DE GRUYTER
OPEN

Perspective

Smart Data for Digital Humanities

Marcia Lei Zeng[†]

School of Library & Information Science, Kent State University, Kent, Ohio, OH 44240, USA

Marcia Lei Zeng is Professor of Library and Information Science at Kent State University. She holds a Ph.D. from the School of Information Sciences at the University of Pittsburgh and an M.A. from Wuhan University in China. Her major research interests include knowledge organization systems (KOS), Linked Data, metadata and markup languages, smart data and big data, database quality control, semantic technologies, and digital humanities. Her scholarly publications consist of more than 90 papers and five books, as well as over 200 national and international conference presentations and invited lectures. Her research projects have received funding from the US National Science Foundation (NSF), Institute of Museum and Library Services (IMLS), OCLC Online Computer Library Center, Fulbright, and other foundations. Dr. Zeng has chaired or served on committees, working groups, and executive boards for the International Federation of Library Associations and Institutions (IFLA), Special Libraries Association (SLA), Association for Information Science and Technology (ASIS&T), the US National Information Standards Organization (NISO), the International Organization for Standardization (ISO), Dublin Core Metadata Initiative (DCMI), International Society for Knowledge Organization (ISKO), and the World Wide Web Consortium (W3C).

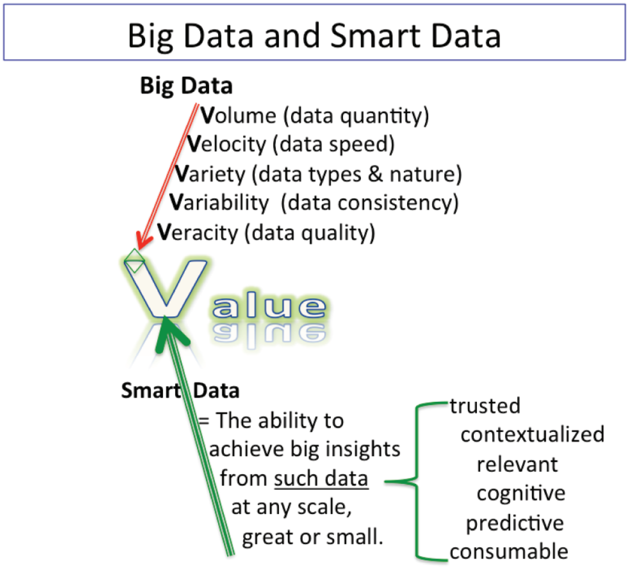
Citation: Marcia Lei Zeng (2017). Smart Data for Digital Humanities.
Vol. 2 No. 1, 2017
pp 1–12
DOI: 10.1515/jdis-2017-0001
Received: Jan. 13, 2017
Revised: Jan. 16, 2017
Accepted: Jan. 16, 2017

Abstract: The emergence of "Big Data" has been a dramatic development in recent years. Alongside it, a lesser-known but equally important set of concepts and practices has also come into being—"Smart Data." This paper shares the author's understanding of *what*, *why*, *how*, *who*, *where*, and *which data* in relation to Smart Data and digital humanities. It concludes that, challenges and opportunities co-exist, but it is certain that Smart Data, the ability to achieve big insights from trusted, contextualized, relevant, cognitive, predictive, and consumable data at any scale, will continue to have extraordinary value in digital humanities.

The emergence of "Big Data" has been a dramatic development in recent years. Alongside it, a lesser-known but equally important set of concepts and practices has also come into being—"Smart Data."

[†] Corresponding author: Marcia Lei Zeng (E-mail: mzeng@kent.edu).

