Library Collections as Humanities Data: The Facet Effect

Padilla, Thomas G., and Devin Higgins. 2014. “Library Collections as Humanities Data: The Facet Effect.” *Public Services Quarterly* 10 (4): 324–35. doi:10.1080/15228959.2014.963780

Many library collections contain digital text, images, and audio. Materials in these forms and the metadata that describe them are frequently the objects of inquiry that Digital Humanists, inside and outside the library, subject to computational analysis to extend their research and pedagogy. Librarians can further enhance use of their digital collections by considering how thinking of them as Humanities data, and promoting them as such, can encourage uses beyond reading, viewing, and listening. For an indicator of what this thinking looks like in practice it is instructive to consider the Library of Congress' effort to make digitized newspaper data openly available through an Application Programming Interface (API), allowing algorithmic interaction in addition to reading through an interface that stands as a surrogate for an analog reading experience (Johnston, 2011, 2014). Michigan State University Libraries has also made modest steps in this direction by making select digitized collections available as bulk downloads (Michigan State University Libraries, 2014). Both efforts are ground in an understanding that data afford new opportunities for user interaction with library collections.

In what follows the authors work through a high level discussion of relevant literature on concepts of information and data to arrive at a definition of Humanities data. Given the scope of this paper and its role in encouraging an understanding of digital collections as Humanities data, this discussion will necessarily be limited to data that is processable by a computer. For an introduction to broader conceptions of information and data consider *Fundamental Forms of Information* (Bates, 2006) and *Semantic Forms of Information* (Floridi, 2013). Following a definition of Humanities data, we proceed to highlight the value that a Humanities data framing offers by working through a series of digital content types and highlighting facets of those data potentially useful for digital humanities research.

DEFINING DATA

Demonstrating the value of a Humanities data framing for digital collections begins with a brief discussion of relevant literature on data. Readers seeking an introduction are well served by Luciano Floridi's *Information: A Very Short Introduction* (2010). The *General Definition of Information* (GDI) Floridi introduces is particularly valuable for those working with digital collections in cultural heritage organizations by virtue of its adoption in Information Science and Database Design communities (2013). The GDI holds that information is composed of data. Intelligibility of information is predicated on *well-formedness*. Data are considered well-formed when arranged according to the syntax, or rules, of a given system. A timeline visualization is an example of data arranged by a temporal syntax. The final component of the GDI holds that well-formed data must be meaningful. Bearing meaning is predicated on conformance to the semantics of a given system. Extending the timeline example, we might say that data represented by a timeline visualization bears the meaning of linearity. Finally, the GDI holds that if the component parts of the information are in place, the meaningfulness of data is present regardless of whether or not an observer can access the meaning of the data. For example, a spreadsheet of data with column headers in an unknown language, with string and numeric data types arrayed below, would indicate the presence of information regardless of whether or not an individual could interpret that information.

Having discussed the GDI, we remain without an actual definition of data. The history of usage in the English language offers initial clues. In *Data Before the Fact,* we learn that the word data is the plural form of the Latin *datum,* which is the neuter past participle of the verb *dare* or to give (Rosenberg, 2013, p. 18). For Rosenberg, this association imparts a notion of data as “something given.” This etymology frames data as the “given” thing that is required to assert facts, with facts being used to elaborate evidence. Rosenberg formally distinguishes facts as ontological, evidence as epistemological, and data as rhetorical (p. 18).

Rosenberg illustrates how data have functioned over time by analyzing word usage across a large corpus of texts drawn from Eighteenth Century Collections Online (ECCO). Rosenberg's analysis reveals a shift from data utilized in argument as agreed upon principles and/or facts derived from religious scripture, to a more contemporary conception of data as materials that are "generated by experiment, experience, or collection" (p. 33). Based on Rosenberg's research, it is possible to define data in part as materials generated by investigation, experience, or collection that serve a rhetorical function in support of argumentation.
Part of the difficulty in defining data lies in parsing definitions. Gregory Bateson described information as "a difference that makes a difference"; Trevor Munoz asserts that "data is a common rhetorical tool for tracing disciplinary practices"; Lisa Gitelman argues in part that data are the product of the norms and methodologies that govern disciplinary imagination (Gitelman, 2013, p. 3); and Luciano Floridi defines data as a lack of uniformity (Bateson, 1972, p. 453; Munoz, 2014; Floridi, 2010, p. 23). Combining the first and last definition it is possible to say that data are the record of difference. By combining all perspectives it is possible to say that data are records of difference arranged, interpreted, and put into the service of argumentation according to disciplinary norms and methodologies.

