywood, 2011). These courses are typically focused on graduate students in the disciplines, teaching them how to handle and present their research findings and data. They are targeted to students who are past their general phase of learning, and who will therefore have many specific ways of handling data visualization that are appropriate and customary to their disciplines. Designing general data visualization content at this stage is therefore difficult. Besides, the data training agenda is crowded, and there is little time to add training to meet new goals. Qin and D'Ignazio (2010) mention data visualization as only one of 20 topics in a science data training context. Despite these difficulties, a brief reminder of the need for thoughtful and effective data visualization may be appropriate in the graduate context, as well as pointers to places to learn more information. Particular disciplines may use data visualization extensively, but the specialized techniques of the discipline are best addressed by specialists, not by generalists like librarians from outside the discipline.

In other contexts, terms such as quantitative literacy or numeric literacy are used to describe the skills needed, placing the focus on the mathematical aspects of understanding data. Certainly quantitative reasoning is an important part of educational goals and may include making numerical sense out of graphs and charts. Other literacies could be adduced, but the purpose of this article is not to provide precise definitions of the boundaries between these interrelated forms of literacy, or to introduce new terminology for data visualization information literacy. Instead, data visualization should be thought of as a component that relates to many aspects of literacy as described. Skills in data visualization support a range of literacies and should be viewed as complementary to them.

Data Visualization as a Basic Component of Information Literacy

If data visualization is not to be taught as separate or specialized content, how can it best be integrated into general education goals? While the LIS literature has recognized that data visualization has potential significance (Thomas, 2012) and is a topic whose time has come (Bell, 2010), specific goals for data visualization have not been articulated. A few skills relevant to data visualization are mentioned by interviewees in the Data Information Literacy Project, but not in a general education context (Carlson, 2013b). This article represents a further step towards defining data visualization's place in general education on information literacy.

Although the ACRL Information Literacy Competency Standards for Higher Education (ACRL, 2014) are undergoing revision, the basic competencies are currently as follows:

1. The information literate student determines the nature and extent of the information needed.
2. The information literate student accesses needed information effectively and efficiently.
3. The information literate student evaluates information and its sources critically and incorporates selected information into his or her knowledge base and value system.
4. The information literate student, individually or as a member of a group, uses information effectively to accomplish a specific purpose.
5. The information literate student understands many of the economic, legal, and social issues surrounding the use of information and accesses and uses information ethically and legally.

The first competency can be considered a preliminary stage that defines the research project. This is not to minimize its importance. Phetteplace (2012, p. 97) states, 'It bears repeating: the first step to good data visualization is good data. Most of the thought and effort should go into to [sic] collecting and analyzing data; playing with visuals until you find a compelling option is the reward for your due diligence.' However, most of the initial research definition, data preparation, and other steps do not directly relate to data visualization itself.

The second competency dealing with access does have a visualization component, but one that is arguably not an analytical one. Because most visualizations will be encountered in the course of general research into a topic, via websites and publications, the information seeker may not need to learn new skills just to tap into data visualizations. While there are many applications of data visualization, such as the mapping of census data, where graphical elements are prominent, and there are some sites that specialize

in creating visualizations of preexisting data, effective and efficient access to data visualization is not a universal or high-priority need. Initial steps in general education for data visualization should lie elsewhere. However, that does not preclude librarians and educators from building data visualization access tools into their repertoire of resources and guides.

The last of the five competencies, relating to ethical and legal considerations, is a general proviso that applies to all information use. Data visualization may engender some unique ethical and legal considerations, such as the safeguarding of individually identifiable information in a graph of social network relationships, or whether derived data distilled into images for distribution abides by terms of use for the data. However, these are more likely to arise in specialized contexts, and are not appropriate for introductory educational efforts.

The competencies relating to evaluation and use (3 and 4) are the most relevant to data visualization in practice, and the most appropriate for incorporation of data visualization goals into introductory outreach. The ACRL Standards focus on the intellectual framework required to achieve the competencies, not the use of specific technologies or tools. For data visualization, the focus on the intellectual framework should remain the same, because the tools will change rapidly.