WHAT is Smart Data? Big data has been characterized by multiple "V"s, with the number of "V"'s still increasing. Volume, Velocity, and Variety have been joined by Variability and Veracity (refer to Figure 1). Big Data can bring big Value, if used appropriately, because it is now possible to find the hidden patterns, the unexpected correlations, and the surprising connections within large datasets through effective processing (Gardner, 2012). The realization of the last "V", Value, is dependent on "Smart Data," the "ability to achieve big insights from trusted, contextualized, relevant, cognitive, predictive, and consumable data at any scale, great or small" (Kobielus, 2016, p. 8). Simply speaking, Smart Data makes sense out of Big Data. It provides value from harnessing the challenges posed by Volume, Velocity, Variety and Veracity of Big Data, in-turn providing actionable information and improving decision making (Sheth, 2014). Smart Data "is the way in which different data sources (including Big Data) are brought together, correlated, analyzed, etc., to be able to feed decision-making and action processes" (Iafrate, 2015, p. 13) (Figure 1).



Compiled based on Kobielus, J. (2016, June). The evaluation of Big Data to Smart Data. Keynote at Smart Data Online 2016

Figure 1. Big Data and Smart Data.

WHY Smart Data? Data in the 21st century, like oil in the 18th century, is an untapped asset that holds immense value for those who can learn to extract and use it. "Data is the new oil " (Humby, 2006) has become a defining phrase used by many in recent years as the evidence became more and more convincing. "However, in its raw form, data is just like crude oil; it needs to be refined and processed in order to generate real value. Data has to be cleaned, transformed, and analyzed to unlock its hidden potential" (TiECON East, 2014). According to a 2012 report on the "digital universe"—a measure of all the digital data created, replicated, and consumed in a single year—"even with a generous estimate, the amount of information in the digital universe that is 'tagged' accounts for only about 3% of the digital universe in 2012, and that which is analyzed is half a percent of the digital universe" (Gantz & Reinsel, 2012, p. 3). Extracting Value from Big Data characterized by the other "V"s presents both great challenges and inestimable opportunities. Only after it has been tamed through organization and integration processes is such data turned into Smart Data that reflects the research priorities of a particular discipline or field. These tamed results, as Smart Data inquiries, can then be used to provide comprehensive analyses and generate new products and services (Gardner, 2012; Mukerjee, 2014; Schöch, 2013; TiECON East, 2014).

HOW to transform Big Data into Smart Data? A look at the topics presented at Smart Data conferences since 2015 may provide a good overview of the technologies involved in Smart Data strategies of achieving big insights from trusted, contextualized, relevant, cognitive, predictive, and consumable data at any scale. These include: cognitive computing, deep learning, machine learning, artificial intelligence, predictive analytics, graph databases, machine intelligence, voice processing, semantic technologies, autonomous vehicles, Big Data, data science, Internet of Things (IoT), text analysis, Resource Description Framework (RDF), knowledge graphs, contextual computing, Linked Data, deep reasoning, ontologies, JSON-LD[1], common sense, natural language processing (NLP), and semantic search (DATAVERSITY, 2017). These topics are closely interrelated and overlapping. For example, deep learning shows the great potential in natural language processing; cognitive computing uses machine learning to find deep patterns (including those not obviously statistical) within complex, unstructured, and streaming data. Some of the topics have moved beyond the original territory conveyed by these labels for years. For example, "artificial intelligence" is a field that has changed dramatically in the 21st century. Meanwhile, the topics of the Smart Data conferences reflect the varied applications of the W3C standards for the Semantic Web, including—but not limited to—RDF, Linked Data, ontologies, graph databases, semantic search, and other semantic technologies (Figure 2).

[1] JSON-LD is a lightweight Linked Data format.