Grammar offers a final vector for shedding light on our effort to define data. Lisa Gitelman notes that data is a mass noun which means that it can take the singular verb form (2013). From a grammatical perspective then, "data is" and "data are" constitute valid use (Rogers, 2010), though Gitelman noted that a Google search for "data is" produced nearly four times as many results as "data are." On the face of it this seems like a fairly innocuous battle of the grammarians, but it highlights an often overlooked feature of data. It becomes remarkably easy to consider data as a unitary totality, thus "data is" rather than "data are," eliding the complexity of a structure that provides internal cohesion and allows linkages to other data. In a nod to Matthew Kirschenbaum's (2008) writings on "screen essentialism," to realize that "data are" is to take a step closer to dispelling a "data essentialism" that conceals the complex and multifaceted nature of data.

DEFINING HUMANITIES DATA

… Humanities data are organized difference presented in a form amenable to computation put into the service of Humanistic inquiry.

Humanists in a broad array of disciplines have been thinking for some time about how computation can be leveraged to study their objects of inquiry. Over the past decade, this effort has coalesced under the broad umbrella of the Digital Humanities. Digital Humanities can be usefully distilled as scholarship presented in digital form(s), scholarship enabled by digital methods and tools, scholarship about digital technology and culture, building and experimenting with digital technology, and a self-reflexive engagement with Digital Humanities research and pedagogy (Honn, n.d.). Where aspects of this thinking may have been relegated to corners of the academy in the past, a contemporary shift toward knowledge products that are encoded, stored, and disseminated digitally forces broader relevance. It is clear that the computational aspects of this research render digital collections as data. It stands to reason that librarians must understand what facets of data are likely to be leveraged in the course of Digital Humanities research.

The authors hold that Humanities data are organized difference presented in a form amenable to computation put into the service of Humanistic inquiry. Production, organization, and interpretation of Humanities data is shaped by disciplinary norms and methodologies. Forms of Humanities data include but are not limited to text, audio, images, and moving images. These data are encoded in digital form and are thus processable by a computer. Processability enables measurement, identification, and extraction of data at the micro level across a macroscopic scale, which in turn supports the ability of Digital Humanists to apply a wide range of visualization and data mining techniques to Humanities data. An example of this application can be seen in Ian Milligan's computational exploration of millions of web pages crawled by the Internet Archive. Part of this work entails identification, extraction, visualization, and analysis of image data such as color values, the tendency for groupings of color to occur across certain types of websites, and image file format adoption rates (Milligan, "Using images," 2014). Data generated in the course of this work can stand alone or be associated with source data through placement into an appropriate metadata schema. Accordingly metadata constitutes a set of data that Digital Humanists can use to derive insight.

Consider Schmidt's (2012) exploration of author gender distribution across LC subject classes, Mullen's (2014) analysis of History dissertation size over time, and Underwood, Black, Auvil, and Capitanu's (2013) use of MARC records to help determine genre across large volumes of heterogeneous text content. In the cultural heritage sector, institutions such as the Cooper Hewitt Museum have used pixel value data from their images, combined with code that operationalizes the concept of Shannon Entropy to sort image collections by visual complexity (Walter, 2013; Weisstein, n.d.).

Cultural heritage organizations large and small are rich repositories of Humanities data. Hallmark projects such as American Memory at the Library of Congress began exploring mass digitization of text, moving image, and audio as early as 1990, later implementing a program that would make more than 5 million digital items available to the public by 2000. Since 1996, The Internet archive has captured, preserved, and provided access to more than 423 billion web pages, as well as large collections of software, text, audio, and moving images. The HathiTrust, which originated as a partnership between universities belonging to the Committee on Institutional Cooperation and the University of California System, now counts 90 members that contribute effort to preserving and providing access to millions of text works (HathiTrust, n.d.). Rhizome, a nonprofit located in the New Museum, focuses considerable effort on preserving digital art (Fino-Radin, 2011). Despite presenting rich repositories of Humanities data, cultural heritage institutions have long limited themselves to displaying their Humanities data (books, images, and so forth), to the neglect of promoting the ability to derive insights from them at scale (Leonard, 2014). In what follows we aim to contribute to closing this gap by discussing facets of text and image data that can be used in Digital Humanities research.

