Author Names Removed
Affiliation(s) Removed

Proficient Use of Open Data Requires These Core Information Skills: An Open Data Community Perspective (Paper)

Abstract:

Expanding access to open data, such as government data and research data, requires that we consider how citizens and stakeholders can best access the value these data hold. Should individuals rely on an intermediary to create information products from the data, or should they dive in and work with raw data? Building on previous work defining a core set of data literacy skills, we convened a workshop with 34 open data professionals to define the core set of skills for working with open data: "open data literacy". Analysis of their perspectives reveals a focus on non-technical skills, like creativity, curiosity, and critical thinking, as a priority over technical skills like coding and visualization. We describe their perspective in detail, and reflect on the significance of our findings for information professionals.

1. Introduction

We are a data-rich society; perhaps even data-driven (Pentland, 2013). The growth of data is fueled, in part, by governments releasing machine-readable data in open formats, and researchers being asked, lobbied, and/or compelled to release raw data publicly, when possible. This open data is believed to have positive impacts that include encouraging an informed populace, supporting government transparency, and enabling value-added services (Davies, 2010). While developers are embracing open data, in combination with analytics and machine learning, to create useful applications (Jetzek et al., 2014; Kitchin, 2014), we suggest it is also important that non-technical users be comfortable working with open data. In addition to the benefits to these users, the engagement of domain experts is essential to ensuring the quality and accuracy of open data (Colborne and Smit, 2017).

The goal is to transition from being data-rich to being information-rich and knowledge-rich, for which we need both data scientists and domain experts capable of working effectively with data. The McKinsey Global Institute suggested that at current training rates, in the US alone there will be 1,500,000 more jobs than “data-savvy” analysts and managers (Manyika et al., 2011); IDC suggests a similar number (Vesset et al., 2014). The more common term for these skills is *data literacy*, or the ability to comprehend, create, criticize, and communicate data. It is the first level of the tri-level literacy, fluency, mastery scale. In short, data-literate individuals have the knowledge, understanding, and skills to connect people to data.

The 2015 report on Strategies and Best Practices for Data Literacy Education (Ridsdale et al., 2015) synthesized a set of competencies comprising data literacy from dozens of primary sources in the literature, yet cautioned that few of these had been validated as essential in practice. While the report mentions open data, there was no effort to identify which competencies were necessary for this general type of data that all citizens have access to, regardless of their background, location, or occupation.

In this paper, we identify and quantify the essential skills for working with open data, based on the input of a group of open data experts. We conclude from the skills identified that information skills, in addition to some technical confidence and ability, are at the core of effective open data skills.

2. Methods

We hosted a workshop titled "Data Literacy and Open Data" at the Canadian Open Data Summit 2016 in Saint John, New Brunswick. The initial goal of this workshop was to communicate the current state-of-the-art in data literacy scholarly work with a set of practitioners in Open Data, to bridge these two communities. To engage the audience, our workshop committed to providing "facilitated breakout discussions on the core skills needed to be data literate, data fluent, and data masters in the context of open data", and to produce a white paper summarizing the workshop outcomes. This paper makes secondary use of the anonymous data collected during this workshop.

The conference attracted attendees from government, the private sector, NGOs, and education; from across the country, with New Brunswick heavily represented. Attendees self-selected for a high level of interest and expertise in open data. We did not collect demographic information from our 34 workshop attendees.

All open data experts attended an introductory presentation defining literacy and fluency, as well as providing background information on data literacy from the 2015 report (Ridsdale et al., 2015). No definition or sets of skills from that report were included in the presentation. The group was then split into six roughly equal groups at six different tables, and provided with flip chart paper and markers. They were asked "What skills / abilities / competencies / understanding does a citizen need to engage directly with Open Data?", and to "brainstorm the skills you believe an individual needs in order to benefit from Open Data". After 30 minutes of thinking time, each group was asked to report their results to the group. Finally, each attendee was given 10 stickers and asked to vote for the skills they personally considered the *most* important. This is a standard "dotmocracy" exercise, and all attendees seemed comfortable with the process. (On average, each attendee used only 8 of their stickers). The result was a list of skills – with inevitable overlap and duplication – and a number of votes for each skill.

After the workshop, we digitized and analyzed the skills and the score for each skill. The skills were aggregated, combining similar skills that originated from different groups. Based on the skills identified, we identified an appropriate grouping ("skill area") to further categorize each skill. The number of votes was calculated for each skill area based on the total number of votes of all skills in that area; this means it was possible for experts to vote more than once for a skill area.

3. Results and Discussion

Originally, there were eighty-one skills identified by the participants. Four groups used a single sheet of flip chart paper (9, 12, 12, and 11 skills); one used two sheets (20 skills); and one used four sheets (17 skills). After skills with similar meanings were combined, and some skills excluded for being too general or specific, there were forty-two skills. Table 1 lists the most popular (by vote) unique skills identified by the participants. These forty-two were then categorized thematically into seven skill areas. The total number of votes for the skills in each is summarized in Table 2; we define each area as follows:

Analysis: The ability to assess data to determine its relevance, or its value, or simply what makes it interesting. Which data sets can you link? What patterns or trends do you see in the data? What conclusions can you reasonably reach based on the data you have?

Attitude: These skills encompass personality traits which are useful when dealing with open data. Examples include patience, curiosity, and common sense. They also included useful mind sets, indicating that engagement and collaborative attitudes were important. Participants indicated during the session they were not always comfortable thinking about these as skills – they were more attributes, or qualities. There was disagreement on whether these attributes could be learned or taught.

