DE GRUYTER
OPEN

Expert Review

Big Metadata, Smart Metadata, and Metadata Capital: Toward Greater Synergy Between Data Science and Metadata

Jane Greenberg†

College of Computing and Informatics, Drexel University, Philadelphia, PA 19104, USA

Citation: Jane Greenberg (2017). Big Metadata, Smart Metadata, and Metadata Capital: Toward Greater Synergy Between Data Science and Metadata.

Vol. 2 No. 3, 2017
pp 19–36
DOI: 10.1515/jdis-2017-0012

Received: Jun. 10, 2017
Revised: Jul. 24, 2017
Accepted: Jul. 26, 2017

Jane Greenberg is the Alice B. Kroeger Professor and Director of the Metadata Research Center① at the College of Computing and Informatics, Drexel University. Her research activities focus on metadata, knowledge organization/semantics, linked data, data science, and economics. She serves on the advisory board of the Dublin Core Metadata Initiative (DCMI) and the steering committee for the NSF Northeast Big Data Innovation Hub (NEBDIH)②. She is a principal investigator on the NSF Spoke initiative, "A Licensing Model and Ecosystem for Data Sharing," and the principal investigator for the Metadata Capital Initiative (MetaDataCAPT'L) and the Helping Interdisciplinary Vocabulary Engineering (HIVE) linked data project. She is also a co-principal investigator (CoPI) for Drexel's NSF Industry/University Collaborative Research Center (NSF-I/UCRC), the Center for Visual and Decision Informatics (CVDI)③. Her research has been funded by US National Science Foundation (NSF), National Institutes of Health (NIH), Institute of Museum and Library Services (IMLS), CVDI, National Consortium for Data Science, Microsoft Research, National Library of Medicine, Library of Congress, OCLC Online Computer Library Center, among other organizational and private sponsors. She has received numerous awards and honors for her research and leadership; most recently she was recognized as a 2016 ELATE at Drexel® Fellow, a 2014 Data Science Fellow at the National Consortium for Data Science, and, in 2012, she held a Chair of Excellence at the University of Juan Carlos III, Madrid, Spain.

Abstract

Purpose: The purpose of the paper is to provide a framework for addressing the disconnect between metadata and data science. Data science cannot progress without metadata research.

† Corresponding author: Jane Greenberg (E-mail: jg3243@drexel.edu).

① http://cci.drexel.edu/mrc/
② http://nebigdatahub.org/
③ http://www.nsfcvdi.org/

Expert Review

This paper takes steps toward advancing the synergy between metadata and data science, and identifies pathways for developing a more cohesive metadata research agenda in data science.

Design/methodology/approach: This paper identifies factors that challenge metadata research in the digital ecosystem, defines metadata and data science, and presents the concepts *big metadata*, *smart metadata*, and *metadata capital* as part of a metadata *lingua franca* connecting to data science.

Findings: The "utilitarian nature" and "historical and traditional views" of metadata are identified as two intersecting factors that have inhibited metadata research. Big metadata, smart metadata, and metadata capital are presented as part of a metadata *lingua franca* to help frame research in the data science research space.

Research limitations: There are additional, intersecting factors to consider that likely inhibit metadata research, and other significant metadata concepts to explore.

Practical implications: The immediate contribution of this work is that it may elicit response, critique, revision, or, more significantly, motivate research. The work presented can encourage more researchers to consider the significance of metadata as a research worthy topic within data science and the larger digital ecosystem.

Originality/value: Although metadata research has not kept pace with other data science topics, there is little attention directed to this problem. This is surprising, given that metadata is essential for data science endeavors. This examination synthesizes original and prior scholarship to provide new grounding for metadata research in data science.

Keywords Metadata research; Data science; Big metadata; Smart metadata; Metadata capital

1 Introduction

Metadata, as a form of data affixed to other data, is indispensable to data science and the interconnected domain of data analytics. Metadata describes data, provides context, and is vital for accurate data interpretation and use by both humans and machines. Given this dependency, it is logical to conclude that metadata innovation ought to have progressed in tandem with advances in big data and data science. To this end, leading data science journals and conferences have been increasing coverage of metadata research and development (R&D). Examples include *Data Science Journal* (v. 15, 2016) dedicated to data models, and a recent issue of the *Journal of Data and Information Science* (*JDIS*) with Li and Sugimoto's (2017) work on long-term maintenance of metadata. Another example is the 2nd IEEE Workshop on Big Data Metadata and Management (BDMM, 2017)[④], to be hosted at the 2017 IEEE International Conference on Big Data in Boston, Massachusetts,

④ http://cci.drexel.edu/bigdata/bigdata2017/Workshops.html

USA. Nevertheless, despite such important developments, metadata R&D has not kept pace within the greater framework of data science.

Research literature and notable reports reveal the metadata research lag in data science. For example, Smith et al. (2014) call for metadata research, emphasizing that "current big data ecosystems lack a principled approach to metadata management." Another clear example is the US report entitled *The Federal Big Data Research and Development Strategic Plan* (NITRD, 2016). This report stresses the need for research on metadata frameworks to ensure data trustworthiness, and identifies a myriad of metadata-related research topics, many of which are found in similar governmental and disciplinary reports worldwide (e.g. ERAC Secretariat, 2016; Ilevbare, Athanassopoulou, & Wooldridge, 2017). Collectively, these examples make clear the absence of a coherent metadata research agenda in data science. This gap raises questions about why there is a disconnect between metadata and the larger sphere of data science research, and how to address this challenge.