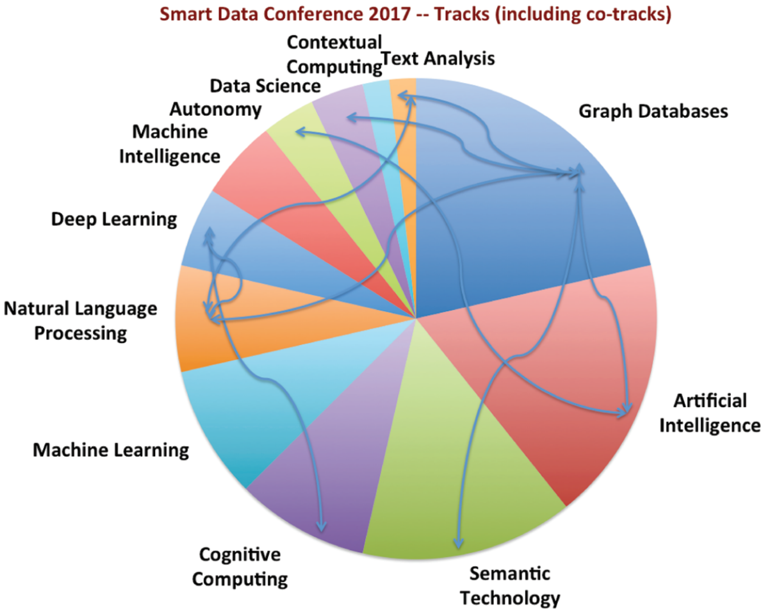
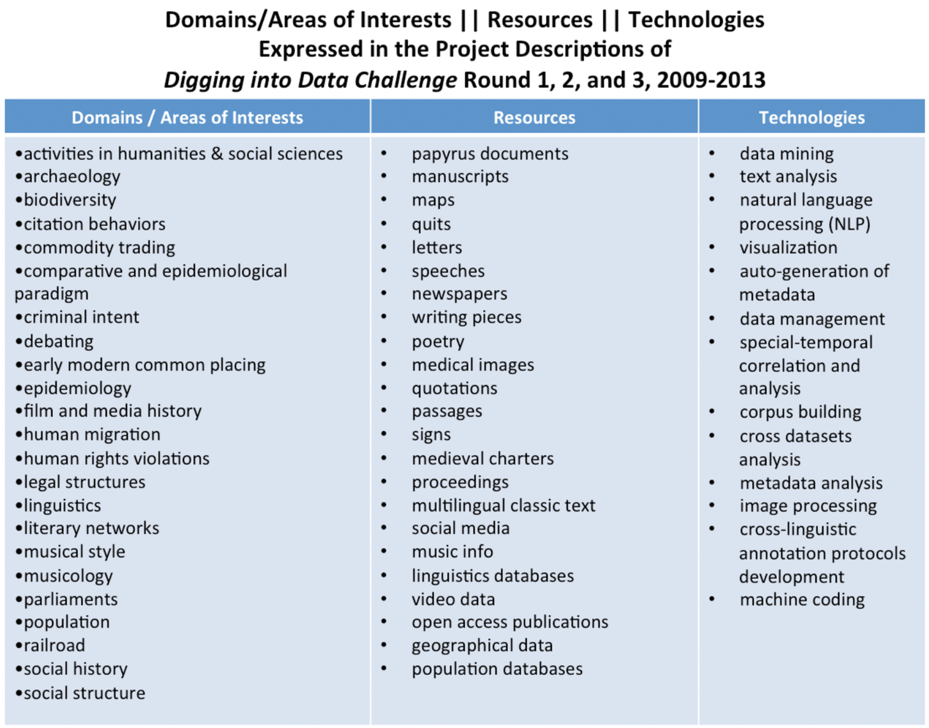


Figure 2. Smart Data Conference 2017 tracks, including combined co-tracks (marked by arrows). Source: Compiled according to the program at http://smartdata2017.dataversity.net/.

WHO is making/using Smart Data? Efforts to tame Big Data using Smart Data strategies have been made by various experts including natural scientists, engineers, business executives and financial analysts, practitioners of medicine, and government agents. In humanities, the word "Smart Data" is not universally used, even though the approaches can be recognized in many research projects over the last six years. Since 2009, through the *Digging into Data Challenge* program (https://diggingintodata.org/), research funders from more than 10 countries have funded dozens of projects aimed at research questions in the humanities and/or social sciences. The sponsors in the USA include the National Endowment for the Humanities (NEH), the National Science Foundation (NSF), and the Institute of Museum and Library Services (IMLS). Based on the project descriptions of the last three rounds (the most recent, the 4th round has not announced final winners as of the date this paper was written), the resources include mainly unstructured data assets originating in ancient times, while structured datasets created in the digital age are also used. The domains and areas of interests are widely spread in the humanities and social sciences. Technologically, large-scale data analyses have been applied to research questions in the fields using the Smart Data approaches (refer to the above "How" section). Methodologically, the projects are interdisciplinary and strive to show how best to tap data in large scale and diverse formats in order to search for key insights while also ensuring access to such data by humanities and social science researchers through new technology-supported tools (Figure 3).



