Title	For Want of a Nail: Three Tropes in Digital Curation
Authors(s)	Shankar, Kalpana
Publication date	2015-12
Publication information	Shankar, Kalpana. "For Want of a Nail: Three Tropes in Digital Curation" 44, no. 4 (December, 2015).
Publisher	De Gruyter
Item record/more information	http://hdl.handle.net/10197/7272
Publisher's version (DOI)	10.1515/pdtc-2015-0019

Kalpana Shankar*

For Want of a Nail: Three Tropes in Data Curation

Abstract: This article explores the role of three key tropes in the data curation profession. Using interviews with digital preservation experts, researchers, public sector statisticians, and social science data archivists as well as popular and professional literature and media, this article discusses how tropes and narratives are used to create shared meaning among data curation stakeholders. The article explores how tropes of abundance / overload, openness, and trust are created and used and concludes with reflections on how such stories articulate professional values and concerns. The article advocates for further attention to the use of narratives and stories as the data curation profession develops.

Keywords: Data Deluge, Digital Curation, Narratives, Open Data, Research Practices

1 Introduction

Data curation as a profession represents an entanglement of technical, social, and organizational practices. As an activity, it stands at the nexus of complicated and interconnected conversations about cultural memory, institutional sustainability, and organizational transparency and accountability. Pennock defines data curation thus: "Data curation, broadly interpreted, is about maintaining and adding value to a trusted body of digital information for both current and future use: in other words, it is the active management and appraisal of digital information over its entire life cycle" (Pennock 34). Curation encompasses a pragmatic and practical set of concerns with all aspects of the life cycle of digital objects, as well as the engagements of institutional and organizational structures required to manage digital objects curated over time. In many cases, conversations, conducted via professional and scholarly publications, formal and informal workshops and classes, white papers, conferences, online forums, and social media, emphasize the technologies needed for acquiring, preserving, describing, and disseminating the digital products: conversations about what could be described as the instrumentality of data curation. Understandably, most data curation writings

focus on this necessary and practical work: the standards, tools, instruments, and other elements of process (Willis, Greenberg, and White; Wilson et al). A second set of conversations focus on institutional support, collaborations, consortia, and intellectual property challenges—conversations needed to make data curation a sustainable and successful endeavor (Heidorn; Erwin and Sweetkind-Singer, Shankar et al). A third set of conversations is often about the stakeholders—reaching out to user communities, legal and organizational mandates, the training of professionals and content creators, costs, and making the case for data curation to increasingly strapped academic institutions (Gabridge; Beagrie, Chruszcz, and Lavoie).

What brings these groups together and helps them bridge their disciplinary divides are stories. Psychologist Jerome Bruner, a pioneer in the fields of cognitive and educational psychology and the power of narratives and stories, asserts that the stories people tell are particularly effective instruments for what he terms "social negotiation" and transcending or bridging different perspectives, needs, backgrounds, and skills (2003). Narrative analysis has been used with great effect to understand how information systems are designed and experienced, how people within the organization make sense of the changes they experience, and share knowledge about information systems. One of the most well-known examples of this work is that of anthropologist Julian Orr (1996) who studied the knowledge-sharing practices of Xerox copy machine repair workers. During his ethnographic studies, he found that over morning coffee in favorite shops, repair workers shared stories of particularly truculent copiers, their successes and failures in debugging obscure error codes, the finicky ways of particular models, and other bits of knowledge that are difficult to translate into formal knowledge management but essential to the day-to-day work of repair staff. Studies of stories, conducted over the decades in other workplace venues and contexts, are also used to understand how interdisciplinary groups come to shared understandings in project and teamwork and collaboration (Linde; Loseke).