This article considers these questions as steps toward advancing the synergy between metadata and data science. The following section describes metadata and data science, followed by discussion of two intersecting factors that challenge metadata research in the digital ecosystem. Next, the paper introduces the concepts *big metadata*, *smart metadata*, and *metadata capital*. These concepts are presented as contributions to the metadata *lingua franca* connecting to the data science space. The conclusion summarizes key discussion points and considers next steps for advancing metadata research in the area of data science.

2 Metadata and Data Science Defined

Exploring the interconnection between metadata and data science requires a review of these two concepts.

2.1 Metadata: A Value-added Language

Metadata has been loosely defined and popularized as data about data, information about information. More comprehensive definitions address metadata as structured data supporting functions associated with an *object*, an object being any "entity, form, or mode" (Greenberg, 2005, 2010; Lytras, Sicilia, & Cechinel, 2013). Examples of metadata functions include data discovery, access, use, provenance tracking, authenticity and security verification, preservation management, and other activities throughout the data lifecycle (UK Data Archive, 2012). Metadata researchers draw on these functions to create typologies identifying different metadata types (Méndez & van Hooland, 2013). Business and data warehousing typologies generally include business and technical metadata, and, at times, also include process and operational metadata (Dong et al., 2016; Shankaranarayanan &

Even, 2006; Vaduva & Dittrich, 2001). The digital library, archive, and repository communities have categories, such as descriptive, technical, preservation, provenance, and usage metadata (Zeng & Qin, 2016). In these cases, the "types" of metadata connect to the lifecycle of the object being represented or tracked.

To understand the full extent of metadata, it is important to recognize that the adjectival label "metadata" is not always used when, in fact, the data of interest has a *meta* status. In other words, data that is meta, an abstraction of another object, is not always labeled as metadata. Common examples include *provenance data*, *linked data*, *contextual data*, and *authenticity data*. These data exist only because of the actuality of other objects, and can only occur as a result of an object's activity.

Beyond labeling and categorization, metadata can more universally be thought of as value-added language that serves as an integrated layer in an information system. When appropriately placed and accessible, by human or machine, metadata language eloquently enables the interplay between an object, such as data, and the desired activity, such as discovery, access, provenance tracking, calculation, or other directives. To understand metadata research opportunities in data science, it is useful to also review the meaning of data science, as follows.

2.2 Data Science: Leveraging Data to Gain New Insights

Data science is an interdisciplinary field that targets studying and leveraging data to gain insights. A data science undertaking may enable one to predict a phenomenon or automate decision-making. The Data Science Association defines data science as the "scientific study of the creation, validation and transformation of data to create meaning" (Data Science Association, 2017). Data science draws upon the full range of data (small, big, static, structured, unstructured, or streaming), and applies scientific and statistical methodologies to learn from data (van der Aalst, 2016).

Data science has many aspects, and the collection of definitions reveals different emphases. For example, Dhar (2013) focuses on the predictive capabilities of data, emphasizing application of statistical methods. Stanton (2012) offers a broader definition, explaining that data science encompasses a full range of activities, including the "collection, preparation, analysis, visualization, management, and preservation of large collections of information." The unifying factor across various definitions is the "science" that comprises defining appropriate questions, selecting and obtaining suitable data, and applying the correct, at times often innovative, modeling, and statistical methods.

The "science" of data science indicates a methodological and systematic approach to leveraging data as part of studying a problem or a phenomenon. Data science endeavors rely not only on data, but accurate description of the data—hence metadata. Given the reliance on metadata, one would anticipate appropriate support for, and recognition of, the value of research addressing metadata processes,

applications, and societal impacts. Unfortunately, there are a number of key impediments to understanding the scientific merit of metadata research. These impediments are reviewed below in the context of challenges to metadata R&D.

3 Challenges to Metadata Research

Information and library science, computer science, and a number of disciplinary domains (e.g. biology, medicine, materials science, and geography to name a few) support a generally tightly-knit, robust metadata community through interest groups within a larger association, several targeted conferences, and focused publications, such as the *International Journal of Metadata, Semantics, and Ontologies*. Despite strong, historical grounding, metadata research in data science, and the larger digital ecosystem, is restrained by not being considered a *true* scientific endeavor. More specifically, challenges to metadata research stem, to a large degree, from two intersecting factors: 1) the utilitarian nature of metadata, and 2) historical and traditional perceptions of metadata.

3.1 The Utilitarian Nature of Metadata

Metadata is generally viewed as a practical application relating to cataloging, indexing, database development, and the recording of digital transactions. This point is underscored in "Metadata in Everyday Life," the first section in NISO[5]'s new primer, *Understanding Metadata* (Riley, 2017). To be clear, seeking pragmatic solutions with metadata is vital to nearly any digital undertaking; however, a pragmatic emphasis can challenge research opportunities. An example here is the rationalist approach pursued for schema design. That is, data dictionaries and metadata application development are commonly based on practical experience, rather than substantive empirical or theoretical approaches.

Another utilitarian aspect affecting perceptions of metadata stems from the pressing need for metadata to accommodate the exponential growth of data and the larger digital ecosystem, which limits resources (time, personnel, and finances) that could otherwise be allocated toward deeper metadata research analyses and theoretical development (Greenberg, 2009). As noted above, there is a robust metadata research community; however, the pragmatic strength and necessities of metadata have very likely impeded development of a more rigorous metadata agenda in data science.

