

Preprint of submission to *The Journal of Community Informatics*, Sept. 2016

## Data literacy conceptions, community capabilities

Paul Matthews

Department of Computer Science and Creative Technologies

UWE Bristol

### Abstract

*To delineate and describe the data literacy concept, a core set of data competencies are identified, and a further four varieties – each with a different focus of attention – are described. These are named research (an academic focus), classroom (a secondary educational focus), carpentry (a practical training focus) and inclusion (a community development focus). It is argued that the inclusion focus helps data literacy to be construed as a community capability. The capabilities approach enables us to see data literacy as conferring a holistic freedom to operate, where the technical gains only make sense within a framework of social functions and goals.*

### Introduction

Data literacy as a concept carries ambiguity (IODC, 2015). A broad definition is it is that which “enables individuals to access, interpret, critically assess, manage, handle and ethically use data” (Prado & Marzal, 2013). Yet a range of alternative, partially overlapping conceptions exist. Sources of variation include the sector of origin, the relative weight and treatment given to different data-related competencies and the extent to which the concept should extend to its *framing* context, including individual or organisational goal-setting. Equally important for community informatics, however, is the extent to which data literacy can be a pathway to new capacities within locally-focused groups. Here, it is useful to consider enhanced knowledge as providing new rights and freedoms over and above possible material benefits. For this reason, steering the concept of data literacy towards a *capabilities* conception is useful.

This article looks at an overarching general competency framework, followed by four conceptions of data literacy – identified by text mining – which vary in their explicit emphasis. Then, Amartya Sen’s *capability approach* from development economics is proposed as a way to help emphasise the necessary features of a widened, community-oriented conception of data literacy.

### A central skills framework

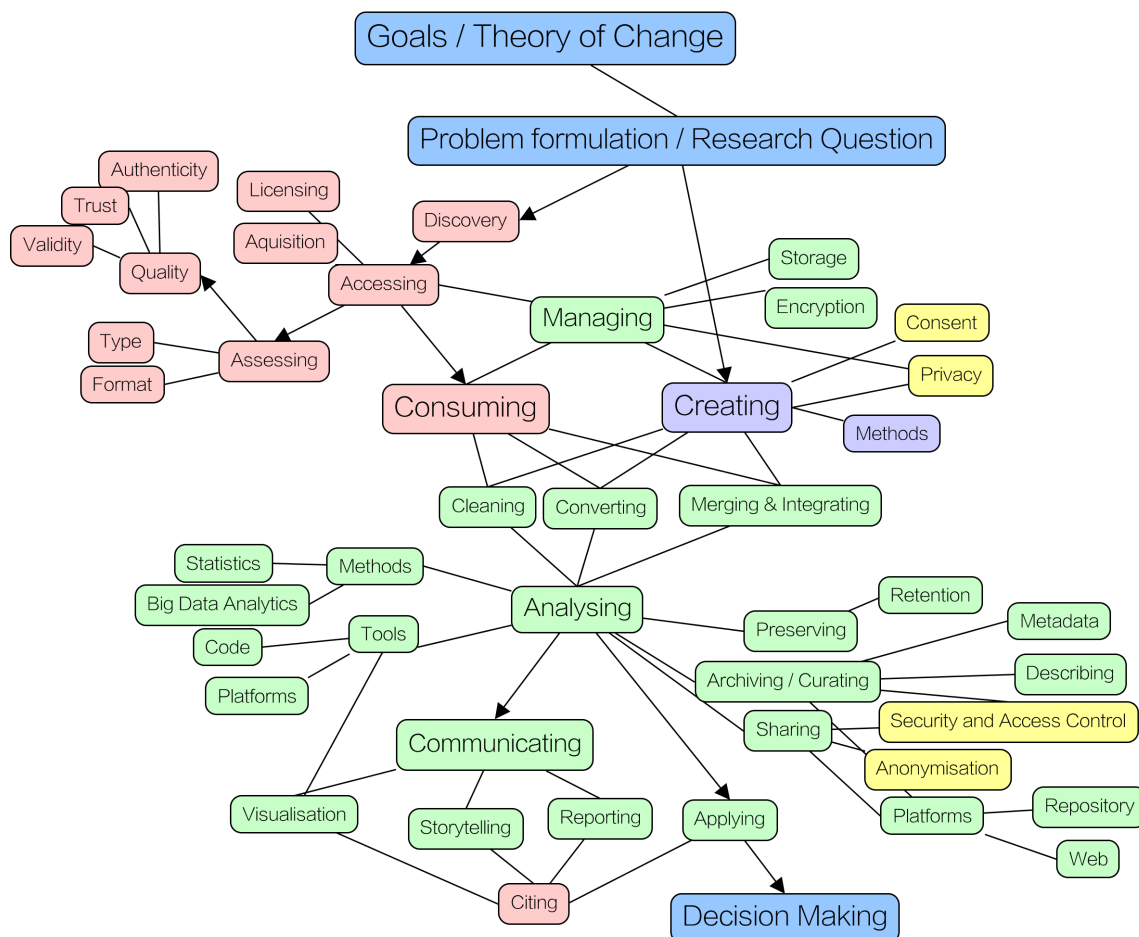
Figure 1, below, provides a generalised conceptual model of the core competencies of data literacy, most of which are cited in common across different sources on the topic.