Domains/Areas of Interests || Resources || Technologies
Expressed in the Project Descriptions of
***Digging into Data Challenge* Round 1, 2, and 3, 2009-2013**

Domains / Areas of Interests	Resources	Technologies
•activities in humanities & social sciences •archaeology •biodiversity •citation behaviors •commodity trading •comparative and epidemiological paradigm •criminal intent •debating •early modern common placing •epidemiology •film and media history •human migration •human rights violations •legal structures •linguistics •literary networks •musical style •musicology •parliaments •population •railroad •social history •social structure	• papyrus documents • manuscripts • maps • quits • letters • speeches • newspapers • writing pieces • poetry • medical images • quotations • passages • signs • medieval charters • proceedings • multilingual classic text • social media • music info • linguistics databases • video data • open access publications • geographical data • population databases	• data mining • text analysis • natural language processing (NLP) • visualization • auto-generation of metadata • data management • special-temporal correlation and analysis • corpus building • cross datasets analysis • metadata analysis • image processing • cross-linguistic annotation protocols development • machine coding

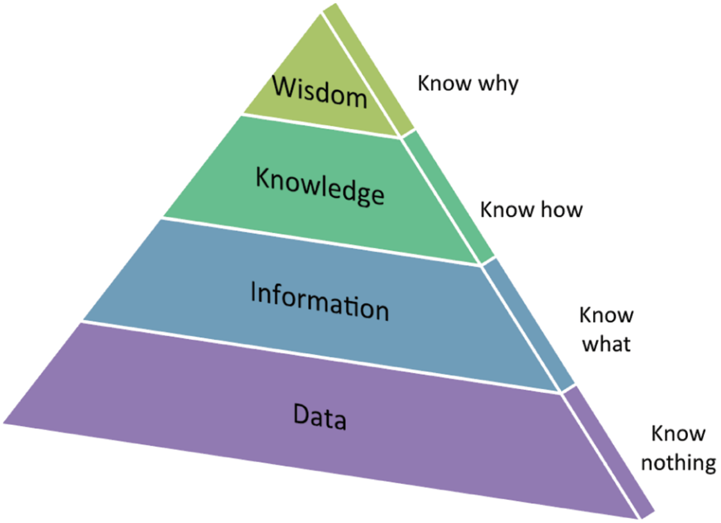
Figure 3. Domains/areas of interests, resources, and technologies expressed in the project descriptions of *Digging into Data Challenge* Round 1, 2, and 3, 2009–2013. Source: Compiled based on the project descriptions retrieved from the website at https://dev.diggingintodata.org/awards.

A newly launched nationwide contest encouraging the use of data from *Chronicling America*'s digital repository of historic US newspapers, as well as the new Humanities Access Grant funded by the NEH (2016), are further signs of initiatives taking place at the intersection between the humanities and digital technologies. While the multifaceted landscape of digital humanities is yet to be fully understood, the highly competitive Digital Humanities (DH) conferences could give us more clues. The self-tagged topics of submissions from DH 2013 to DH 2016 conferences reflected a multi-disciplinary nature: *text analysis* was number one in submission count, followed by *historical studies*, *data mining/text mining*, *archives & repositories*, *literary studies*, and *data visualization*. The 2017 count, which separated topics from disciplines, shows that those top topics are joined by *interdisciplinary collaboration* and *corpora and corpus activities*. The disciplines that have more than 100 submissions are: *computer science, literary studies*, *library and information science*, *cultural studies*, and *historical studies*. A notable finding is that submissions from *film and media studies* have greatly increased compared to previous years, as have other *non-textual* disciplines. There has also been a steady increase of new authors entering the field and of co-authorship of submissions (Weingart, 2016; 2017). Overall, a wide range of disciplines and approaches are seen in the humanities to reach "bigger smart data" or "smarter big data" (Schöch, 2013), as demonstrated by the outcomes presented at digital humanities conferences, the government funded research projects, and new initiatives and publications all over the world in the past six years.