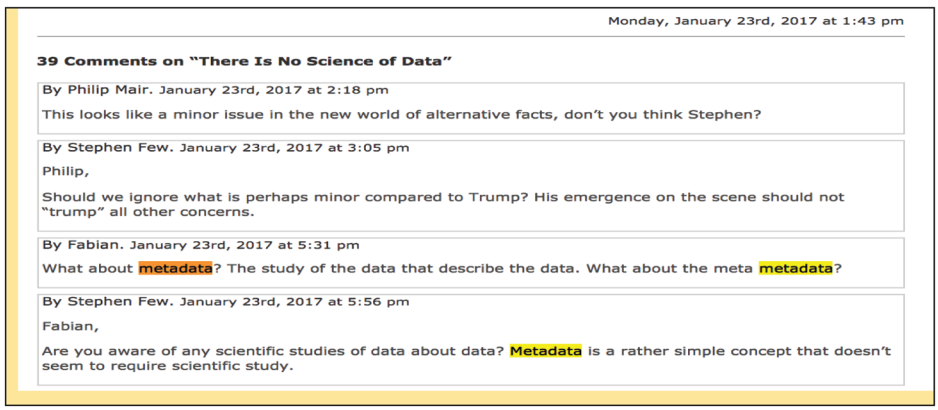
3.2 Historical and Traditional Perceptions of Metadata

Metadata carries baggage similar to that of cataloging (Coleman, 2005; Tennant, 2002). Specific criticisms address the Semantic Web, with claims that ontologies

[5] National Information Standards Organization (http://www.niso.org/home/)

cannot support automatic reasoning (Shirky, 2005), the mark-up is excessive (Manian, 2011), and that the goals underlying linked data are unrealistic. There is also concept of "metacrap," coined by Doctorow (2001), referring to the impossibility of "exhaustive, reliable metadata" due to "insurmountable obstacles," and proclamations that automated methods will take over, obviating the need to investigate metadata (Dimitrova, 2004). The metadata community has internal critics as well, as demonstrated by Beall's "*Dublin Core: An Obituary*" (2004), and his later piece, "*Dublin Core is Still Dead*" (2014). Both articles lambaste the Dublin Core metadata standard, despite the fact that it is one of the most universally adopted, cross-disciplinary, and internationally used metadata standards.

Traditional perceptions are further reflected in differing opinions about metadata and what constitutes a science. An illustrative example is found in "There Is No Science of Data," a discussion on *Visual Business Intelligence: A blog by Stephen Few*, wherein the author states, "Metadata is a rather simple concept that doesn't seem to require scientific study" (Few, 2017; see Figure 1).


Monday, January 23rd, 2017 at 1:43 pm

39 Comments on "There Is No Science of Data"

By Philip Mair. January 23rd, 2017 at 2:18 pm

This looks like a minor issue in the new world of alternative facts, don't you think Stephen?

By Stephen Few. January 23rd, 2017 at 3:05 pm

Philip,

Should we ignore what is perhaps minor compared to Trump? His emergence on the scene should not "trump" all other concerns.

By Fabian. January 23rd, 2017 at 5:31 pm

What about metadata? The study of the data that describe the data. What about the meta metadata?

By Stephen Few. January 23rd, 2017 at 5:56 pm

Fabian,

Are you aware of any scientific studies of data about data? Metadata is a rather simple concept that doesn't seem to require scientific study.

Figure 1. *Visual Business Intelligence: A blog by Stephen Few* (January 23, 2017).

Few (the blog author) has over 30 years of experience in business intelligence and information design, and his viewpoint clearly illustrates that many simply equate metadata with the nuts-and-bolts of an information system, rather than a research-worthy topic. Few continues his blog discussion by observing that information science and data science are also misnomers; despite the fact that a discussion contributor, named Konrad, shares, "Actually there is a whole academic discipline dedicated to the study of information. . . ," (Konrad. January 24, 2017 at 1:11 am). Further, this participant references the Wikipedia entry for information science, which is substantive with credible references confirming existence of the discipline. Although continued discussion of what is a science extends beyond the

scope of this *JDIS* metadata-focused paper, it is important to recognize that differing opinions impact views on what is a research worthy topic.

3.3 Summary: Moving Past the Impediments

Overall, the discussion above provides insight into why metadata research faces impediments in data science, and other disciplines. Nevertheless, the value of metadata cannot be denied. In fact, the significance of metadata became a mainstream media topic with Edward Snowden's whistleblowing on the US government's surveillance of personal phone record metadata, without individual consent or knowledge of this activity (Greenwald, 2013). In advocating for greater attention to metadata research, the following section presents three concepts to foster dialogue about metadata, and help provide a framework for metadata-focused research in data science.

4 Metadata Concepts Relating to Data Science

Every domain has its *lingua franca*; that is, a language that community members understand and use to correspond. A domain specific language may select or co-opt concepts or terms from another community, and tailor terms for their own needs. Zeng and Qin's "*Metadata Research Landscape*" (Chapter 2, in *Metadata*, 2016) helps document the current metadata research *lingua franca*, covering metadata architecture, modeling, semantics, and data-driven aspects. Commonly used research methods are also part of this *lingua franca*. Examples include content analyses that generally target metadata quality (e.g. Zavalina, 2011), cross-walk analyses for metadata scheme development (e.g. Gaitanou et al., 2016), experiments comparing automatic metadata generation approaches, and semantic mapping assessments (e.g. Vlachidis et al., 2013; White, Willis, & Greenberg, 2014). The existing *lingua franca* forms an important footing for metadata R&D in data science. In advancing the dialogue and advocating for further metadata research, the following section presents three concepts that provide a framework for metadata-focused research in data science.

4.1 Big Metadata

The data science enterprise has been motivated by the availability of massive amounts of digital data and new capacities for data-driven solutions. These ideas are central to the "fourth paradigm," a dimension coined by Microsoft Research visionary Jim Gray, and captured by Hey, Tansley, and Tolle (2009), to explain the growing, unprecedented opportunity for data-driven science. Metadata is a vital component of the fourth paradigm, although the significance of metadata is often overlooked or only noted in a limited way. Metadata can garner new research attention if it is understood as *big metadata*.