HUMANITIES DATA | TEXT

… digital text abundance combined with widespread availability of computational methods and tools allow extension and automation of sense-based faculties.

It is not outlandish to claim that digital text are the most ubiquitous Humanities data that cultural heritage institutions hold. Mass digitization efforts along with contemporary practices of knowledge production bias collection composition heavily toward this content type.

Models attempting to define what digital text are vary from a sequence of graphic characters to procedural coding and graphic characters, geometric shapes, an image, and an Ordered Hierarchy of Content Objects (OHCO) (Renear, 2004). For our intents and purposes, providing a model for text is less important than understanding what facets digital text offers that can be leveraged for research purposes. For analysis of single passages, Ted Underwood has written that our "wrinkled protein sponge" is sufficient to the task (2012). When scholars seek to push past the single passage to consider texts at scale, they quickly reach the limits of sense-based ability to capture and retain facets of text data like line length, part of speech, and relationships between words (Higgins, 2014). It is at this threshold that scholars like Franco Moretti (2007), with his computational exploration genre, have realized the value of digital text.
Digital text abundance combined with widespread availability of computational methods and tools allow extension and automation of sense-based faculties to enable interaction with and derivation of insight from Humanities text data at scale. Broadly, Ted Underwood (2012) has written that computability of text allows scholars to categorize text by features such as word frequencies, which in turn support assertions about genre and similarity between texts; contrast vocabulary between texts in different genres to explore distinctiveness or similarity; trace word and phrase usage over time; group documents by features that tend to occur together like sequences of words; and apply named entity extraction to identify things such as personal names and geographic locations. Trevor Owens (2014) has characterized utilization of these methods and tools like these as less borg and more exo-suit, where exo-suit "illustrates a vision of technology that extends the capabilities of its user."

Prime examples of exo-suits in action include Robert K. Nelson's *Mining the Dispatch* (2014), Andrew Goldstone and Ted Underwood's *The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us* (2014), Micki Kaufman's *Everything on Paper will be Used Against Me:*

Quantifying Kissinger (2014), and Mark Sample's *Hacking the Accident* (2012). Nelson used a tool called MALLET to apply a statistical method called topic modeling to a corpus of digitized Civil War era dispatches comprising 112,000 pieces and 2.4 million words. Topic modeling allowed Nelson to categorize the corpus by topic, or patterns of words that tend to occur together. With these topics in place scholars are able to navigate the corpus by variation in topic expression over time. Goldstone and Underwood applied a similar method to seven literary studies journals, published 1889–2013, to uncover thematic and rhetorical trends at the article level. With this data they were able to argue for the presence of distinct changes in disciplinary focus and practice over time. Kaufman's work with the Digital National Security Archive's Kissinger Collection further illustrates the multifaceted nature of Humanities data. Kaufman scraped documents and associated metadata from the DNSA website. Kaufman proceeded to extract text data from the PDFs. The extracted text data was put into AntConc to generate word frequency and collocation data, had topic modeling applied to it using MALLET, and was run through a sentiment analysis program. Data generated through each of these tools were in turn used to support visualizations that served to enhance exploration of topics, themes, and relationships surfaced from DNSA Kissinger Collection. Mark Sample's (2012) *Hacking the Accident* takes a slightly different approach to working with Humanities data. Rather than focusing analysis on measurements of source data, Sample deforms the data through application of an N + 7 algorithm that "replaces every noun—every person, place, or thing— … with the person, place, or thing—mostly things—that comes seven nouns later in the dictionary" (2012). This act of deformance produces a new object of inquiry.

There is much debate as to what skills and competencies are needed to engage in this type of research. The authors argue for a spectrum of requisite skill and competency. Frederick Gibbs and Trevor Owens have both written on the value of low barrier data exploration. For example, a Historian might fruitfully examine Google N-Grams to discern interesting paths and probable dead ends, without need for computer programming or statistical skills (Gibbs & Owens, 2013; Owens, 2012). Similarly, it has been argued that use of a tool that obscures the inner workings of an algorithm beneath a graphical user interface is suitable, where ease of initial use serves to incentivize a user to commit requisite time and energy to acquire needed competencies in computer programming and statistics (Padilla, 2014). Generally speaking, the need for higher levels of algorithmic understanding, statistical reasoning, and computer programming competency run proportional to desire to put data into the service of a research claim.