There is a common consensus that data literacy skills need to encompass both the creation and use of data (Carlson, Fosmire, Miller, & Nelson, 2011) That said, many of the competencies following the acquisition of data from either route are common to both. To begin with, however, one needs to be able to frame a research question or problem that can be addressed with application of data (“Problem Formulation” in

Figure 1). This question itself will stem from a stated goal of some sort (Anderson, Gummer, Mandinach, & Parton, 2015). Thus, the first principle of data literacy is seen by many to be smart goal-setting (“Goals/Theory of Change”), followed by the informed analysis of data relevant to goal monitoring or validation (Tygel, Campos, & de Alvear, 2015).

If data already exists, then skills are required to source and extract it, then to further assess quality and credibility (“Discovery”, “Accessing”, “Assessing”). Creation of new data clearly requires knowledge of the relevant method of production and any relevant technology or methodology (“Creating”, “Methods”). Some of the most widely discussed or core competencies then centre around pre-analysis steps concerned with cleaning and transforming data, as well as combining data from multiple sources (“Cleaning”, “Converting”, “Merging”). This *data munging* is repeatedly noted to be an underestimated skill, yet a considerably extensive part of any data work (Coulton, Goerge, Putnam-hornstein, & de Haan, 2015; Kingsley, Pettit, & Hendey, 2013).

**Figure 1: Generalised conceptual model of data literacy activities.**



Starting point Koltay (Koltay, 2015): Competencies connected to consumption are shown in red, creation shown in purple, common competencies shown in green. Ethical considerations shown in yellow.

Analysis is usually seen to require the *sub-literacy* of statistics to test out hypotheses and explore relationships (Carlson et al., 2011). This stage is often linked to the use of platforms and tools to help with the analysis (“Methods” and “Tools”). While such tools and platforms for data cleaning, analysis and publication are increasingly powerful and easy to use, they can only be considered as part of a solution and advanced knowledge in other aspects of data work will also be needed (Davies, Perini, & Alonso, 2016).

Post-analytical data competencies include the sensible archiving of data for reuse, which should involve considerations about access and security (“Preserving”, “Archiving”), in addition to adequate description (Federer, Lu, & Joubert, 2016).

The final area – communication of outcomes – is one where a great deal of emphasis is placed, particularly in light of the wealth of visualisation and exploration tools now available (“Communicating”). There is also a notable trend toward *storytelling* with data, where narrative is seen as the most effective way of getting across the whole lifecycle from goal to result and the new insights afforded by the analysis (IODC, 2015).