WHERE is the distinctive mark in Digital Humanities? It might be natural that, when thinking about digital humanities in the data-intensive research projects, people would look for distinctive marks toward the direction of technologies. However, as Schöch (2013) pointed out, the distinctive mark of Big Data in the humanities seemed to be a methodological shift rather than a primarily technological one. Further scrutiny of the methodological shift in humanities highlights the role of Big Data and Smart Data for every field of knowledge. In short, the relationship between Big Data and Smart Data can be characterized as "what it is" and "what it is for" (Iafrate, 2015). This view of turning Big Data into Smart Data brings us back to the well-known Data-Information-Knowledge-Wisdom (DIKW) pyramid (Zeleny, 1987; Ackoff, 1988) which represents the most basic strategy for understanding a world that far exceeds our brains' capacity by filtering, winnowing, and otherwise reducing it to something more meaningful (Figure 4).

Nevertheless, the Smart Data approach is not simply a replication of the DIKW path because Smart Data is based on Big Data's methodology, which assumes the ability to reveal the *unknown-unknowns* (Borne, 2013) instead of taking the approach that one knows to do something in order to prove or disapprove the *known-unknowns* (Figure 5). This is a fundamental advancement of Smart Data that distinguishes it from other approaches that follow the more traditional blueprint of hypothesizing, modeling, and testing (Anderson, 2008).

One good example of revealing the unknown-unknowns through Smart Data is the research project "A network framework of cultural history" published in *Science* and also on *Nature Video* (Schich et al., 2014a; 2014b). A multidisciplinary research team provided a macroscopic perspective of the cultural history of Europe and North America across 3,000 years, using simple—but large—datasets of the birth and death locations of more than 150,000 notable individuals, which revealed previously undocumented human mobility patterns and cultural attraction dynamics. Incorporating this analyzed data, the 3,000 year, large-scale patterns in European



Compiled based on Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, *16*(1), 3-9. and Zeleny, M. (1987). Management support systems: towards integrated knowledge management. *Human Systems Management*, *7*(1), 59-70.

Figure 4. The Data-Information-Knowledge-Wisdom (DIKW) pyramid.

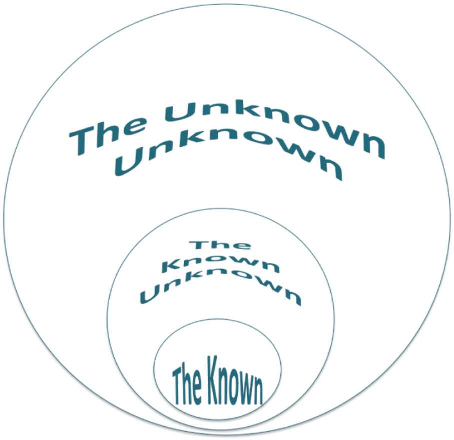


Figure 5. The unknown-unknowns.

and American cultural life are visualized and brought to life, enlightening the formation of intellectual and cultural centers, the rising and crumbling of empires, and other influential factors, all beyond the scope of specific events or narrow time intervals (Recommend watching the video at https://www.youtube.com/watch?v=4gIhRkCcD4U). The value of the knowledge is incredible and the big insights are achieved from trusted, contextualized, relevant, cognitive, predictive, and consumable data (the original sources are structured data from Freebase (now Wikidata), the General Artist Lexicon (AKL), and the Getty Union List of Artist Names (ULAN)) (Schich et al., 2014a). This example not only demonstrates the potential of the Smart Data approach in sociology, anthropology, and history in general but also indicates a significant methodological advancement in the humanities.