Expert Review

Big metadata is both a first-class object and an auxiliary associated with the wide, seemingly countless variety of data formats, types, and genres. Simon's piece, "*Too Big to Ignore: The Business Case for Big Data*" (2013), validates the importance of metadata for big data. He also confirms the existence of "big metadata" in reference to the wide diversity of data. The concept big metadata has also appeared in the research literature. For example, Smith et al. (2014) discuss big metadata in relation to the US government's trove of big data; and Zhao et al. (2014) identify big metadata as a vital data source that can give insight into real world traffic problems.

Beyond an association with big data diversity and size, *big metadata* reflects the wide range of data lifecycle activities found among projects, settings, and systems. Data lifecycle scenarios extend from simple (data creation, capture, storage, and preservation) to complex (data use, reuse, repurposing, and modification), using both human and automatic processes. And, at the data lifecycle meta level is the metadata lifecycle, which generates big metadata.

Big metadata is defined below in Table 1 by the *volume*, *velocity*, *variety*, *variability*, and *value*, built on the common 5Vs used to define big data.

Table 1. The five Vs of big metadata.

Five Vs	Definition
Volume	The quantity and usefulness of metadata generated daily confirms the existence of big metadata. At times metadata is less than or equal to the extent of the data it describes in size (bytes). During other times the metadata exceeds the data being described or tracked, due to the complexity of the data lifecycle activity. Linked data offers an example, with metadata renderings that can be larger than the volume of data object(s) being represented. Like big data, not all big metadata is useful, and a challenge is to identify the big metadata that is useful for data science and analytic endeavors.
Velocity	Metadata is generated via automatic processes at immense speed correlating with rate of digital transactions. For example, searching Google, answering an email, purchasing an item online, and day-to-day office activities such as word processing of all log data, as well as associated metadata.
Variety	Metadata reflects the wide variety of data formats, types, and genres along with the extensive range of data and metadata lifecycles. In addition, the different types of metadata (e.g. discovery, technical, preservation, etc.) as well as unique domain specific metadata requirements intensify the variety.
Variability	There is an unmistakable unevenness of metadata across the digital ecosystem. Lack of uniformity is extensive for data descriptions across different domains, systems, and processes. This unevenness can even be profound within domains, given economic factors supporting metadata generation, competing standards, or, simply, differing adoption policies. For example, two organizations may use the same metadata standard, but have different implementation practices. Even when standardization is imposed, an organization, process, and human activity can contribute to inconsistencies.
Value	*If data is the new black gold*—akin to petroleum requiring purification, but also a money maker, then metadata is the new platinum—a malleable substance that keeps its toughness, and can serve as a catalyst, sparking a reaction*. Metadata, as the *new platinum*, can be modified, while remaining a strong, independent data type. Metadata stands as a durable data object that triggers various functions—the catalyst, and achieves results—a reaction. Metadata is vital to accurate data interpretation and use by both humans and machines, and the value of metadata for data science endeavors cannot be overstated or diminished.

Note. *Singh (2013) identified data as the new black gold on Wired.com.

Table 1 draws from the commonly applied 5Vs (Marr, 2014), although other big data frameworks with nuanced or even different criteria likely apply to big metadata. Clearly, data science is not limited to big data; however, exploring the framework above is warranted inasmuch as it helps define big metadata and identify research pathways. Smart metadata, discussed in the following section, offers another fresh insight into metadata in the area of data science.

4.2 Smart Metadata

Metadata is inherently smart data because it provides context and meaning for data. One of the earliest uses of "smart metadata" was for a special session entitled "Smart Metadata" at the 2003 Dublin Core Conference, Seattle, Washington (DCMI, 2003). Themes in this special session included interoperable metadata, Semantic Web support, accessibility, and ontologies. Around the same time, van Hemel et al. (2003) promoted the idea of smart metadata in reference to the Semantic Web and the use of the Resource Description Framework (RDF) for topic maps. In 2007, Kogen, Miller, and Schobbe (2007) of the Microsoft Corporation used the term smart metadata as part of a patent description for a technique supporting metadata field management in a taxonomy system. Since that time, there does not appear to be a clear path for using the term "smart metadata" although research and discussions acknowledge metadata as a value-added factor supporting smart search, and as an enabler or characteristic of the Semantic Web and linked data (e.g. Fatima, Luca, & Wilson, 2014; Oh, Yi, & Jang, 2015). Zeng underscores this point in her work on smart data in the humanities, specifically in a recent discussion segment entitled "*How to Transform Big Data into Smart Data*?", where she identifies Semantic Web standards along with other semantic technologies (Zeng, 2017) as part of the solution.

A related aspect of smart metadata is the alignment with smart technology, including smart, mobile devices, and appliances. Examples include *mobile health technology*, such as the Fitbit, tracking heartrate, calories burned, miles that one has walked or run; *smart buildings*, using sensors to control lighting or the heating, ventilation, and air conditioning (HVAC) unit; the innovation of *smart cities*, powered by a smart grid and interlinking to the Internet of Things (IoT); or the more recently proposed phenomenon, the Internet of Everything (IoE). From smart technology to the more encompassing, smart environment, there is reliance on the collection of data, including metadata, feeding data-driven algorithms and launching intelligent, actionable processes.

Smart metadata has received attention within smart technology research. For example, Abbasi, Vassilopoulou, and Stergioulas (2017) used the phrase "smart metadata" to identify research directions and new tools supporting better use of

digital media and the larger IoT. Contractor et al. (2015) refer to smart metadata in their analysis of the Learning Content Hub, a content management system supporting automatic metadata assignment, and the use of analytics to build customized educational applications. Similarly, researchers identify smart metadata as part of their design for a personalized, recommendation engine for TV programs (Thyagaraju & Kulkarni, 2011). In all these cases, metadata is smart in that it enables an action that draws on the data being represented or tracked. The action depends on good quality metadata that is accessible, preserved over time, and trusted. These ideas translate into the principles presented in Figure 2, forming a smart metadata matrix.