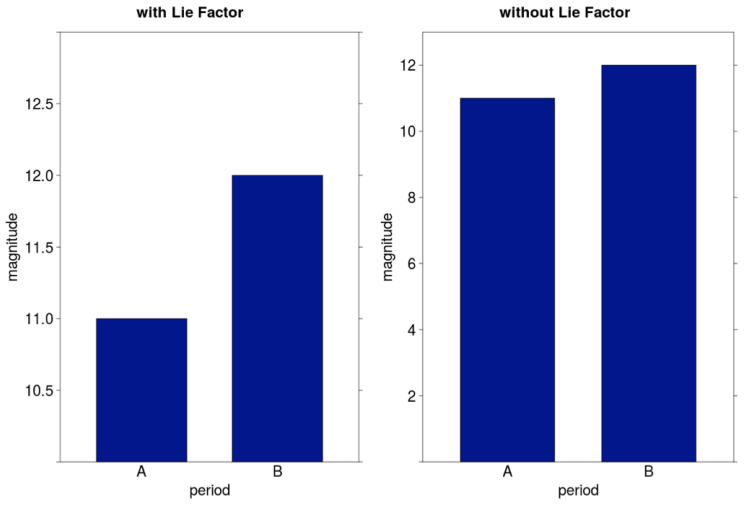
The ACRL has also developed Visual Literacy Competency Standards for Higher Education (ACRL, 2011), but Visual Literacy as defined in these standards focuses more on the interpretation and use of visual imagery and media outside of the data context. These standards tangentially refer to the context of data visualization in Standard Four, part 1.f, as follows: "Determines the accuracy and reliability of graphical representations of data (e.g., charts, graphs, data models)". Elsewhere, the standards remain focused on images in general. Still, there are parallels with the broader ACRL Information Literacy standards and the data visualization concepts discussed in this paper. Standard Three specifies the ability to "interpret and analyze" visual imagery, which is related to the concept of critique discussed below. Standard Four of the Visual Literacy Standards deals with evaluation, and Standard Five deals with use. See Hattwig et. al. (2013) for further discussion of the Visual Literacy standards.

Evaluation, Critique, and Use

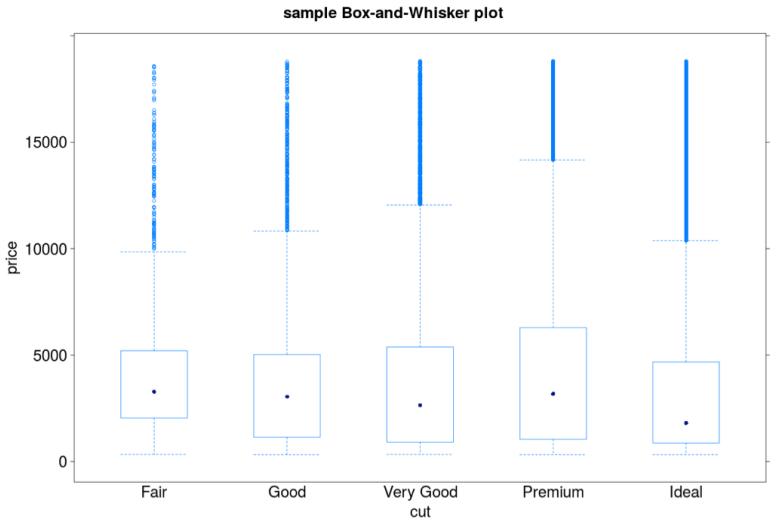
The third ACRL Information Literacy Standard requires students to evaluate sources critically, and to incorporate them into their knowledge base. There is enough work in both the evaluation and critique of data visualization resources that considering these elements separately is justified.