## Four conceptual varieties

Aside from the commonalities noted above, it is important to also delineate the variety of approaches to data literacy taken by different authors. This serves to make clearer the different settings in which the concept is studied and also potential limitations in scope and vision between varieties.

To investigate these varying emphases, 13 sources dealing directly with data literacy or seeking to define it were analysed following the document clustering technique described by Rose (Rose, 2016). Sources were selected using snowball sampling from Koltay (2015) and from the School of Data bibliography<sup>1</sup> and focused on material written in the last five years which had the term “data literacy” in the title.

Text was extracted using pdftotext<sup>2</sup> software and headers, footers and references were removed. A reference set of tokens across all texts was constructed resulting in 69,153 items. A document x term matrix was constructed and tf/idf (a measure of the term’s importance within the document in proportion to the set of all documents) calculated for each term. Then, a K-Means cluster analysis with a target of 4 clusters was executed on this matrix<sup>3</sup>. Four clusters were suggested by a visual inspection of the documents and by iteration (A greater number of clusters led to the “Research” cluster being split into very similar/significantly overlapping clusters). Clustering results are given in Table 1, below, with cluster names chosen by the author and further elaborated below. In Table 2, the top distinctive terms for each cluster are listed in order of occurrence.

---

<sup>1</sup> <http://schoolofdata.org/2016/02/11/research-results-part-6-data-literacy-research-references-and-resources/>

<sup>2</sup> <https://en.m.wikipedia.org/wiki/Pdftotext>

<sup>3</sup> The repository for this analysis is available at <https://github.com/paulusm/datalit-documents>.

**Table 1: Sources used for document clustering**

Source	Main Focus Cluster
Anderson, R., Gummer, E. S., Mandinach, E. B., & Parton, B. M. (2015). Ethical and Appropriate Data Use Requires Data Literacy. <i>Phi Delta Kappan</i> , 96(5), 25–28.	2-Classroom
Bradshaw, L. (2014). Beyond Data Science: Advancing Data Literacy – Made by Many – Medium. Retrieved September 2, 2016, from <a href="https://medium.com/the-many/moving-from-data-science-to-data-literacy-a2f181ba4167#.ayhkrcoop">https://medium.com/the-many/moving-from-data-science-to-data-literacy-a2f181ba4167#.ayhkrcoop</a>	4-Inclusion
Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2011). Determining Data Information Literacy Needs: A Study of Students and Research Faculty. <i>Portal: Libraries and the Academy</i> , 11(2), 629–657.	1-Research
Data-Pop Alliance. (2015). <i>Beyond Data Literacy: Reinventing Community Engagement and Empowerment in the Age of Data</i> .	4-Inclusion
Erwin, R. (2015). Data Literacy: Real-World Learning Through Problem-Solving With Data Sets. <i>American Secondary Education</i> , 43(2), 18–26.	1-Research
Federer, L. M., Lu, Y.-L., & Joubert, D. J. (2016). Data literacy training needs of biomedical researchers. <i>Journal of the Medical Library Association : JMLA</i> , 104(1), 52–7.	1-Research
Fonticiaro, K., & Oehrli, J. A. (2016). Why data literacy matters. <i>ALA Knowledge Quest</i> , 44(5), 20–28.	1-Research
Koltay, T. (2015). Data literacy: in search of a name and identity. <i>Journal of Documentation</i> , 71(2), 401–415.	1-Research
Mandinach, E. B., & Gummer, E. S. (2013). A Systemic View of Implementing Data Literacy in Educator Preparation. <i>Educational Researcher</i> , 42(1), 30–37.	2-Classroom
Martin, E. (2014). What Is Data Literacy? <i>Journal of eScience Librarianship</i> , 3(1), 1–2.	1-Research
Slater, D. (2016). Research Results Part 1: Defining Data Literacy. Retrieved September 2, 2016, from <a href="http://schoolofdata.org/2016/01/08/research-results-part-1-defining-data-literacy/">http://schoolofdata.org/2016/01/08/research-results-part-1-defining-data-literacy/</a>	4-Inclusion
Teal, T. K., Cranston, K. A., Lapp, H., White, E. P., Wilson, G., Ram, K., & Pawlik, A. (2015). Data Carpentry: Workshops to Increase Data Literacy for Researchers. <i>International Journal of Digital Curation</i> , 10, 135–143.	3-Carpentry