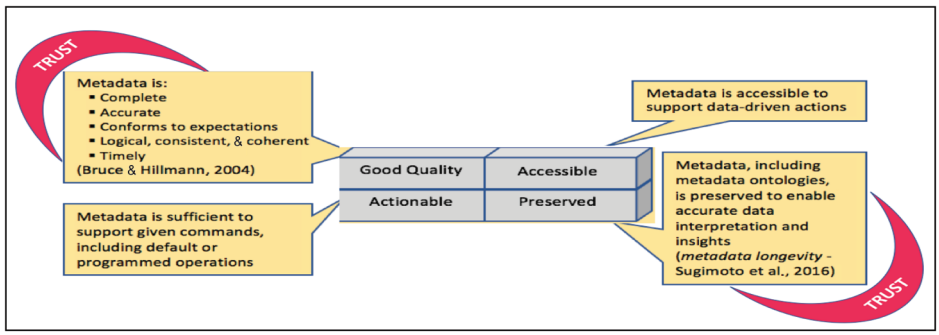


Figure 2. Smart metadata matrix of principles.

Principles of *smart data* (Figure 2) are defined as follows:

- **Good quality** Smart metadata is good quality metadata. A number of researchers have identified criteria that define good quality metadata. Bruce and Hillmann (2004) present one of most well-known criteria for determining metadata quality. Figure 2 identifies five of Bruce and Hillmann's criteria that are essential for smart metadata. Good quality metadata is also trusted metadata, and produced by a reliable source.
- **Accessible** Smart metadata is accessible, along with data being represented, to support data-driven activities. There are multiple system levels connecting metadata and access. First, metadata specifies technical requirements for accessing and using data, such as technologies needed. Second, metadata indicates the access policy, such as necessary required permissions, rights, and other protocols that enable metadata and data access and use. Third, metadata, as a "smart" asset, is accessible along with the data being represented, so that both data sources—metadata and data—can be used for data science inquiries.
- **Actionable** Smart metadata is actionable. That is, smart metadata is formatted so that it can be ingested and understood by humans and/or machines, as

required, to invoke or execute an operation. The consumable state of smart metadata needs to also be reflected in the data being represented.

- **Preserved** Smart metadata is preserved in a useful manner. This step is critical for identifying data patterns over time. Big data is volatile and metadata is often modified, enriched, or deleted in sync with change. Interpreting data change over time is difficult or even impossible when previously generated metadata is absent. Metadata must be preserved by a trusted, dependable source; this includes the preservation metadata vocabularies, such as data dictionaries and attribute descriptions. Research on metadata longevity (Li & Sugimoto, 2017; Sugimoto et al., 2016) has resulted in a framework solution for preserving metadata. Additionally, Kunze et al. (2016a; 2016b) present complementary work developing a persistence vocabulary. These are significant initiatives that can help further formalize our understanding of preservation as a principle for smart metadata.
- **Trust** The last smart metadata principle, trust, connects across all principles, although it primarily links with quality and preservation. As noted above, good quality metadata is trusted metadata, and produced by a reliable source; and metadata that is preserved must be overseen and maintained by a dependable entity.

The smart metadata principles defined here qualify metadata as value-added data. The next section of this paper explores value relating to metadata more thoroughly through the concept metadata capital.

4.3 Metadata Capital

"Metadata capital" is a concept that emerged through research on data and metadata reuse in the Dryad data repository (Greenberg, Swauger, & Feinstein, 2013). Capital, broadly speaking, is understood as an asset with value; and the value may be financial, intellectual, social, or defined in other ways. Capital is most commonly associated with finance and wealth, and draws from work such as Adam Smith's *An Inquiry into the Nature and Causes of the Wealth of Nations*, published in 1776 (Smith, 1776). Smith's (1776) emphasis is on market value, or financial aspects. The financial component of capital has been explored more specifically through the Metadata Capital Initiative. This research was predicated on the fact that value, as a financial indicator, can be measured. The incipient effort has chiefly applied the modified capital gains equation (Greenberg, Swauger, & Feinstein, 2013; Greenberg et al., 2014a; 2014b), and calculated costs associated with metadata creation and reuse to determine value. Specifically, metadata reuse demonstrates a greater return on investment (ROI) by adding value to the initial metadata cost.

It is important to point out that cost and value are not always aligned; this is because a product can cost more than it is worth, or be assigned a price that is below its worth. Even so, financial cost can be calculated. The metadata capital work postulates that when a purchased item is reused, over time, it is worth more than its original cost. Analogies to consider include a top-end stainless steel pot that is used over and over, without any change, and always supporting cooking to perfection; or an antique chest that has been passed down generation after generation, and is used to store sweaters in the summer, while also serving as a piece of furniture, becoming more valuable with age.

As stated above, capital, wealth, and value do not solely apply to financial matters, despite the fact that much of the big data and data science coverage is associated with economic incentive and opportunity. The broader interpretation of capital extends to knowledge (intellectual capital), and friendships—personal and professional relationships (social capital), as well as other areas, including some still likely to be discovered. Drawing on this broader context of value, a formalized definition of "metadata capital" is as follows, which was originally published in the *Bulletin of the Association for Information Science and Technology* (Greenberg, 2014).