Evaluation, as used here, refers to the basic questions that must be asked of a particular data visualization to establish its quality, accuracy, and reliability. The danger inherent in a visual medium is that the power of the image will overwhelm the substantive content that it represents. When presented with a data visualization, the user should 'interrogate the image' and establish the source of the data, the reliability of the source, and the appropriateness of the visualization for the kind of data. If the underlying data is of poor quality, no amount of elegant graphics can compensate for this. Understanding the methodology that produced the data is also essential (Gray, 2004).

Concepts such as Edward R. Tufte's Lie Factor (Tufte, 2001) can be introduced to provide a framework for systematically checking the level of distortion inherent in an image. The Lie Factor is computed by dividing the size of the effect shown in the graphic by the actual size of the effect in the data. For example, an increase in magnitude from 11 to 12 can be made to appear as a doubling, if we set the baseline at 10 (+1 vs. +2 over the baseline).



Students should be introduced to a basic range of visualization types (bar, line, scatterplots, box and whiskers plots, etc.) and learn appropriate uses for each. For example, connecting points into lines to show a time series trend is a good idea, while connecting points in a scatterplot usually has no meaning and can be misleading. A box-and-whisker plot can summarize the variation of a dense dataset. The R package ggplot2 includes a sample dataset of 50,000 diamond prices with related characteristics. For example, in Figure 2, the box-and-whisker plot of diamond prices classified by the cut of the diamond shows a considerable number of high-priced outliers beyond the "whiskers", but a relatively compact central range of prices covering the 25th to 75th percentile of the data within the boxes.



Students should be aware that there are many other methods available to them, recognizing that time constraints may limit what is presented in an introduction to the topic. One resource describing such methods is the Periodic Table of Visualization Methods (2014) which displays an extensive and suggestive classification of available techniques, even if the classification is not quite as rigorous as the Periodic Table of Elements.

Students should learn to evaluate data visualizations that they plan to incorporate into their research, just as they weigh and evaluate textual sources to cite. Evaluation answers the fundamental question of whether or not a particular data visualization is sound and reliable to use as a basis for scholarship.

Critique, in the sense proposed here, is evaluation raised to the next level, and attempts to answer the question of whether or not a particular data visualization is among the best possible in its domain for a particular application.

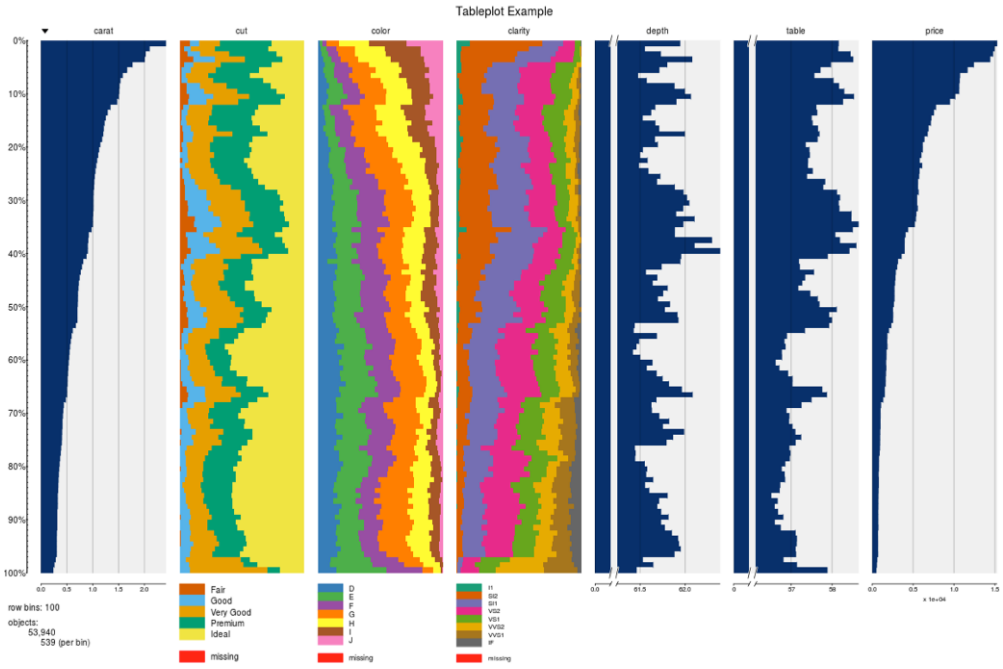
Fox and Hendler (2011) argue, among other things, that as web technologies have improved the ease of implementing more complex and interactive data visualizations, science should make greater use of these techniques for the masses to explore and interpret data. As research uses more sophisticated visualization techniques, students will need to understand and appreciate these nuances.