### **1) A research focus**

This cluster includes the most common approaches to data literacy, with an emphasis on a further education setting and on librarians as stewards of data and advocates of good data management practice (e.g. Fonticiaro & Oehrli, 2016). Data skills are seen as one part of wider information literacy abilities, or even merged into a wider *data information literacy*:

“Data information literacy ... merges the concepts of researcher-as-producer and researcher-as-consumer of data products. As such it builds upon and reintegrates data, statistical, information and science data literacy into an emerging skill set.”

(Carlson et al., 2011, p. 6)

In the academic research community, data is seen as a key asset, being both a product of a research project or programme and a source for new investigative efforts. A distinctive research view of data literacy focuses also on the processes for storing and sharing data, alongside the application of metadata for discovery and potential reuse (e.g. Federer et al., 2016).

### **2) A classroom focus**

Within secondary education, data literacy is promoted as a means for students to become familiar with data manipulation in problem-based learning (Erwin, 2015). The core learning objectives here are question-asking, evidence gathering and performing relatively simple analysis and visualisation.

“participating students demonstrated a satisfactory grasp of fundamental concepts of the scientific research process, the process of organizing and cleaning the data, the use of descriptive statistics to better understand the data, and the use of spread-sheet software to efficiently and accurately generate statistical analyses.”

(Erwin, 2015, p. 20)

The hope for data literacy within early education is that it will instil enthusiasm for evidence-based decision making and engagement with data connected to real-world problems. The outcomes, then, are seen to be as much to do with an appreciation of societal challenges as a set of data manipulation competencies.

### **3) A “carpentry” focus**

Data carpentry focuses very much on the practical means toward developing data analysis capacity, with hands-on engagement with relevant datasets being the core activity. Perhaps most importantly, the focus is on data relevant to workshop participants’ own domain of work, so learning is of immediate practical relevance.

“we developed Data Carpentry as two-day workshops to meet these data training needs and focus on standard steps in the data workflow – organizing, managing and analyzing data in a more efficient and reproducible way.

Additionally, because people learn best when new skills are building on an existing framework, Data Carpentry workshops are designed to be domain-specific so researchers can learn more quickly and effectively, and see more immediately how to implement these skills and approaches in their own work.”

(Teal et al., 2015, p. 138)

In the data carpentry conception, we see less emphasis given to high level literacy goals, and more given to tools and techniques – to a steep learning curve being surmounted by deliberate practice.

#### 4) An inclusion focus

A broader conception, and arguably one more relevant to community informatics, is one I am calling an *inclusion* conception. Here, a key goal of data literacy is to overcome skill and knowledge inequalities, to enable community members and organisations to access data and to put it to work on local and personal concerns.

“Data literacy focused on building social inclusion offers a doorway to understanding, interpreting, and managing data-driven decisions and arguments for all people. The alternative future we must strive for is one where people are incentivized and empowered to control their own data and its use. This is data inclusion”

(Data-Pop Alliance, 2015, p. 23)

While all of the core competencies remain relevant to the inclusion conception, a distinctive feature is this widening of the concept to better capture the end aspiration of building literacy skills within communities. A further goal is that marginalised people are *directly* involved, rather than simply being the intended beneficiaries of data projects (Davies et al., 2016)

**Table 2: Most common terms in paper clusters corresponding to alternate data literacy conceptions (“data”, “literacy”, and stop words removed)**