1. An asset that contains contextual knowledge about content.
   a. Content is the data or information contained in any information object (any "entity, form, or mode").
   b. Context is who, what, where, when, how, why, etc., which can be captured via metadata attributes (Kunze, 2001).
2. A product or service generated by human labor and/or machine-driven processes with value that increases over time or that enables the value increase of other assets.
3. A good (a service facilitator) supporting a range of functions such as discovery, provenance tracking, rights management, authentication, preservation and other functions associated with lifecycle management and access.
4. A public good if the product (metadata) is open, following which the services can be open.

Metadata capital is defined as an asset, a product, a good, including a public good, which enables gain through knowledge, access, and services. Metadata capital connecting to this broader interpretation associates with the promise of big data when considering the unprecedented opportunity to address real world problems in energy, health, and the environment (Greenberg & Garoufallou, 2013). Metadata is essential for using data to compare new energy approaches; track the progression and decline of a health crisis, such as the Ebola virus; or study climate change.

The biggest challenge with metadata valuation in this broader spectrum, and even with financial aspects, is the formidable task of substantiating value. In pursuing metadata capital as a financial topic, costs can be identified, or at least estimated, by adding system expenses, staff salaries distributed by hours dedicated to metadata tasks, and other incurred costs. However, determining exactly where to begin measuring cost is not an easy task. Does cost start with the metadata system design, the salary of the person who had the idea to build the metadata system, the person or team that implemented workflow design, or the cost of the code library that allowed the system to be built? Assessing social and intellectual value is even more daunting. How can we determine long-term consequences for metadata created today that allows for a major health discovery five or ten years from now?

There are more questions than answers in pinpointing or even approximating the value of metadata; it is predicament that underscores a significant challenge and invites research. Metadata capital requires further study, including drawing upon valuation and appraisal frameworks from other disciplines. What frameworks exist for measuring value across the domains of energy, health, and the environment? How do people assess the value of knowledge, personal friendships, and professional contacts? Although there is no single answer, drawing upon valuation research from other domains can help chart metadata research directions, and, future, demonstrate the value of metadata entrenched in data science.

5 Summary and Conclusions

Metadata, while applauded by many, has not been vigorously pursued as a research topic in data science, compared to statistical modeling, algorithm testing, data mining, and visualization. To be clear, there is metadata research; however, metadata focused scientific and scholarly output in data science venues has not kept pace with these other topics. Articulating a problem is one of the first steps to addressing a challenge. This paper pursued initial steps to addressing this challenge in the following ways:

1) **The "utilitarian nature" and "historical and traditional views" of metadata were identified as two intersecting factors that have inhibited metadata research.** Having a clear understanding of barriers is important for addressing existing challenges. The pragmatic nature of metadata is paramount, and applied research ought to be shared more, rather than minimized. Additionally, fundamental approaches can interconnect with applied work, as research matures. As for traditional views, there is always an opportunity for change. The cultural shift taking place in data sharing is evidence of change.

Sharing metadata research impacts in data science can attract more interest and support. Although this paper identified two key factors, there are very likely other intersecting factors to consider in future work.

2) **Contextual definitions were presented for both "metadata" and "data science" to help further dialogue and research on metadata in the data science domain.** As noted above, articulating a problem is a first step to addressing a challenge. A second step is to define the context. Both metadata and data science were defined as part of this goal. The definitions given draw from other published work, synthesizing common themes and ideas. Given the work pursued in this paper, defining these two terms was an obvious choice, although it is likely that providing contextual definitions for additional terms will aid future research.
3) ***Big metadata*, *smart metadata*, and *metadata capital* were presented as part of a metadata *lingua franca* to help frame research in data science research space.** These concepts are not commonly discussed in data science, although they appear in research, and the examination of these concepts integrates original work in this paper, along with ideas and outcomes from other scholarship to provide grounding. Admittedly, the concepts introduced may warrant refinement; and there are other significant metadata concepts that also deserve focus. Even so, the presentation of these terms together can offer support and provide a pathway for metadata research within data science.

The immediate contribution of this work is, simply, that it may elicit response, critique, or revision. A more impactful contribution is that this work may motivate more researchers to consider the significance of metadata as a topic worthy of research within data science and the larger digital ecosystem. In a recent discussion, my colleague at Drexel University, Dr. Rosina Weber, asked me, "Can you imagine data science without metadata?" I cannot think of a statement more profound than this to motivate next steps. This question needs to be considered by anyone who applauds or dismisses the value of metadata.

Data science cannot progress without metadata research; and while an extensive range of metadata topics are important, researchers need to ask: *which metadata topics are most pressing to pursue*? In other words, let's prioritize metadata research so that data science can successfully address our most significant societal challenges, and more fully contribute to the greater good. In conclusion, the framework presented in this paper, defining big metadata, smart metadata, and metadata capital, can help researchers, across multiple disciplines, prioritize next steps and collectively advance metadata research in data science.

References

Abbasi, M., Vassilopoulou, P., & Stergioulas, L. (2017). Technology roadmap for the creative industries. Creative Industries Journal, 10(1), 40–58.

Beall, J. (2004). Dublin Core: An obituary. Library Hi Tech News, 21(8), 40–41.

Beall, J. (2014). Dublin Core is still dead. Library Hi Tech News, 31(9), 11–13.

Bruce, T.R., & Hillmann, D.I. (2004). The continuum of metadata quality: Defining, expressing, exploiting. ALA Editions. Retrieved on July 31, 2017, from http://ecommons.cornell.edu/handle/1813/7895.

Coleman, A.S. (2005). From cataloging to metadata: Dublin Core records for the library catalog. Cataloging & Classification Quarterly, 40(3–4), 153–181.

Contractor, D., Negi, S., Popat, K., Ikbal, S., Prasad, B., Kakaraparthy, S., Sengupta, B., Vedula, S., & Kumar, V. (2015). Smarter learning content management using the Learning Content Hub. IBM Journal of Research and Development, 59(6), 3:1–3:9.