Critique involves comparison among different data visualizations in order to develop understanding of which visualizations exemplify best practices. General principles such as striving for clarity, avoiding clutter, and emphasizing the most relevant data apply to most visualizations. In addition, the best visualizations enable rich understanding of complex datasets with relative ease. The techniques developed to produce these visualizations are both an art and a science, and should be appreciated and emulated by students, who should also learn to be cautious of oversimplification and approaches that sacrifice features of the data in favor of graphical elegance. For example, Fisher, Dempsey and Marousky (1997) show that despite more complex 3D forms being appealing to the eye, simpler 2D graphics were preferred when the task required actual extraction of information from the graphs. New methods, such as tableplots, are also being developed to simplify the visualization of large datasets (Tenneke, de Jonge and Daas, 2013). Figure 3 is a tableplot based on the ggplot2 diamond price dataset which provides a snapshot of the relationship among variables in this large dataset.

Data visualization should remain a tool in service of scientific knowledge, not an end in itself, in spite of the undeniable aesthetic attractiveness of many of the best visualizations. A good discussion of the difference between pretty infographics and high quality and accurate representations of data is Beauty is as Beauty Does (Lyons, 2011). Senay and Ignatius (1999) provide a useful schema of data visualization types and their attributes, along with extensive guidelines for their use at Rules and Principles of Scientific Data Visualization. Their rules are distilled from the works of several pioneers in the field such as Cleveland, Tukey, and Tufte. Senay and Ignatius emphasize the need for further research, saying 'to gain better understanding of the effectiveness and expressiveness of various visualization primitives, it is essential that empirical studies of visualization techniques should be undertaken. Otherwise, the techniques that are generated may be viewed as meaningless 'pretty pictures.' ' Lessons distilled from the work of researchers in the field should be used to transmit some of these concepts to students at the undergraduate level and higher, using relevant examples.

Use is the third proposed area of focus. Use puts the emphasis on putting data visualization into practice. With the menu of visualization types and rules for their use previously introduced, students should have the opportunity to practice doing their own data visualizations. Guiding students through introductory examples, working in sandbox environments, and using various demos and examples will lead students through the process of actually developing their own visualizations based on the choices before them. More experience with actual creation of data visualizations will develop skill and wisdom in making good selections, and will reinforce the concepts learned about evaluation and critique. The actual form that the 'Use' component takes will be determined by current technology and research needs in a particular setting.

Owen et. al. (2013) classify data visualizations into three areas: 1) scientific or data visualization, in which the data dimensions correspond to physical reality (e.g., remote sensing); 2) information visualization, for multi-dimensional data from a defined field of interest; and 3) visual analytics, which is massive and heterogeneous. These boundaries are fluid and may overlap, but this is one potentially useful schema for types of



visualization. Each of these types will have its own software and design decisions. Owen et. al. go on to address several areas in the creation of visualizations: the User, the Design Stage, Visual Presentation, Interaction Techniques (required for visual analytics), Communication, Collaboration, Evaluation, and Displays. Not all of these categories are concerned with the analytical, intellectual literacy skills that relate to information literacy, but again this can serve as a template for developing practical examples that allow students to create and use visualizations.