Structurally, narrative analysts position stories as "bounded slices of communication that present themselves as coherent wholes and can be isolated for analysis" (Fenton and Langley 1175). Stories may vary in scope: big stories, often called "master narratives" or "grand narratives," may span space and time to encompass whole communities of practitioners. Small or local stories may be limited to

*Corresponding author: Kalpana Shankar,
e-mail: kalpana.shankar@ucd.ie

specific organizations or work groups (Callon and Law; Doolin). Stories are fruitful units of analysis because they reflect and structure peoples' understandings of what they are doing and who they are (Squire; Labov and Waletzky). Examining narratives is a way to explore many dimensions of professional and organizational ethos and practice. The study of narratives suggests a focus on how the sequence of events, the characters and plot, and tensions are expressed and how those elements become embedded in the technical artifacts of work.

Kunda and Anderson-Wilk (2011) make a convincing case for the digital curation profession to position itself as storytellers. They note that people find stories more powerful than raw data or observations. "Community stories," they argue, are used in museum exhibit design, interface design, and other venues where stories are used to deliver knowledge. Like those pursuing such projects, the multiple communities engaged in digital curation do not necessarily share a working language at the outset of a project but must come to some common understandings. Those who are managing and curating data are usually information professionals using the language of standards, frameworks, metadata, and statistics, but are tasked with making these topics comprehensible and engaging to data creators in different disciplines. In turn, the researchers and other data creators who are being exhorted to maintain and deposit data are domain specialists in academic disciplines, civil service, and professional arenas (Zimmerman). These individuals and groups create data in the course of their activities, are usually aware that there is some need for maintaining their data over time, but may lack the time, training, and resources to do so effectively. Institutional administrators, granting agencies, and other funders may know something or nothing of all of these topics but are often tasked with marketing, institutional sustainability, and repository or research agency policy development and implementation (Witt).

This paper explores one way in which meaning is negotiated among data curation stakeholders—through the tropes of abundance, openness, and trust. Tropes (for the purposes of this paper) are recurrent metaphors that are widely recognizable even if the specific instances are new. These tropes are often about idealized views of what *ought* to be but also what *is* and represent structures through which humans organize and share their worldviews (Morgan). The stories or narratives that are used to flesh out these tropes allow people to express aspirations, moral imperatives, anxieties, identities—and ultimately, possibilities for change. They "make concrete sets of connections and practices in the everyday world" (May and Fleming 1094). When we pay attention to tropes, we also pay attention to the recurring anecdotes and stories individuals tell to make sense of their lives. For data curation, when actors are grappling with institutional changes, new organizational mandates, or their own evolving work practices, narratives are often invoked to cross-disciplinary boundaries in order to establish and justify professional identity and practices. Some of the most frequently repeated stories and narratives, the ones that become truisms, embed within them the assumptions that people make about themselves and each other (Dawson and McLean). This paper will explore three data curation tropes in interrelated ways: as embedded narratives (visual and oral), their roles in constructing and maintaining data curation as practice, and as devices by which different stakeholders engage each other. The paper will conclude with some reflections on paying attention to how narratives become embedded in the standards, practices, and technologies that comprise data creation and curation.

2 Research Design

The empirical data for this paper comes from two sources. This first is an ongoing study of the historical and contemporary aspects of institutional and data sustainability in some of the most long-lived examples of data archives: Social Science Data Archives (SSDA). Social science data archives predate computers and the Internet, and provide a unique opportunity to examine perceptions about what makes an archive sustainable over long periods of time. The social sciences have enjoyed stable and successful data archives since the 1940s (Green and Gutmann). Corti (2012) suggests that SSDA are the foundation of contemporary efforts in open data, data archiving, and data curation.

Ten semi-structured interviews, each lasting approximately 45 minutes, were conducted in winter and spring 2014 at two SSDA with current and former staff, researchers who have participated in governance, institutional managers, major depositors, and repository users. Depending upon their role, respondents were asked about their professional responsibilities, their concerns and challenges for maintaining social science data, the development and use of technologies and standards, and changes in institutional needs over time. In addition, the researchers gathered governance meeting minutes, annual reports, strategic plans and grant proposals, and other institutional documents. Data was coded inductively for emergent themes but also coded deductively to focus on sustainability, particularly

with respect to business models, economics, standards development, changes in users / practices over time.