Data Science Association (DSA). (2017). About data science. Retrieved on June 18, 2017, from http://www.datascienceassn.org/about-data-science.

DCMI. (2003). Special session: Smart metadata. In 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice-Metadata Research & Applications, Seattle, Washington. Retrieved on June 30, 2017, from http://dublincore.org/workshops/dc-2003/smartDC.html.

Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64.

Dimitrova, N. (October–December, 2004). Is it time for a moratorium on metadata? IEEE Multimedia, 11(4), 10–17.

Doctorow, C. (2001). Metacrap: Putting the torch to seven straw-men of the meta-utopia. Retrieved on June 28, 2017, from http://chnm.gmu.edu/digitalhistory/links/pdf/preserving/8_17.pdf.

Dong, R., Su, F., Yang, S., Xu, L., Cheng, X., & Chen, W. (2016, September). Design and application on metadata management for information supply chain. In the 16th International Symposium on Communications and Information Technologies (ISCIT) (pp. 393–396). Washington, DC: IEEE Computer Society Press.

ERAC Secretariat. (2016). European Research Area and Innovation Committee. European Union. Brussels, February 3, 2016. Retrieved on June 18, 2017, from http://data.consilium.europa.eu/doc/document/ST-1202-2016-INIT/en/pdf.

Fatima, A., Luca, C., & Wilson, G. (2014, March). New framework for semantic search engine. In 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation (UKSim) (pp. 446–451). Washington, DC: IEEE Computer Society Press.

Few, S. (2017). Visual business intelligence: A blog by Stephen Few. There is no science of data, January 23, 2017. Retrieved on July 7, 2017, from https://www.perceptualedge.com/blog/?p=2560.

Gaitanou, P., Gergatsoulis, M., Spanoudakis, D., Bountouri, L., & Papatheodorou, C. (2016). Mapping the hierarchy of EAD to VRA Core 4.0 through CIDOC CRM. In the 10th International Conference on Metadata and Semantics Research (MTSR 2016) (pp. 193–204). Cham, Switzerland: Springer International Publishing.

Greenberg, J. (2005). Understanding metadata and metadata schemes. Cataloging & Classification Quarterly, 40(3–4), 17–36.

Greenberg, J. (2009). Theoretical considerations of lifecycle modeling: An analysis of the dryad repository demonstrating automatic metadata propagation, inheritance, and value system adoption. Cataloging & Classification Quarterly, 47(3–4), 380–402.

Greenberg, J. (2009). Metadata and digital information. In M.J. Bates & M.N. Maack (Eds.), Encyclopedia of Library and Information Sciences (pp. 3610–3623). Boca Raton, FL: CRC Press.

Greenberg, J. (2014). Metadata capital: Raising awareness, exploring a new concept. Bulletin of the Association for Information Science and Technology, 40(4), 30–33.

Greenberg, J., & Garoufallou, E. (2013). Change and a future for metadata. In MTSR-2013: Proceedings of the 7th Metadata and Semantics Research Conference (pp. 1–5). Cham, Switzerland: Springer International Publishing.

Greenberg, J., Murillo, A.P., Ogletree, A., Boyles, R., Martin, N., & Romeo, C. (2014a). Metadata capital: Automating metadata workflows in the NIEHS Viral Vector Core Laboratory. In MTSR-2014: Proceedings of the 8th Metadata and Semantics Research Conference (pp. 1–13). Cham, Switzerland: Springer International Publishing.

Greenberg, J., Ogletree, A., Murillo, A.P., Caruso, T.P., & Huang, H. (2014b). Metadata capital: Simulating the predictive value of self-generated health information (SGHI). In 2014 IEEE International Conference on Big Data (pp. 31–36). Washington, DC: IEEE Computer Society Press.

Greenberg, J., Swauger, S., & Feinstein, E.M. (2013). Metadata capital in a data repository. In DC-2013: the International Conference on Dublin Core and Metadata Applications (pp. 140–150). Lisbon, Portugal: Dublin Core metadata initiative.

Greenwald, G. (2013). Edward Snowden: The whistleblower behind the NSA surveillance revelations. The Guardian. Retrieved on June 18, 2017, from https://www.theguardian.com/world/2013/jun/09/edward-snowden-nsa-whistleblower-surveillance.

Hey, T., Tansley, S., & Tolle, K. (2009). The fourth paradigm. Redmond, Washington: Microsoft Research.

Ilevbare, I., Athanassopoulou, I., & Wooldridge, J. (2017). UK Workshop on Data Metrology and Standards. The National Physical Laboratory and partners at the University of Huddersfield and University of Cambridge. March, 2017. Retrieved on June 18, 2017, from http://www.bigdata.cam.ac.uk/files/npl-industry-workshop-on-data-metrology-standards/npl-industry-workshop-on-data-metrology-standards-report.

Kogan, D.E., Miller, P.C., & Schobbe, G.A. (2007). Techniques to manage metadata fields for a taxonomy system. US 20080301096 A1. (Also published as WO2008150619A1). Retrieved on June 28, 2017, from http://www.freepatentsonline.com/y2008/0301096.html.

Kunze, J. (2001). A metadata kernel for electronic permanence. In International Conference on Dublin Core and Metadata Applications, North America, DC2001. Retrieved on July 31, 2017, from http://dcpapers.dublincore.org/pubs/article/view/656.

Kunze, J., Calvert, S., DeBarry, J., Hanlon, M., Janée, G., & Sweat, S. (2016a). Persistence statements: Describing digital stickiness. California Digital Library. Retrieved on July 20, 2017, from http://escholarship.org/uc/item/2zm9x47c.