As a concluding example from the literature, Kelleher and Wagener (2011) offer 10 simple guidelines that apply to almost any kind of data visualization. These guidelines are distilled from the theoretical literature in a reliable way, and are very useful for basic training in data visualization. They range from [#1] 'create the simplest graph that conveys the information you want to convey' to [#10] 'select an appropriate color scheme based on the type of data', and are illustrated with examples of effective and poor practice. Could this serve as an Elements of Style (Strunk and White, 1979) for graphics? Perhaps. While no one would argue that Strunk and White are forever definitive and prescriptive, few would argue against the benefits of attempting to identify and reinforce specific fundamental and useful principles. More importantly, practicing such rules reinforces literacy. Such preferred practices have evolved for static graphics, but recommendations for the new world of interactive and dynamic graphics have not yet been distilled into concise and widely accepted principles. As education and literacy efforts for data visualization grow, the body of knowledge describing these best practices will grow in parallel.

Conclusion

As argued here, evaluating, critiquing, and using data visualizations have become an essential literacy, one that is now required to understand and make use of the information products of our data-driven age. By focusing on a limited set of the most fundamental principles of evaluation, critique, and use, data visualization can be incorporated into introductory information literacy efforts targeted at undergraduates and general learners. Librarians and other educators should be equipped to instruct in these areas. Data librarians and other data professionals are clearly positioned to lead these efforts.

While 'the devil is [still] in the details' (Womack, 2014), and the implementation of actual instruction programs will depend greatly on the preferred technologies and topics appropriate to each institutional environment, helping students make better use of information as presented via data visualization supports the core goals of information literacy. Other quantitative, data, and numerical literacies will also benefit from a dose of data visualization. Most importantly, students exposed to more sophisticated data visualization training will be better able to understand, not only data visualizations, but the world around them, and to develop the skills to influence their world.

References

ACRL (2011) Visual Literacy Competency Standards for Higher Education. [Online] Available from: *http://www.ala.org/acrl/standards/visualliteracy*. [Accessed: 21 September 2014].

ACRL (2014) Information Literacy Competency Standards for Higher Education. [Online] Available from: *http://www.ala.org/acrl/standards/informationliteracycompetency*. [Accessed: 22 June 2014].

Bell, M. (2010) Do You See What I See? Multimedia and Internet @ Schools. March/April 2010, pp. 39-41.

Carlson, J. et. al. (2013a) Developing an Approach for Data Management Education: A
Report from the Data Information Literacy Project. International Journal of Digital Curation. 8(1), pp. 204–217. *http://dx.doi.org/10.2218/ijdc.v8i1.254*

Carlson, J. et. al. (2013b) [Online] Findings from the DIL Interviews: Data Visualization and Representation. Available from: *http://docs.lib.purdue.edu/dilsymposium/2013/dilcompetency/12/*. [Accessed: 22 June 2014].

Cleveland, W. (1985) The Elements of Graphing Data. Wadsworth.

Cleveland, W. (1993) Visualizing Data. AT&T Bell Laboratories.

"Findings from the DIL Interviews: Data Visualization and Representation"
http://docs.lib.purdue.edu/dilsymposium/2013/dilcompetency/

Fisher, S., Dempsey, J. and Marousky R. (1997) Data Visualization: Preference and Use of Two-Dimensional and Three-Dimensional Graphs. Social Science Computer Review. 15, p. 256. <*http://dx.doi.org/doi:10.1177/089443939701500303*>

Fox, P. and Hendler J. (2011) Changing the Equation on Scientific Visualization. Science. 331, 11 February 2011, pp. 705-708.

Gray, A. (2004) Data and statistical literacy for librarians. IASSIST Quarterly. 28 (2/3), pp. 24-9.

Hattwig, D. et. al. (2013) Visual Literacy Standards in Higher Education: New Opportunities for Libraries and Student Learning. portal: Libraries and the Academy. 13(1), pp. 61-89. <http://dx.doi.org/doi:10.1353/pla.2013.0008>

Huang, W. (ed.) (2014) Handbook of Human Centric Visualization. Springer.