The second project involves the development of a new environmental sustainability research institute on the author's campus. There are more than eighty principal investigators from engineering, the social sciences, and natural sciences, as well as IT and other support staff; fifteen interviews based on the Research Data Management Plan checklist developed at the United Kingdom Data Curation Centre (http://www.dcc.ac.uk/resources/policy-and-legal) were conducted in the spring and summer of 2012. The author was brought in to interview researchers about their current data practices, anticipated changes over time, and concerns about data. The interviews were coded for these themes, but also for specific answers to questions raised by the Research Data Management checklists. The transcripts were inductively coded for concerns specific to the new institute: buy-in for developing a data archive for the institute, researchers' attitudes and needs with respect to open data, acquiring funding for data archiving, and for emergent themes (Emerson, Fretz, and Shaw) of research interest to the author.

Although there are other themes that could be explored, the three tropes in this paper appeared in some version in all of the interviews regardless of the role of the interviewee—curator, data creator, or other data curation stakeholder. Second, the particular tropes discussed are the most evocative of the sociotechnical complexity of data curation and appear in mass media as well as professional literature about data curation (which provides the third source of "data" for this paper, although no attempt was made to conduct a systematic review or content analysis). Lastly, they all share narrative structures—all of them contain "data stories" about the creation of data, its growth, its governance, and the need for data curation in some fashion or other.

3 Managing Abundance: Making the Case for Data Curation

The need for data curation begins with abundance and overabundance. Simply put, a need for data curation exists because digital data exists in bewildering quantity with increasing rates of growth (Borgman; Meyer and Schroeder). Institutions and individuals by extension are increasingly vested in its effective management, use, and reuse. A brief and extremely unscientific search for the term "data deluge" in Google Scholar reveals

approximately 4000 instances of the term in the titles or text of professional and scholarly articles, primarily from the natural sciences, computer science, and information studies (with several hundred instances of "data curation" and "data deluge" occurring together). A similarly cursory and unscientific search on the main Google search engine increases the instances of the term to over 390,000, with the top article in—of all places—*The Economist*. In that article, one reads: "[T]he world contains an unimaginably vast amount of digital information which is getting ever vaster ever more rapidly. This makes it possible to do many things that previously could not be done: spot business trends, prevent diseases, combat crime, and so on. Managed well, the data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments to account" (*The Economist*). The article goes on to cite examples from fields as diverse as big-box store retail and astronomy to illustrate that the Data Deluge is clearly Big Business. The metaphorical deluge is illustrated thus: "A male figure [is] holding an inverted umbrella toward a sky from which streams of binary data rain down. With his umbrella, the man catches some of the streams, using the distillate to water a flower by his side. Other zeros and ones rain down without bothering the man or his plant" (Coopmans, 38). The World Economic Forum, also no Luddite when spying new opportunities for economic growth, has referred to personal data as "the new oil" that will demand new ways of thinking about people (World Economic Forum).

That data is growing (and at an accelerated pace) is repeated so often it has become foundational—an origin story for the data curation community, but one whose significance goes well beyond its immediate practitioners. The "data deluge" as metaphor has much to recommend it. It is onomatopoeic and alliterative (thus easily remembered and repeated), and easily visualized as invoked because it evokes the myths of "information overload" (Tidline) that most dwellers in ICT-instrumented societies are intimately familiar with. "Data deluge" conjures up a mythic, Noah's Flood-inspiring sense of scale (Bassi, Denazis, and Giacomin). This scale is often greeted with awe. Nevertheless, the bombastic rhetoric, particularly when the data deluge term first became more commonly used in the literature, is more reflective of profound anxieties about the nature of this new technologically driven and deterministic production of research data. Data does not "behave" as other commodities do. With other commodities, scarcity shapes value and use. However, data behaves in a maddeningly opposite fashion—its abundance must be tamed if it is to be useful. The mere quantity generates concerns because there is so much of

it. In spite of the abundance, this arena is no level playing field. Johns (2013) writes that the "data-savvy" and those with "deep analytical talent" (and the education to enable its realization) may leave the rest of us behind.