Research Focus	Classroom Focus	Carpentry Focus	Inclusion Focus
Librarians Libraries Information Management Faculty Projects Learned Standards Relevant Service Research	Teacher Decisions Course State Parent Prepared Know Students Policy Training Privacy	Workshops Training Instructors Software Lessons Materials Participating Surveys Taught Reproducibility Allow	Social Participating Design Inclusion Stories Narrative Journals Big Contextualize Able Term
7 documents e.g.: (Carlson et al., 2011)	2 documents, e.g. (Mandinach & Gummer, 2013)	1 document: (Teal et al., 2015)	3 documents, e.g. (Data-Pop Alliance, 2015)

## A Capabilities Conception

The inclusion conception developed by Data-Pop pushes data literacy toward the form of a capability. The capabilities approach (CA) developed by Amartya Sen was a response to a prevailing view of human development as mere material gain (to be measured in GDP per capita) and the desire for intrinsic rather than instrumental measurements of the success of development programmes (Robeyns, 2011). Sen proposed instead the concept of development goals as the winning of new freedoms enabled by enhanced *functionings* (Sen, 2005). These Sen left deliberately unspecified, though they were seen to vary from the simple to the sophisticated. An important feature of the approach is that a capability is not necessarily practiced, but that the ability to practice it confers a freedom of opportunity on the individual. Equally important is the idea that capabilities may vary between people even given the same set of *personal means* (access to resources/primary goods), due to other factors such as interpersonal differences, environmental factors and societal positions.

While Sen argued that capabilities did not need to be specified and were necessarily context-specific, his peer Nussbaum proposed that a defined core set of capabilities was necessary for advocacy purposes. Within Nussbaum's proposed core capabilities, we can see where data literacy would fit, notably in her "practical reason" and "political and material control over one's environment" capabilities (Nussbaum, 2003). Nussbaum's *practical reason* can be seen to be enhanced by evidence/data-based decision making in life decisions. Similarly, political and material control can be conferred through participation, including the appreciation and active use of data relating to politics, land tenure and employment.

The CA has been highly influential in the development world (Robeyns, 2011). It has certainly been seen to be relevant to community informatics, though it has been persuasively argued that it is too individualistic in focus and that social and environmental capability dimensions also need to be considered (Stillman & Denison, 2014). In a similar vein, Lanzi notes how capabilities may be divided, for better clarity, into *S-caps*, or concrete skills and knowledge; *E-caps*, relating to external norms and social roles; *O-caps* or option capabilities; and *M-caps*, relating to moral/ethical principles and life choices. While S-caps and E-caps afford basic skills and professional competencies, all of the above capability types are required in concert to afford complex functionalities (which include social change, self-learning, knowledge management, and problem solving) (Lanzi, 2007)

### Fostering and Evaluating Social Capabilities

Given the range in competencies required in data work, it is no surprise that these are often seen as most efficiently distributed amongst a team or organisation's staff, with no one individual likely to be sufficiently knowledgeable (Bradshaw, 2014). Several authors attest to the importance of such multidisciplinary teams to the success of data-intensive projects (Atenas, Havemann, & Priego, 2015; Slater, 2016). While data specialist roles may not be formalised, it may suffice to recognise individuals already using data effectively (Bhargava, 2015).

Similarly, data intermediaries are considered an important part of the ecosystem of community development projects at the present- working alongside community-based organisations or individuals (Davies et al., 2016; IODC, 2015). Intermediaries bring technical expertise and analytic abilities to bear on domain-specific problems. Importantly, for capacity-building to occur effectively, it works best for intermediaries to have a long-term commitment to a community and to earn trust (Kingsley, Coulton, & Pettit, 2014)

Social capabilities within communities can be raised by activities associated with social capital, namely: developing interdependence; fostering frequent interactions and creating space for conversations and interactions (Ansari, Munir, & Gregg, 2012)

Shinn notes the importance of *variability* in outcomes as a measure of the success of community programmes. That is, if capacity building is geared toward one particular development outcome, then it will severely limit the ability to be inspired in one's choice of life goals. This *capacity to be inspired* is raised by “practice, repetition, exploration, conjecture and refutation” (Shinn, 2015). Shinn gives the example of employment programmes for those experiencing mental illness – where a broad set of capabilities is a preferred outcome, in contrast to a limited set of job roles that participants might be ready for on completion.