WHICH DATA can be found in supporting research and scholarship in Digital Humanities? When putting Big Data and Smart Data into the context of digital humanities, a key concept that needs to be agreed upon is the use of the term "data." In the digital age, it is common for people to only think of data in terms of digitally available formats. The connection between digital data and data analytics is correct, but we need to fully understand that the terms "data" and "digital data" are not equivalent. The types of data are also not limited to quantitative data. The Reference Model for an Open Archival Information System (OAIS) defined data as a "reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing" while offering examples of data as: a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen. This definition of "data" was given within the context of "information," which is "Any type of knowledge that can be exchanged. In an exchange, it is represented by data" (Consultative Committee for Space Data Systems, 2012, p. 1–10 and p. 1–12). After a comprehensive review of the definitions and terminology for "data" in her book titled *Big data, little data, no data: Scholarship in the networked world*, Borgman (2015) presented an overarching summary that "data are representations of observations, objects, or other entities used as evidence of phenomena for the purpose of research or scholarship" (Borgman, 2015, p. 28).

In the data resources that are usually served through libraries, archives, and museums (LAMs) and other information institutions, the types of data, which are available in the largest quantity, have the diversity in type, nature, and quality, and are the most challenging to process, belong to *unstructured data* found in documents and other information-bearing objects (textual or non-textual, digitized or non-digitized) in all kinds of formats (examples can be found in Figure 3). These primary data resources are held in special collections, archives, oral history files, annual reports, provenance indexes, and inventories, to name just a few. The nature of such data is quite different from, for instance, that of the data used by the "digital universe" that "is made up of images and videos on mobile phones uploaded to YouTube, digital movies populating the pixels of our high-definition TVs, banking data swiped in an ATM, security footage at airports and major events such as the Olympic Games, subatomic collisions recorded by the Large Hadron Collider at

CERN, transponders recording highway tolls, voice calls zipping through digital phone lines, and texting as a widespread means of communications" (Gantz & Reinsel, 2012, p. 1). Such a "digital universe" may not be the major or only source for humanities researchers.

In fact, one primary challenge in applying the Smart Data approach to digital humanities is the availability of data resources for those in need of historical data that one could not obtain through Web crawling or scraping. There is no doubt that Smart Data approaches that have been tested and implemented in business and industry can be applied to the digital humanities. Nevertheless, how to "datafy" the unstructured data (i.e. turn the heritage materials into not only machine-readable but also machine-processable resources, and reconstruct through digitization pipelines) before the researchers can make use of data analytics technologies? This fundamental question might explain why, for digital humanities, the Smart Data approach emphasizes the organization and integration processes to transform unstructured data to structured and semi-structured data (Kaplan, 2015; Mayer-Schönberger & Cukier 2013; Schöch, 2013).

In addition to the *unstructured data* discussed above, LAMs and other information institutions also provide tremendous opportunities for humanities researchers to dig nuggets of gold from *semi-structured data* (examples include the intellectual works encoded following the Text Encoding Initiative (TEI) guidelines, archival finding aids, value-added or tagged resources that exist in all kinds of formats, and the unstructured portions of otherwise structured datasets) as well as *structured data* (including bibliographies, indexing and abstracting databases, citation indexes, catalogs of all kinds, special collection portals, metadata repositories, curated research datasets, and name authorities) (Figure 6).

These datasets might be relatively small in volume and have limited heterogeneity in comparison with Big Data, but they are clean, explicit, trusted, and value-added, and their creation is governed mostly by human decisions. More promisingly, they are among the resources most likely to be freely accessible (non-proprietary and non-commercial). These make them treasures for all humanities researchers and beyond. In his speech titled "Contextual Computing with Knowledge Graphs and the Web of Entities" at Smart Data Online 2016, Richard Wallis, a well-known pioneer of the library community's Linked Open Data movement, provided his vision of *contextual computing*, in which he listed elements such as meaning, syntax, time, location, appropriate domain, regulations, user's profile, process, task, and goal. The revolutionary work of WorldCat Linked Data and the WorldCat Entities experiment at OCLC are among the successful cases. By providing millions of entities of intellectual works, places, concepts, persons, organizations, events, and other types of tamed data together, the WorldCat Entities shows how the structured