Jeffryes, J. and Johnston, L. (2013) An E-Learning Approach to Data Information Literacy Education. 120th ASEE Annual Conference and Exposition, Paper #6956, June 23-26 2013.

Kelleher, C. and Wagener, T. (2011) Ten guidelines for effective data visualization in scientific publications. Environmental Modelling & Software. <http://dx.doi.org/doi:10.1016/j.envsoft.2010.12.006>

Linsen, L. , et. al. (eds.) (2012) Visualization in Medicine and Life Sciences II: Progress and New Challenges. Springer.

Lyons, R. (2011) Beauty is as Beauty Does. [Online] Available from: *https://libperform.wordpress.com/2011/10/28/beauty-is-as-beauty-does/* [Accessed: 24 June 2014].

Marchese, F. and Banissi E. (eds.) (2013) Knowledge Visualization Currents: From Text to Art to Culture, Springer.

Marey, E. (1878) La Méthode graphique dans les sciences expérimentales et principalement en physiologie et en médecine. G. Masson.

Owen, G. et. al. (2013) How Visualization Courses Have Changed over the Past 10 Years. IEEE Computer Graphics and Applications, July/August 2013, pp. 14-19.

Periodic Table of Visualization Methods (2014) [Online] Available from: *http://www.visual-literacy.org/periodic_table/periodic_table.html* . [Accessed: 24 June 2014].

Phetteplace, E. (2012) Effectively Visualizing Library Data. Reference and User Services Quarterly. 52(2), pp. 93-97.

Playfair, W. (1786) Commercial and Political Atlas: Representing, by Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, during the Whole of the Eighteenth Century republished in The Commercial and Political Atlas and Statistical Breviary, 2005, Cambridge University Press.

Qin, J. and D'Ignazio, J. (2010) The Central Role of Metadata in a Science Data Literacy Course. Journal of Library Metadata. 10, pp. 188–204. <*http://dx.doi.org/10.1080/19386389.2010.506379*>

Rice, R. and Haywood, J. (2011) Research Data Management Initiatives at University of Edinburgh. The International Journal of Digital Curation. 6 (2), pp. 232-244.

Schield, M. (2004) Information Literacy, Statistical Literacy and Data Literacy. IASSIST Quarterly. 28 (2/3) pp. 6-11.

Senay, H. and Ignatius, E. (1999). [Online] Rules and Principles of Scientific Data Visualization. Available at *https://www.siggraph.org/education/materials/HyperVis/percept/visrules.htm* [Accessed: 24 June 2014].

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

Strunk, W. and White, E. (1979) The Elements of Style. Third edition. Macmillan.

Tennekes, M., de Jonge, E. and Daas, P. (2013) Visualizing and Inspecting Large Datasets with Tableplots. Journal of Data Science. 11, pp. 43-58.

Thomas, L. (2012) Think Visual. Journal of Web Librarianship. 6, pp. 321–324. <*http://dx.doi.org/doi:10.1080/19322909.2012.729388*>

Tufte, E. (2001) The Visual Display of Quantitative Information. Second edition. Graphics Press.

Tukey, J. (1977) Exploratory Data Analysis. Addison-Wesley.

Visual Complexity (2014) [Online] Available from: *http://www.visualcomplexity.com/vc*/. [Accessed: 22 June 2014].

Womack, R. (2014) Data Visualization and Information Literacy, IASSIST Annual Conference, Toronto, Canada, June 5, 2014. 2014. <http://dx.doi.org/doi:10.7282/T37P8WM1>

Wright, S. et. al. (2012) A Multi-Institutional Project to Develop Discipline-Specific Data Literacy Instruction for Graduate Students. Libraries Faculty and Staff Presentations. Paper 10. *http://docs.lib.purdue.edu/lib_fspres/10*

NOTES

1. Ryan Womack is Data Librarian at Rutgers, The State University of New Jersey (New Brunswick campus). Correspondence may be addressed to *rwomack@rutgers.edu* or 169 College Avenue, New Brunswick, NJ 08901 [USA].