## Conclusion

Data literacy then, is a set of practical competencies situated in a wider context of personal and social goals, as well as challenges such as subjectivity and bias. Without an ability to contextualise, data analysis efforts are often rudderless or misapplied (Bradshaw, 2014). For this reason, the core data literacy competencies cannot be considered in isolation and the intrinsic benefits of data-informed functioning need to be linked to the technical gains.

Properly directed, all four of the data literacy flavours detailed above can build toward capability. Good *research* data practice in the academy will serve to improve the transparency and reproducibility of research and possibilities for public engagement. Starting early in the *classroom* with simple data work will establish cognitive links between use of evidence and societal issues. *Carpentry* builds an important bridge between domain expertise and personal data skills. And *inclusion* rightly seeks to make contextualised data skills a universal aspiration.

A simple common thread between all approaches is the building of familiarity through practice. Practice engenders experience and comfort around data (Slater, 2016) and a deep knowledge of the context around particular datasets (Kingsley et al., 2014). This in turn enables the flexible and inspired application of data to personal and community concerns. In this way, a new freedom is won through the development of a technical literacy.

The challenge of truly inclusive data literacy should of course not be understated (Data-Pop Alliance, 2015) though locally-based, locally-committed technical data intermediaries can provide a part of the infrastructure (Kingsley et al., 2013). The local data ecosystem may include established institutions in new data roles such as public libraries, which are well placed to act as local data hubs (Bertot, Butler, &



Travis, 2014). Importantly, community capabilities will likely be raised as much through socialisation as through technical training.

## Acknowledgements

This article was inspired by a discussion session on data literacy at Open Data Camp 3 in Bristol in May 2016 convened by Johanna Walter and Nat Foo (<http://odcamp.org.uk/>)

## References

- Anderson, R., Gummer, E. S., Mandinach, E. B., & Parton, B. M. (2015). Ethical and Appropriate Data Use Requires Data Literacy. *Phi Delta Kappan*, 96(5), 25–28. <http://doi.org/10.1177/0031721715569465>
- Ansari, S., Munir, K., & Gregg, T. (2012). Impact at the “Bottom of the Pyramid”: The role of social capital in capability development and community empowerment. *Journal of Management Studies*, 49(4), 813–842. <http://doi.org/10.1111/j.1467-6486.2012.01042.x>
- Atenas, J., Havemann, L., & Priego, E. (2015). Open data as open educational resources : Towards transversal skills and global citizenship. *Open Praxis*, 7(4), 377–389. <http://doi.org/http://dx.doi.org/10.5944/openpraxis.7.4.233>
- Bertot, J. C., Butler, B. S., & Travis, D. M. (2014). Local big data. *Proceedings of the 15th Annual International Conference on Digital Government Research - Dg.o '14*, 17–23. <http://doi.org/10.1145/2612733.2612762>
- Bhargava, R. (2015). Architecting for Data. Retrieved September 2, 2016, from [http://ssir.org/articles/entry/architecting\\_for\\_data](http://ssir.org/articles/entry/architecting_for_data)
- Bradshaw, L. (2014). Beyond Data Science: Advancing Data Literacy – Made by Many – Medium. Retrieved September 2, 2016, from <https://medium.com/the-many/moving-from-data-science-to-data-literacy-a2f181ba4167#.ayhkrcoop>
- 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 and the Academy*, 11(2), 629–657. <http://doi.org/10.1353/pla.2011.0022>
- Coulton, C. J., Goerge, R., Putnam-hornstein, E., & de Haan, B. (2015). *Harnessing Big Data for Social Good : A Grand Challenge for Social Work*.
- Data-Pop Alliance. (2015). *Beyond Data Literacy: Reinventing Community Engagement and Empowerment in the Age of Data*.
- Davies, T., Perini, F., & Alonso, J. (2016). Researching the emerging impacts of open data ODDC conceptual framework. *The Journal of Community Informatics*, 12, 1–37.
- Erwin, R. (2015). Data Literacy: Real-World Learning Through Problem-Solving With Data Sets. *American Secondary Education*, 43(2), 18–26. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=103298736&site=ehost-live>
- Federer, L. M., Lu, Y.-L., & Joubert, D. J. (2016). Data literacy training needs of biomedical researchers. *Journal of the Medical Library Association : JMLA*,