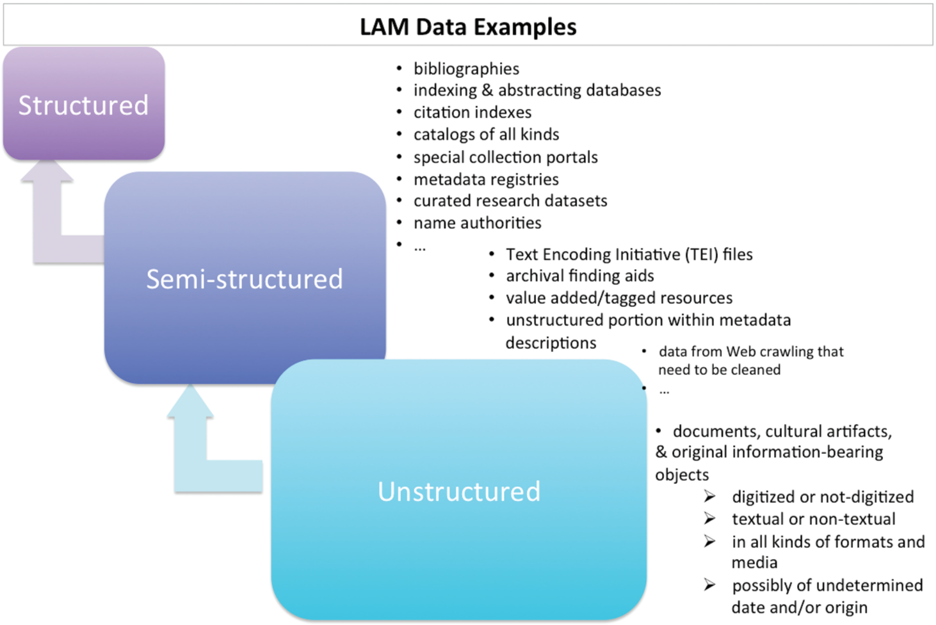

LAM Data Examples
Structured
Semi-structured
Unstructured
• bibliographies
• indexing & abstracting databases
• citation indexes
• catalogs of all kinds
• special collection portals
• metadata registries
• curated research datasets
• name authorities
• …
• Text Encoding Initiative (TEI) files
• archival finding aids
• value added/tagged resources
• unstructured portion within metadata descriptions
• data from Web crawling that need to be cleaned
• …
• documents, cultural artifacts, & original information-bearing objects
➢ digitized or not-digitized
➢ textual or non-textual
➢ in all kinds of formats and media
➢ possibly of undetermined date and/or origin

Figure 6. Examples of the data resources provided by libraries, archives, and museums (LAMs).

data provided by LAMs can enrich knowledge graphs and Linked Open Data datasets infinitely (Wallis, 2016).

In the processes that transform unstructured data to structured and semi-structured data, the Smart Data strategy drives data service providers to aim at machine-*understandable*, -*processable*, and -*actionable* (instead of merely machine-*readable*) data, to provide accurate data in the processes of interlinking, citing, transferring, rights-permission management, use and reuse, and to enable both one-to-many usages and high efficiency processing of data for digital humanities.

Conclusion

Today, advanced technologies, under the umbrella of Big Data and Smart Data, allow researchers of the humanities to join the mainstream of the digital age with new abilities as never before: to access and reuse large volumes of diverse data; to unearth patterns and connections formerly hidden from view; to reconstruct the past; to discover the impact and value of qualitative and quantitative variables in both real and virtual environments; and to bring the knowledge of the complex intricacies of human society to light. Challenges and opportunities co-exist, but it is certain that Smart Data, the ability to achieve big insights from trusted, contextualized, relevant, cognitive, predictive, and consumable data at any scale, will continue to have extraordinary value in digital humanities.

References

Ackoff, R.L. (1989). From data to wisdom. Journal of Applied Systems Analysis, 16(1), 3–9.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. Wired, 16(7). Retrieved on December 5, 2016, from https://www.wired.com/2008/06/pb-theory/.

Borgman, C. (2015). Big data, little data, no data: Scholarship in the networked world. Cambridge, MA: MIT Press.

Borne, K. (2013). Big data, small world: Kirk Borne at TEDxGeorgeMasonU [Video file]. Retrieved on December 15, 2016, from https://www.youtube.com/watch?v=Zr02fMBfuRA.