HUMANITIES DATA | IMAGE

… the data-driven approach allows us to "see" the image from a new, often estranging angle.

Crispin Sartwell (2011) describes images as "breathtakingly transparent," pointing out that "what we don't see is that we understand images *too well*" (p. 162). Photographic images in particular "hit" us, in an almost physical way, seemingly without mediation, as direct expressions of reality. Sartwell goes on to describe Flint Schier's position that the distinguishing feature between "word" and "text" is "ease of interpretation" (p. 162). Words are encoded in a particular language that needs to be decoded, whereas images happen to us directly, and seductively. Schier's (1986) phrase for the way in which images work on us automatically, even in cases where the image is symbolic and requires an act of interpretation to decipher, is "natural generativity" (p. 43). Images and their associated meanings seem to spring up almost spontaneously, perhaps, as Schier suggests, to an even greater extent than the meanings of words do.
In light of this passive processing of images we are subject to as human viewers, there has been a countervailing push to intellectually engage with the image as the site of "a discrepancy, a dissemblance" (Rancière, 2006, p. 7). Contrary to any notion of transparency, the image " … presents a deceptive appearance of naturalness and transparence concealing an opaque, distorting, arbitrary mechanism of representation, a process of ideological mystification" (Mitchell, 1984, p. 504). The image, even the photograph, does not in itself *tell the truth* and therefore needs to be investigated with greater rigor for all its indications to the contrary.

By the same token, it is still not uncommon to come across media accounts decrying our "image-saturated," "media-driven" society, by which we are meant to understand that encounters with images act as surrogates

for some missing and more fully real set of experiences, in which no veil of images would lie between us and the world. The image, under this view, distances and alienates us from reality. As Guy Debord (1994) said almost 50 years go, "… everything that was directly lived has moved away into a representation" (p. 1). The image has often been conceived of as something like a sliver or surface, a "caul" or "membrane" in the Lucretian view (Elkins & Naef, 2011, p. 3); a way of thinking which would seem to render the term "opaque" irrelevant. Where might we find a hidden depth, if the image is already defined by a lack thereof? What else is there to an image besides what we already see?

Data-driven approaches provide librarians and digital humanists (among others) access to a new means of approaching the *substance* of the image. This substance is not a new layer of depth, per se, but rather a different facet with its own configuration. A typical digital image (excepting, of course, SVG and other vector graphics) is a matrix of pixels, where each pixel possesses a property of color or grayscale which when taken together within a frame create forms and entities which have meaning to a human viewer. Approaching this image from a technical, materialist angle, as a matrix of pixels, rather than as human viewers, brings new facets of this meaning to light and, potentially, undermines others. The image is no longer just a surface for us to witness, but an alien sort of object for us to examine and analyze, under circumstances through which, in fact, our strictly visual interaction with the image may be occluded. The image has a potentially alternate reality as data. Treating the digital image as an instance of data allows evasion of a unitary representation. The data-driven approach allows us to "see" the image from a new, often estranging angle.

Considering images as grids of color data has enabled new modes of film visualization and analysis. Projects such as *The Colors of Motion* and *Movies in Color* reveal the color palettes of films by analyzing individual frames (Clark, 2014; Radulescu, n.d.). Precisely specifying the color palette of a film allows critics to more concretely discuss the aesthetic effects it achieves. Lev Manovich (2013) has used sophisticated visualization techniques in service of critical analysis of the films of Dziga Vertov. Manovich examined individual frames of the films under investigation, as well as shot sequences over time, a further affordance of diachronic content types such as moving images, which expand the repertoire of techniques by which digital media can be processed, analyzed, and interpreted.