Skill	Votes
Basic statistics & math	19
Pattern Recognition	17
Basic computer literacy	16
Ability to judge sources; where is the data coming from	15
Skepticism	13
Communication Skills	13
Data Visualization	11
Patience & Focus	10
Curiosity	10
Awareness of bias	10

Table 1: The top individual skills required to work directly with open data, as identified by a panel of experts.

Awareness: These skills involve general awareness about the availability of open data, combined with understanding the big picture of a specific domain or context. Included would be skills like knowing where and how to access open data, how to search for data sources, and knowing when data is truly "open".

Communication: These skills deal with the communication of data—both by publishing open data, and about communicating with other team members about the data. This included being able to produce, and comprehend, data visualizations.

Computer Knowledge: This concerns the technical aspects of data, such as software and programming knowledge, technical data skills, and data management/sharing experience.

Critical Thinking: This is concerned with being able to assess and analyze the datasets themselves—framing questions appropriately, understanding bias, being skeptical, and in general understanding the limits of data.

Numeracy: Open data often involves numbers; participants identified various levels of statistics knowledge as essential, some of which might strain the definition of "numeracy". Participants wanted to see some understanding of descriptive statistics, data distributions, and various types of data (e.g. quantitative versus qualitative data).

Some skills listed on the pages were not categorized, as they were too specific ("datatype disambiguation") or too general ("ability to read/write", "Internet access").

Experts appeared to agree there is a broad range of skills drawn on when working with open data, and a set of essential attributes that make an individual more effective. We would suggest a single individual is unlikely to possess all of these skills and attributes, and that collaborative work may be required even for casual, citizen-led projects.

While "technical" skills were identified as important, it was often at a basic level: while 16 participants identified some level of computer literacy as being important (e.g. basic Excel skills), only 6 voted for "some kind of programming skills", with 1 voting for "coding (helpful but not required)". Similarly, math and statistics were identified as important, but the concepts named are taught at the high school level.

Skill Area	Number of Skills	Total Votes
Analysis	7	35
Attitude	7	32
Awareness	7	38
Communication	5	34
Computer Knowledge	6	27
Critical Thinking	5	47
Numeracy	5	41
TOTAL	**42**	**254**

Table 2: The aggregate Skill Areas and their total vote counts

Of note is the prevalence of "soft skills", both in the brainstorming activity and in the voting stage. Skepticism, awareness of bias, and understanding the source of data are all in the top 10, and all are references to the general skill area of critical thinking. These, along with communication skills and focus, would be advertised as learning outcomes of most post-secondary degree programs. There is an implicit suggestion that technical skills are important up to a point, but that non-technical or "soft" skills are equally crucial. In fact, only two of the seven skill areas (and 68 of the 254 votes) would be considered technical skills, and the level of skill in these categories was often identified as being basic computer or numeric literacy. (It should be noted that different people have different definitions for "literacy" in these areas.)

These results confirm many of the competencies mentioned in the 2015 data literacy report (Ridsdale et al., 2015). They also support the contention in the literature that data literacy shares the same theoretical grounding as information and statistical literacies (Hogenboom, Holler Phillips, & Hensley, 2011; Koltay, 2014).

There are limitations to this data, most notably the relatively casual conduct of the workshop due to its intended function as a knowledge translation activity, rather than a data collection activity. The sample of experts included was a convenience sample, and may not reflect the open data community more broadly.

4. Conclusion

The skills identified align quite closely to the competencies of a modern information professional. Future work could examine the learning outcomes of ALA-accredited degrees to establish this link more reliably, though our experience is that practicing information professionals already perceive and acknowledge this link. The remaining question is how to communicate these competencies to the general public, to improve access to and comprehension of the volume and variety of open data available to us. Information professionals should, and must, play a key role in ensuring that these competencies are communicated to a broader population.

Reference List:

Colborne, A. and Smit, M. (2017). Identifying and mitigating risks to the quality of open data in the post-truth era. In *Proceedings of the IEEE International Conference on Big Data (Big Data)*, p. 2588–2594.

Davies, T. (2010). Open data, democracy and public sector reform: A look at open government data use from data.gov.uk. Edited version of Masters dissertation available from http://practicalparticipation.co.uk/odi/report/, University of Oxford.

Hogenboom, K., Holler Phillips, C.M., and Hensley, M. (2011). Show me the data! Partnering with instructors to teach data literacy. *ACRL 2011,* Philadelphia, Pennsylvania. 410 417.

Jetzek, T., Avital, M., and Bjorn-Andersen, N. (2014). Data-Driven Innovation through Open Government Data. *Journal of Theoretical and Applied Electronic Commerce Research*; Curico, 9(2):100–120.

Kitchin, R. (2014*). The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Koltay, T. (2014). Big data, big literacies? *TEMA, 24*, 3 8.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity* [report]. McKinsey Global Institute. Pentland, A. (2013). The data-driven society. *Scientific American*, 309(4):78–83.

Ridsdale, C., Rothwell, J., Smit, M., Hassan, H. A., Bliemel, M., Irvine, D., Kelly, D., Matwin, S., and Wuetherick, B. (2015). *Strategies and best practices for data literacy education: Knowledge synthesis report*. Dalhousie University.

Vesset, D., Olofson, C. W., Schubmehl, D., McDonough, B., Woodward, A., Stires, C., Fleming, M., Nad- karni, A., Zaidi, A., and Dialani, M. (2014). *IDC FutureScape: Worldwide Big Data and Analytics 2015 Predictions* [report]. International Data Corporation.