- 104(1), 52–7. <http://doi.org/10.3163/1536-5050.104.1.008>
- Fonticiaro, K., & Oehrli, J. A. (2016). Why data literacy matters. *ALA Knowledge Quest*, 44(5), 20–28.
- IODC. (2015). *Enabling the Data Revolution*. Retrieved from <http://1a9vrva76sx19qtvvg1ddvt6f.wpengine.netdna-cdn.com/wp-content/uploads/2015/11/opendatacon-report-en-web.pdf>
- Kingsley, G. T., Coulton, C. J., & Pettit, K. L. S. (2014). *Strengthening Communities with Neighborhood Data*. <http://doi.org/10.1007/s13398-014-0173-7.2>
- Kingsley, G. T., Pettit, K. L., & Hendey, L. (2013). *Strengthening Local Capacity For Data-Driven Decisionmaking*. Retrieved from <http://www.urban.org/UploadedPDF/412883-Strengthening-Local-Capacity-For-Data-Driven-Decisionmaking.pdf>
- Koltay, T. (2015). Data literacy: in search of a name and identity. *Journal of Documentation*, 71(2), 401–415. <http://doi.org/10.1108/JD-02-2014-0026>
- Lanzi, D. (2007). Capabilities, human capital and education. *Journal of Socio-Economics*, 36(3), 424–435. <http://doi.org/10.1016/j.socec.2006.12.005>
- Mandinach, E. B., & Gummer, E. S. (2013). A Systemic View of Implementing Data Literacy in Educator Preparation. *Educational Researcher*, 42(1), 30–37. <http://doi.org/10.3102/0013189X12459803>
- Nussbaum, M. (2003). Capabilities As Fundamental Entitlements: Sen and Social Justice. *Feminist Economics*, 9(2–3), 33–59. <http://doi.org/10.1080/1354570022000077926>
- Prado, J., & Marzal, M. (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri*, 63(2), 123–134. <http://doi.org/10.1515/libri-2013-0010>
- Robeyns, I. (2011). The Capability Approach. Retrieved September 16, 2016, from <http://plato.stanford.edu/entries/capability-approach/>
- Rose, B. (2016). Document Clustering with Python. Retrieved September 12, 2016, from <http://brandonrose.org/clustering>
- Sen, A. (2005). Human Rights and Capabilities. *Journal of Human Development*, 6(2), 151–166. <http://doi.org/10.1080/14649880500120491>
- Shinn, M. (2015). Community Psychology and the Capabilities Approach. *American Journal of Community Psychology*, 55(3–4), 243–252. <http://doi.org/10.1007/s10464-015-9713-3>
- Slater, D. (2016). Research Results Part 1: Defining Data Literacy. Retrieved September 2, 2016, from <http://schoolofdata.org/2016/01/08/research-results-part-1-defining-data-literacy/>
- Stillman, L., & Denison, T. (2014). The Capability Approach Community Informatics. *The Information Society*, 30(3), 200–211. <http://doi.org/10.1080/01972243.2014.896687>
- Teal, T. K., Cranston, K. A., Lapp, H., White, E. P., Wilson, G., Ram, K., & Pawlik, A. (2015). Data Carpentry: Workshops to Increase Data Literacy for Researchers. *International Journal of Digital Curation*, 10, 135–143. <http://doi.org/10.2218/ijdc.v10i1.351>
- Tygel, A. F., Campos, M. L. M., & de Alvear, C. A. S. (2015). Teaching Open Data for Social Movements: a Research Strategy. *The Journal of Community Informatics*, 11(3).