Digital images can also be manipulated *en masse*. As viewers of images, we are limited by our faculties of perception and memory: We can neither memorize nor process all of the details of even one image, let alone thousands or millions of them. Computational approaches allow for the systematic analysis of images on the Big Data scale: hundreds of thousands of images from Instagram, for instance. As part of the *Selfiecity* project, a team of researchers examined over 600,000 photos to produce a dataset of 640 self-portraits posted on the social media platform Instagram from each of five world cities: Bangkok, Berlin, Moscow, New York, and Sao Paolo. Human viewers determined which photos were self-portraits, after which algorithms were used to perform "automatic face analysis, supplying … algorithmic estimations of eye, nose and mouth positions, [and] the degrees of different emotional expressions" (*Selfiecity*, 2014). Data about the emotional register of individuals appearing within a set of images constitute a form of metadata that blurs the lines between the technical and the descriptive. Technical metadata have typically consisted of metadata used to ensure "interoperability of systems" and as a form of documentation of "the creation or storage encoding processes or formats" of a particular resource, whereas descriptive metadata have been employed in service of discoverability, including information such as title, creator, and subject heading (National Information Standards Organization, 2004, pp. 12–16). The *Selfiecity* project is an exemplar of a new attitude to resource description and access in which the boundaries of these metadata categorizations are pushed to the breaking point, where metadata can be simultaneously technically derived and descriptive. The suite of features of an image, its technical specifications, do not only exist on a technical, data curatorial plane, but also point toward descriptive facets of the content of the image. Approaching an image as a specific form of data allows us to see the continuation between, and ultimate congruence of, the image as *substance* and the image as *picture*.

The distance between the machine and human understanding of digital imaging is shortening. The Library of Congress presents over 20,000 photographs on Twitter and invites users to tag them or leave comments ("Library of Congress Photos," 2012). Given the variety of photos available, training an algorithm to specify the content of these images is still out of reach. Yet one can imagine future algorithms with the

capability, verging on artificial intelligence, to describe and classify images of arbitrary content. Google has begun the process of doing just that, using an unsupervised machine learning algorithm, essentially a brain-like computer program that looks for patterns without human guidance, to distill images of particular forms or objects that a computer could recognize. Google's project used 16,000 computer processors to pore over 10 million video screenshots, looking to stabilize visual definitions of particular repeated forms, and succeeded, most famously, by figuring out how to recognize a cat (Markoff, 2012; Le et al., 2011). The ability to computationally process images is exciting new terrain for librarians. Algorithmically derived technical and descriptive metadata have the potential to create new paths to library collections and items, in ways that not only serve traditional searching and browsing but also rise to meet the research interests of Digital Humanists looking for sets of images that correspond to criteria not only based on content, but on particular visual features related to color and form.

CONCLUSION

Research conducted under the broad umbrella of the Digital Humanities illustrates the many facets of Humanities data that serve purposes aside from reading and viewing. With greater awareness of the facets that are leveraged in the course of mining, visualizing, and generating new objects of inquiry, comes the ability for librarians to better promote the value of their collections. Encouraging use of these collections is bolstered by development of APIs to interact with data and provision of bulk data downloads. Finally, librarians are in a natural position to promote collections by offering training in the skills, tools, and methods needed to take advantage of Humanities data. Undoubtedly, thinking about, promoting, and providing access to collections in this manner represents a significant challenge, yet the challenge is well worth it for the opportunity it affords to articulate the broader relevance of library collections.

Notes

Comments and suggestions should be sent to the Column Editor: Christopher Guder, Ohio University, Alden Library, 30 Park Place, Athens, OH 45701. E-mail: guder@ohio.edu

REFERENCES

- **1.** Bates , M. (2006). Fundamental forms of information . *Journal of the American Society for Information Science and Technology* , 57 (8), 1033 – 1045 . [CrossRef], [Web of Science ®]
- **2.** Bateson , G. (1972). *Steps to an ecology of mind* . New York , NY : Ballantine .
- **3.** Clark , C. (2014). *The colors of motion* . Retrieved from http://thecolorsofmotion.com
- **4.** Debord , G. (1994). *The society of the spectacle* . New York , NY : Zone Books .
- **5.** Elkins , J. , & Naef , M. (Eds.). (2011). *What is an image?* University Park : Pennsylvania State University Press .
- **6.** Fino-Radin , B. (2011). Digital preservation practices and the rhizome artbase. Retrieved from http://media.rhizome.org/artbase/documents/Digital-Preservation-Practices-and-the-Rhizome-ArtBase.pdf
- **7.** Floridi , L. (2010). *Information: A very short introduction* . New York , NY : Oxford University Press . [CrossRef]
- **8.** Floridi , L. (2013). *The philosophy of information* . Oxford , England : Oxford University Press .
- **9.** Gibbs , F. , & Owens , T. (2013). The hermeneutics of data and historical writing . In K. Nawrotzki & J. Dougherty (Eds.), *Writing history in the digital age* . Ann Arbor : University of Michigan Press . Retrieved from http://writinghistory.trincoll.edu/data/gibbs-owens-2012-spring/
- **10.** Gitelman , L. (2013). *"Raw data" is an oxymoron* . Cambridge , MA : The MIT Press . Retrieved from http://ieeexplore.ieee.org/xpl/bkabstractplus.jsp?bkn=6451327