Those who are responsible for creating the data and those called upon to manage and curate it invoke this urgency of abundance. "Good narratives" have the ring of truth and data floods seem evident on the face of lived experiences of researchers and curators, and even the public. Researchers may want to be seen as being in fields that have data abundance; Sawyer (2008) posits that data wealth and poverty in research will increasingly become the demarcator of research impact. It is not surprising that information professionals and libraries look to these data floods and deluges both with anticipation and dismay. Librarians and allied information professionals have developed and implemented tools for over one-hundred years to harness and manage the objects in their purview. They have provided intellectual and physical access through robust organizational structures, all kinds of economic downturns, new user communities, and new information resources. Stories of success—that libraries and archives know how to manage abundance—are potentially useful in enlisting allies in their ongoing work. The director of one Social Science Data Archives (SSDA) noted that there are interrelated needs for research data: journals, funding bodies, and other institutions that are requiring data deposit, and researchers themselves who want to increase citations to their work. He went on to tell a "success story" of a university researcher who experimented with data deposit in the SSDA—even though the researcher was not a social scientist, the SSDA provided an effective infrastructure. Her citation count increased and the visibility of her data was cited as a positive in a successful grant application.

This new world of research data, however, does not necessarily behave in the same way as the other resources librarians have learned to manage. While underlying professional principles may remain the same, the library community has become used to a world in which the objects it manages came to them through acquisition, licensing, gifts, and similar mechanisms. Research data requires new kinds of metadata, preservation mechanisms, access controls, and potential privacy / confidentiality concerns. Managing research data also require that libraries work with new communities to acquire data in the first place; many of the researchers had never considered their own institutional libraries as appropriate repositories for their research data (Wallis and Borgman).

Information professionals are clearly not alone with these anxieties; the researchers who generate the "data deluge" have their own anxieties and concerns about their data. Some express anxieties about the roles and abilities of librarians and libraries to manage data and provide effective, ethical, access (both physical and intellectual) to data. One architecture researcher discussed the relatively short-lived institute in which he was a member and wondered aloud what would happen to the data afterwards. He said, "I was a member of a now-defunct institute and the data there is gone. We didn't know who to talk to about making the data available." Jahnke, Asher, and Keralis (2012) describe this gap between the librarians increasingly tasked with data management and the researchers' perception of the libraries as "a dispensary of goods ... rather than a locus for real-time research / professional support" (p. 4). They argue that this gap effectively compromises the ability of those in the library and information science field to collaborate with those same researchers. Kouper (2013), in a reflection on her experience as a Council on Library and Information Resources (CLIR) Fellow in Data Curation, describes the data curator as a hybrid professional—a researcher, librarian, technology adopter, and policymaker. However, conveying these multiple roles to the researchers requires convincing them that libraries are more than passive dispensaries.

Anxiety, expressed through stories of the deluge, is not the only response. Data abundance is data opportunity. *The Economist* article discussed above and other mass media publications convey a breathless sense of excitement about the possibilities for new scientific research possibilities, government accountability through open government data initiatives and, of course, new economic growth potential for businesses to exploit data. The word "revolution" is similarly invoked in conjunction with the "data deluge" to represent the new potential and paradigmatic possibilities of so much data. The former editor of the magazine *Wired*, Chris Anderson, even wrote an oft-cited—often-criticized as deterministic—article in 2008 entitled "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," in which he argued that the plethora of digital data and the appropriate tools would render the creation of hypotheses from previous knowledge, the painstaking testing of them, and arriving at new conclusions, a relic of the past because with this new data-driven science, "simply" mining data for extent patterns would be sufficient to generate new insights (and of course, create a market for new kinds of knowledge and data scientists). The researchers are understandably not immune to the anxieties generated by the deluge. Consider this quote from *The Economist* article cited above: "Alex Szalay, an astrophysicist at Johns Hopkins University, notes that the proliferation of data is making them increasingly inaccessible. 'How to make sense of all these data? People should be worried about how we train the next

generation, not just of scientists, but people in government and industry,' he says" (*Economist* 2010).