Consultative Committee for Space Data Systems. (2012). Reference model for an open archival information system (OAIS): Recommended practice (CCSDS 650.0-M-2: Magenta Book). Washington, DC: CCSDS. Retrieved on December 15, 2016, from http://public.ccsds.org/publications/archive/650x0m2.pdf.

DATAVERSITY Education, LLC. (2017). Smart Data Conference (website). Retrieved on January 12, 2017, from http://smartdata2017.dataversity.net.

Digging into data challenge. (n.d.) Retrieved on January 10, 2017, from https://diggingintodata.org/.

Gardner, D. (2012). An ocean of data [Introduction]. In R. Smolan, & J. Erwitt (Eds.), The Human Face of Big Data (pp. 14–17). Sausalito, CA: Against All Odds Productions.

Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East. IDC iView, December 2012, 1–16. Retrieved on January 10, 2017, from http://www.dedupecentral.co.uk/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

Humby, C. (2006). Data is the new oil. Talk given at the Association of National Advertisers (ANA) Senior Marketer's Summit, Kellogg School. (Source: M. Palmer, M. (2006 Nov. 3). Data is the New Oil (Web log post)). Retrieved on January 10, 2017, from http://ana.blogs.com/maestros/2006/11/data_is_the_new.html.

Iafrate, F. (2015). From big data to smart data. London: ISTE Ltd., and Hoboken, NJ: John Wiley & Sons, Inc.

Kaplan, F. (2015). A map for big data research in digital humanities. Frontiers in Digital Humanities, 2, p. 1. Retrieved on January 10, 2017, from https://owl.english.purdue.edu/owl/resource/560/10/.

Kobielus, J. (2016, June). The evolution of big data to smart data [PowerPoint slides]. Keynote at Smart Data Online 2016.

Mayer-Schönberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. New York, NY: Eamon Dolan/Houghton Mifflin Harcourt.

Mukerjee, P. (2014). Introduction to data science [PowerPoint slides]. Retrieved on January 10, 2017, from http://www.slideshare.net/prithwis/01-intro2-datascienceyantrajaalblog.

National Endowment for the Humanities (NEH). (2016). Grants. Retrieved on January 10, 2017, from https://www.neh.gov/grants.

Schich, M., Song, C., Ahn, Y.Y., Mirsky, A., Martino, M., Barabási, A.L., & Helbing, D. (2014a). A network framework of cultural history. Science, 345(6196), 558–562.

Schich, M., Song, C., Ahn, Y.Y., Mirsky, A., Martino, M., Barabási, A.L., & Helbing, D. (2014b, July 31). Charting culture. Nature Video [Video file]. Retrieved on January 10, 2017, from https://www.youtube.com/watch?v=4gIhRkCcD4U.

Schöch, C. (2013). Big? smart? clean? messy? Data in the humanities. Journal of Digital Humanities, 2(3), 2–13.

Sheth, A. (2014). Transforming big data into smart data: Deriving value via harnessing volume, variety and velocity using semantics and semantic web [PowerPoint Slides]. Keynote at 30th IEEE International Conference on Data Engineering (ICDE) 2014. Retrieved on January 10, 2017, from http://ieeexplore.ieee.org/document/6816634/.

TiECON East. (2014). Data is the new oil. Retrieved on January 10, 2017, from http://www.tieconeast.org/2014/big-data-analytics.

Wallis, R. (2016). Contextual computing with knowledge graphs and the Web of Entities. Presentation at Smart Data Online 2016.

Weingart, S. (2016). Submissions to DH2016 (pt. 1) [Web log post]. Retrieved on January 10, 2017, from http://www.scottbot.net/HIAL/index.html@tag=dhconf.html.

Weingart, S. (2017). Submissions to DH2017 (pt.1) [Web log post]. Retrieved on January 10, 2017, from http://scottbot.net/submissions-to-dh2017-pt-1/.

Zeleny, M. (1987). Management support systems: Towards integrated knowledge management. Human Systems Management, 7(1), 59–70.