Kunze, J., DeBarry, J., Hanlon, M., Scout, C., & Sweat, S. (2016b) A vocabulary for persistence. In SciDataCon 2016. September 11–13, 2016, Denver Colorado. Retrieved on July 21, 2017, from http://www.scidatacon.org/2016/sessions/103/paper/109/.

Li, C., & Sugimoto, S. (2017). Provenance description of metadata vocabularies for the long-term maintenance of metadata. Journal of Data and Information Science, 2(2), 41–55.

Lytras, M.D., Sicilia, M.Á., & Cechinel, C. (2013). The value and cost of metadata (chapter I. 3). In M.A. Sicilia (Ed.), Handbook of Metadata, Semantics and Ontologies (pp. 41–62). Hackensack, N.J., World Scientific Publishing Company.

Manian, D. (2011, Nov. 11). Our pointless pursuit of semantic value. Retrieved on June 29, 2017, from https://www.smashingmagazine.com/2011/11/our-pointless-pursuit-of-semantic-value/.

Marr, B. (2014). Big data: The 5 Vs everyone must know. LinkedIn: Big data. Retrieved on June 18, 2017, from https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know.

Méndez, E., & van Hooland, S. (2013). Metadata typology and metadata uses (chapter I.2). In M.A. Sicilia (Ed.), Handbook of Metadata, Semantics and Ontologies (pp. 9–40). Hackensack, N.J., World Scientific Publishing Company.

NITRD. (2016). The Federal Big Data Research and Development Strategic Plan. The Networking and Information Technology Research and Development Program, May 2016. Retrieved on June 15, 2017, from https://www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf.

Oh, S.G., Yi, M., & Jang, W. (2015). Deploying linked open vocabulary (LOV) to enhance library linked data. Journal of Information Science Theory and Practice, 2(2), 6–15.

Riley, J. (2017). Understanding metadata. Bethesda, MD: NISO Press.

Shankaranarayanan, G., & Even, A. (2006). The metadata enigma. Communications of the ACM, 49(2), 88–94.

Shirky, C. (2005). Ontology is overrated: Categories, links, and tags. Economics & Culture, Media & Community. Retrieved on June 20, 2017, from http://www.shirky.com/writings/ontology_overrated.html?goback=.gde_1838701_member_179729766.

Simon, P. (2013). Too big to ignore: The business case for big data (Vol. 72). Hoboken, NJ: John Wiley & Sons.

Singh, A. (2013). Is big data the new black gold? Wired. Retrieved on July 7, 2017, from http://www.wired.com/2013/02/is-big-data-the-new-black-gold.

Smith, A. (1776). An inquiry into the nature and causes of the wealth of nations. London: W. Strahan and T. Cadell.

Smith, K., Seligman, L., Rosenthal, A., Kurcz, C., Greer, M., Macheret, C., . . . & Eckstein, A. (2014). Big metadata: The need for principled metadata management in big data ecosystems. In Proceedings of Workshop on Data Analytics in the Cloud (pp. 1–4). New York: ACM.

Stanton, J.M. (2012). Introduction to data science. Syracuse University. Retrieved on June 6, 2017, from https://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf.

Sugimoto, S., Li, C., Nagamori, M., & Greenberg, J. (2016). Permanence and temporal interoperability of metadata in the linked open data environment. In Proceedings of the International Conference on Dublin Core and Metadata Applications 2016 (pp. 45–54). Retrieved on June 28, 2017, from http://dcevents.dublincore.org/IntConf/dc-2016/paper/view/430.

Tennant, R. (2002). MARC must die. Library Journal, 127(17), 26–27.

Thyagaraju, G.S., & Kulkarni, U.P. (2011). Family aware TV program and settings recommender. International Journal of Computer Applications, 29(4), 1–18.

UK Data Archive. (2012). Research data lifecycle. Retrieved on June 15, 2017, from http://www.data-archive.ac.uk/create-manage/life-cycle.

Vaduva, A., & Dittrich, K.R. (2001). Metadata management for data warehousing: Between vision and reality. In 2001 International Symposium on Database Engineering and Applications (pp. 129–135). Washington, DC: IEEE Computer Society Press.

van der Aalst, W. (2016). Process mining: Data science in action. Berlin: Springer-Heidelberg.

van Hemel, S., Paepen, B., & Engelen, J. (2003). Smart search in newspaper archives using topic maps. In Proceedings of the 7th ICCC/IFIP International Conference on Electronic Publishing. Retrieved on June 29, 2017, from http://elpub.scix.net/data/works/att/0333.content.pdf.

Vlachidis, A., Binding, C., May, K., & Tudhope, D. (2013). Automatic metadata generation in an archaeological digital library: Semantic annotation of grey literature. In Computational Linguistics (pp. 187–202). Berlin: Springer-Heidelberg.

White, H., Willis, C., & Greenberg, J. (2014). HIVEing: The effect of a semantic web technology on inter-indexer consistency. Journal of Documentation, 70(3), 307–329.

Zavalina, O.L. (2011, September). Free-text collection-level subject metadata in large-scale digital libraries: A comparative content analysis. In International Conference on Dublin Core and Metadata Applications (pp. 147–157). Retrieved on June 18, 2017, from http://dcevents.dublincore.org/IntConf/dc-2011/paper/view/50/19.

Zeng, M.L. (2017). Smart data for digital humanities. Journal of Data and Information Science, 2(1), 1–12.

Zeng, M.L., & Qin, J. (2016). Metadata. New York: Neal-Schuman Publishers, Inc.

Zhao, X., Ma, H., Zhang, H., Tang, Y., & Fu, G. (2014, October). Metadata extraction and correction for large-scale traffic surveillance videos. 2014 IEEE International Conference on Big Data (Big Data) (pp. 412–420). Washington, DC: IEEE Computer Society Press.