The sustainability center's principal investigators expressed similar concerns, about the numerous kinds of skills needed to think about their own data (which they do not consider at the core of their profession. One scientist, when asked how she worked with data and private companies, explained that she had concerns about a new project with a private sector company that surfaced numerous (and at the time unresolved) tensions about data ownership. She went on to say, "[Research ownership] becomes a much grayer area in one of our new projects that we're just starting because we're agreeing to generate it together, and actually now we're about to start going through negotiation in advance about what's the rules. That's going to be very interesting because before that's been very clear for me ... but that changes now."

The trope of data deluge provides strong motivation for stakeholders to think and argue seriously for the professional curation of data. It is hard to argue against the data deluge as an organizing principle; the lived experience of digital overload seems quite real. However, data abundance raises concerns for those who are creating and managing it, and who must argue for it (and work together) to make curation happen. Until relatively recently in many fields, there was little need to worry about it. However, the complexity and volume of data generated in many fields and the anxieties engendered provide impetus for the professionalization of data management (where the data curators step in, and work with the researchers to manage abundance and convey the need to do so to those funding their work). However, data abundance also introduces actors with different motivations—government agencies concerned about the availability of publicly funded research, professional and scholarly publications and societies concerned about transparency and accountability, and private sector companies eager to capitalize on data as a commodity continue.

4 Managing Openness: Open Data, Open Potential, or Open for Disaster?

Openness—creating free access to data for transparency, accountability, and reuse—is invoked as an essential part of the data curation process and profession, even though the process itself it never "free" (that is, without cost). The trope of openness argues for making good data widely available, and unlocks the potential development of new applications and tools by both the private and public sectors for economic development, social justice, health, and other areas. The Open Data Foundation, the Research Data Alliance, and even high profile individuals such as Sir Tim Berners-Lee have to some extent pushed for the open release of data. As a well-understood trope, open data is hard to argue with as an emancipatory good. As a set of narratives, however, there are cautionary tales, horror stories, and far fewer tales of success.

In fact, as much as the overabundance of data that needs looking after provides a motivation for the data curation profession, openness seems to complicate that motivation. The narratives are conflicted ones. Confusion over the definitions of openness and its ramifications exist in a world where some stakeholders (particularly on the subscription-based repository model side) fear a "Wild West of open data" (as one SSDA staff member put it) where Excel spreadsheets and text files, devoid of metadata or interfaces, proliferate, making the work of the data curation professional null and void. However, some SSDA staff believed that open data does not mean uncurated data and with the right approach, openness could mean more usefulness for archives. Researchers, too, express conflicting perspectives. Some advocate for the availability of more data, but with no clear idea of how it will be made available and who will make it so. Others express concern that they themselves will be called upon to make those spreadsheets open without the time or resources to do so (and open themselves and their research up for abuse). The ethos of open data is universal—again, this trope is universally recognized and invoked, but its implementations are deeply entangled with professional anxieties and possibilities.

More specifically, the push for open research data generated significant concerns among SSDA staff about the future role of data curatorial services. During interviews at SSDA, several staff expressed concern that even as more data was being generated, their own role as institutions would be diminished or even eliminated in the world of open research data. Several people expressed the worry that the "value-add" provided by archives would somehow be diminished or even eliminated if researchers just put their data sets on Websites or servers without any "clean-up" and researchers who were interested in secondary data would use the data as is.

However, when researchers were asked if open data created concerns for them, they were vehement about the need for digital curatorial services and the importance of trained individuals and robust curatorial institutions. Few researchers, even with the prevalence of tools for making data available and the greatest knowledge of their own data sets, felt themselves qualified to make their own data

broadly available over time (or that they had the monetary or human resources to do so), but hoped that "trusted bodies" would be available for them to make their data open.