- **11.** Goldstone , A. (n.d.). *A topic model of literary studies journals* . Retrieved from http://rci.rutgers.edu/~ag978/quiet/#
- **12.** Goldstone , A. , & Underwood , T. (2014). *The quiet transformations of literary studies: What thirteen thousand scholars could tell us* . Retrieved from https://www.ideals.illinois.edu/handle/2142/49323
- **13.** HathiTrust Digital Library . (n.d.). Retrieved from http://www.hathitrust.org/partnership
- **14.** Higgins , D. (2014). Reading and non-reading: Text mining in critical practice . In *The top technologies every librarian needs to know: A LITA guide* . American Library Association Editions . Chicago, IL: ALA TechSource.
- **15.** Honn , J. (n.d.). *A guide to digital humanities | Northwestern University* . Retrieved from http://sites.library.northwestern.edu/dh/
- **16.** Internet Archive: Digital Library of Free Books, Movies, Music & Wayback Machine . (n.d.). Retrieved from https://archive.org/index.php
- **17.** Johnston, L. (2011, November 4). From records to data: It's not just about collections any more. *The Signal: Digital preservation*. Retrieved from http://blogs.loc.gov/digitalpreservation/2011/11/from-records-to-data-its-not-just-about-collections-any-more/
- **18.** Johnston , L. (2012). *Digital collections as Big Data*. Presented at Digital Preservation 2012, Library of Congress. Retrieved from http://www.digitalpreservation.gov/meetings/documents/ndiipp12/BigData_Johnston_DP12.pdf
- **19.** Kaufman , M. (2014). Everything on paper will be used against me: Quantifying Kissinger. Retrieved from http://blog.quantifyingkissinger.com/
- **20.** Kirschenbaum , M. (2008). *New media and the forensic imagination* . Cambridge , MA : MIT Press .
- **21.** Le , Q. V. , Ranzato , M. , Monga , R. , Devin , M. , Chen , K. , Corrado , G. S. , … Ng , A. Y. (2011). Building high-level features using large scale unsupervised learning. *arXiv:1112.6209 [cs]*. Retrieved from http://arxiv.org/abs/1112.6209
- **22.** Leonard , P. (2014). *Mining large datasets for the humanities*. Presented at the IFLA WLIC 2014, Lyon, France. Retrieved from http://library.ifla.org/930/
- **23.** Library of Congress American Memory . (n.d.). Retrieved from http://memory.loc.gov/ammem/dli2/html/lcndlp.html
- **24.** *Library of Congress Photos on Flickr* . (2012). *Library of Congress* . Retrieved from http://www.loc.gov/rr/print/flickr_pilot.html
- **25.** MALLET . (n.d.). Retrieved from http://mallet.cs.umass.edu/
- **26.** Manovich , L. (2013). *Visualizing Vertov* . Retrieved from http://lab.softwarestudies.com/2013/01/visualizing-vertov-new-article-by-lev.html
- **27.** Markoff , J. (2012). In a big network of computers, evidence of machine learning . *The New York Times* . Retrieved from http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html
- **28.** Michigan State University Libraries. (n.d.). Humanities data. Retrieved from https://www.lib.msu.edu/dh/humdata/
- **29.** Milligan , I. (2014). *Colour analysis of web archives* . Retrieved from http://ianmilligan.ca/2014/08/22/colour-analysis-of-web-archives/
- **30.** Milligan , I. (2014). *Using images to gain insight into web archives?* Retrieved from http://ianmilligan.ca/2014/08/11/using-images-to-gain-insight-into-web-archives/
- **31.** Milligan , I. (n.d.). *SSHRC proposal* . Retrieved from http://ianmilligan.ca/the-next-project/sshrc-proposal/
- **32.** Mitchell , W. J. T. (1984). What is an image? *New Literary History* , 15 (3), 503 – 537 . doi: 10.2307/468718 [CrossRef], [Web of Science ®]
- **33.** Mullen , L. (2014). *Analyzing historical history dissertations* . Retrieved from http://lincolnmullen.com/research/history-dissertations/
- **34.** Munoz , T. (2014 , July). *An introduction to humanities data and data curation*. Presented at the 2014 CLIR/DLF Postdoctoral Fellowship Summer Seminar, Bryn Mawr, PA. Retrieved from https://speakerdeck.com/trevormunoz/humanities-data