But the stories they told, while expressing their outward enthusiasm for open data, often tell a different tale. One data curator wryly observed that researchers are very keen to make *other* researchers' data open, but not their own. In fact, the kinds of concerns researchers expressed focused on narratives and stories of misuse. One climate researcher, who had made her published articles available on her Website, said that climate change deniers had deliberately misinterpreted her findings. She expressed concern that making the primary data available would only encourage more such misuses that she would then be responsible for tracking. One civil engineering researcher, who considers herself an enthusiastic open data advocate, noted that one of the concerns in her discipline was of terrorism, which to her seemed misplaced since it was a statistically infinitesimal possibility. She noted that many cities quietly removed publicly available datasets from their Websites after 9/11, citing concerns for infrastructure. She said that other researchers were concerned that making structural data about the built environment broadly available would play into the hands of those who would use such data for attacking weak points in physical structures. She noted that "There's a common refrain in my field—'Osama bin Laden was a civil engineer'."

In some disciplines, data sharing is *de rigueur*, while it is met with suspicion and distrust in others. Nevertheless, the rhetoric of openness and data sharing runs through the curation community, even when the resulting data sets are not going to be made open in the purest sense. Data curators point to studies that show higher citation rates (Piwowar, Day, and Fridsma) for studies linked to open data. While studies have extensively documented many other kinds of fears of making research data open (particularly time and resources to prepare data, lack of control once the data has been released "into the wild", and fears of lack of recognition), asking researchers and curators if they knew of any such incidents proved fruitless—none did.

5 Trust: Creating Shared Engagement

To obtain ongoing financial and institutional support for their work, data curators must engender trust in their profession and their repositories and data creators must trust in the professional capabilities and skills of the creators. To do so, actors must align their different frames of reference, to create shared meaning, to speak to wider audiences, and thus (hopefully) foster collaboration. The trope of trust in process, institutions, and individuals is created through conversations that are by necessity taking place at the same time, often with the same groups and institutions. What they share in common is the need for *enrollment*. Callon and Law (1982) define enrollment as the set of strategies by which actors use texts, technologies, and other means by which to "enlist" other stakeholders to their ways of doing things. To translate: data curators get content creators to "create responsibly" (with easily managed formats, good metadata, and robust documentation, for which they will get scholarly credit and/or fulfill legal and disciplinary obligations), get funders to understand the importance of the data curation activity (and then fund it!), and get everyone involved to "be on the same page" with respect to standards, outreach, and access. In short, the data curation community (writ large) must engage numerous stakeholders—the creators of digital and digitized content, the funders and administrators who must be enrolled to support the activity at the institutional or supra-institutional level, and information technology and other professionals who are "doing the work."

The examination of trust in digital information, especially research data, has a long history in the field (van House). It remains an important conversation in the data curation community because so many communities of practice and interest are engaged—getting researchers to "trust" curatorial services and institutions and the quality of data provided by other researchers such that the data can be reused with confidence, getting institutional buy-in from those funding curatorial services, and other stakeholders. Understanding and trust are rooted in the trustworthiness of institutions and standardized processes and products, education, collaboration, and outreach.

Most of the staff from SSDA who were interviewed emphasized their trustworthiness and how they use narratives of "good data practice" to assure their depositors that they knew how to handle volumes of data. One SSDA staff member said: "It's scary for folks that are having a hard time with all these blogs and data. We've processed it and we've put it out there. We've been the one-stop shop. We've got quality data." Another staff member echoed her words: "What we have to sell is that you can have access or you can have access to stuff that works. It's been tested. It's been run through. We've gone and asked the questions, spent the time to curate it, to make it quality and to make it usable the way our audience wants it. It's usable now, not usable two or three weeks, four weeks down the road."