- **35.** Moretti , F. (2007). >*Graphs, maps, trees: Abstract models for literary history* . New York , NY : Verso .
- **36.** National Information Standards Organization . (2004). Understanding metadata. Retrieved from http://www.niso.org/publications/press/UnderstandingMetadata.pdf
- **37.** Nelson, R. K. (n.d.). Mining the dispatch. Retrieved from http://dsl.richmond.edu/dispatch/
- **38.** Owens , T. (2012). *Discovery and justification are different: Notes on science-ing the humanities* . Retrieved from http://www.trevorowens.org/2012/11/discovery-and-justification-are-different-notes-on-sciencing-the-humanities/
- **39.** Owens , T. (2014). *Mecha-archivists: Envisioning the role of software in the future of archives* . Retrieved from http://www.trevorowens.org/2014/05/mecha-archivists-envisioning-the-role-of-software-in-the-future-of-archives/
- **40.** Padilla , T. (2014). *Tooling: Paths toward sparking interest and curiosity in DH* . Retrieved from http://www.thomaspadilla.org/2014/04/27/tooling/
- **41.** Radulescu , R. (n.d.). *Movies in color* . Retrieved from http://moviesincolor.com/
- **42.** Rancière , J. (2006). *The politics of aesthetics: The distribution of the sensible* . London , England : Continuum .
- **43.** Renear , A. (2004). Text encoding . In S. Shreibman , R. Siemens , & J. Unsworth (Eds.), *A companion to digital humanities* (pp. 218–239). Malden, MA: Wiley-Blackwell . [CrossRef]
- **44.** Rogers , S. (2012, July 8). Data are or data is? *The Guardian*. Retrieved from http://www.theguardian.com/news/datablog/2010/jul/16/data-plural-singular
- **45.** Rosenberg , D. (2013). Data before the fact . In L. Gitelman (Ed.), *Raw data is an oxymoron* (pp. 15 – 40). Cambridge , MA : MIT Press .
- **46.** Sample , M. (2012). *Notes towards a deformed humanities* . Retrieved from http://www.samplereality.com/2012/05/02/notes-towards-a-deformed-humanities/
- **47.** Sartwell , C. (2011). Assessments . In J. Elkins & M. Naef (Eds.), *What is an image?* (pp. 162 – 165). University Park : Pennsylvania State University Press .
- **48.** Schier , F. (1986). *Deeper into pictures: An essay on pictorial representation* . Cambridge , England : Cambridge University Press . [CrossRef]
- **49.** Schmidt, B. (2012, May 8). Sapping attention: Women in the libraries. Retrieved from http://sappingattention.blogspot.com/2012/05/women-in-libraries.html
- **50.** Selfiecity . (2014). *Selfiecity* . Retrieved from http://selfiecity.net/
- **51.** Underwood , T. (2012). *Where to start with text mining* . Retrieved from http://tedunderwood.com/2012/08/14/where-to-start-with-text-mining/
- **52.** Underwood , T. , Black , M. L. , Auvil , L. , & Capitanu , B. (2013). Mapping mutable genres in structurally complex volumes. *arXiv:1309.3323 [cs]*. Retrieved from http://arxiv.org/abs/1309.3323
- **53.** Walter , M. (2013). *Default sort, or what would Shannon do?* Retrieved from http://labs.cooperhewitt.org/2013/default-sort-or-what-would-shannon-do/
- **54.** Weisstein , E. W. (n.d.). *Entropy* . Retrieved from http://mathworld.wolfram.com/Entropy.html

This is an electronic version of an article published in:

Padilla, Thomas G., and Devin Higgins. 2014. “Library Collections as Humanities Data: The Facet Effect.” *Public Services Quarterly* 10 (4): 324–35. doi:10.1080/15228959.2014.963780.

Public Services Quarterly is available online at: doi:10.1080/15228959.2014.963780