The importance of using multiple approaches to convey the complicated processes of data curation (and enroll other stakeholders to their "cause") was echoed throughout interviews with preservation and infrastructure developers, who spoke of the time and energy and resources needed to build trust across communities. One developer told the interviewer that working with a team of interested individuals to create a model that "made sense" was one of his finest moments, a component of a narrative that Labov and Waletzky (1967) describe as *evaluation*. He described the situation thus: "I had a great time spending all this time with all these bright people and helping them move forward. By the end of the summer, I had a diagram on a big wall that was three feet high, 30 feet long, and, I mean, you had to kind of get up close to it. They could look at the diagram and say, how can we make more sense of this? They invited me back to work with that committee and support that, and after a few rounds, we managed to do a lot ... without even writing any software or anything, or any kind of automation improvement, just by identifying confusion." A preservation expert explained her role this way: "The policies make it possible for the next person to have a game plan. Part of it is around the people. Part of it is around the technologies ... that you're using durable technologies. For preservation purposes, you don't use cutting edge. You don't use that thing that shows up on your cell phone. That's an access thing. The preservation piece is to make sure that the technologies can convey handshakes across generations of technology that the content can ride around."

Trust in data is as much about trustworthy curatorial processes as it is about the creation of "good data." Such trust requires a sense of community, which is developed through the practice of shared meaning construction across institutional and other boundaries, even within one's own institution (van House, Butler, & Schiff). One data curator observed that "[Y]ou're working in an international community, and it's a virtual community, and that's great. The question is, how do we actually realize the benefits of working a little bit closer to home? What extra can we get out of it? That's where all of these questions come up. What we're trying to do is foster a little bit of more collaboration locally, but the fact that people are collaborating internationally, I think that that's really excellent." One of the institute researchers interviewed made the argument that in her discipline, it is important to curate data because data is expensive to collect. However, she noted that the reuse of curated data is still challenging if one does not personally know the data creator. She and other scientists remarked that there is a big gap between working as an individual or very small group (2–3 people) and working in a group of 8–14 people, much less transitioning to collaborations of hundreds of researchers, a not uncommon pattern in some disciplines.

6 Discussion

A trope is a literary device or metaphor that is so easily recognized and acknowledged that we "know" where the story is going, at least to the point that we can predict how the story is going to work even when we have not encountered the specific instance before. This paper examined three commonly invoked tropes in data curation—the data deluge, the virtues of openness, and the necessity of trust in process and product. Within these tropes, data creators, curators, policy makers, and other actors use narratives to articulate where they position themselves vis-à-vis each other, convey their own understandings to other stakeholders, and transcend disciplinary barriers to shared understanding. At the same time these narratives also express the ways in which different professional communities make sense of what others are doing or should be doing and why. The tropes are prevalent in the field level literature, in popular discourse, on Websites, and in "small narratives": anecdotes and stories that researchers, curation staff, and administrators tell to describe their work to each other and to an interviewer, and how their work fits into larger narratives that are central to the data curation process: an overabundance of data that must be managed, tropes of openness, and trust and mutual engagement between data creators and curators.

More critical attention should be paid to the "grand narratives" of data that transcend the profession of data curation. These narratives are powerful in their ability to construct and shape policy, practice, and thinking on a large scale (Shanahan et al.). There are several that require further unpacking. One that is far broader than the data curation profession is the narrative of transformation: that in the best of all possible worlds, data is transformative and empowering. Data bring us a more rational, open, and transparent society. Gitelman (3) for one asserts that our "zeal for more and more data can become a faith in their neutrality and autonomy, their objectivity." The time for an *anticipatory ethics* (Shilton) of data curation is *now* as societies confront the sheer quantity of data being created by governments, companies, individuals, and even technologies that operate and interact on their own (such as sensor networks and automated high-frequency financial trading). Another grand narrative is the role of data curation and its discourses in shaping cultural policy and a vision for a

future. While the availability of more data is often touted as an emancipatory good, the rhetoric of the "data-driven society" is also troubling because our recent experiences—with the work of Julian Assange, Aaron Swartz, and Edward Snowden—also suggests that the same data can become an omnipresent, omniscient instrument of control.

At the opposite end from the grand is the "small"—what Lampland and Star (2009) characterize as "boring things." How do the narratives and stories, big and small, become embedded in the quotidian? We can pay more attention to standards, tools, practices, and techniques of data curation as well as the texts we create to understand how technical structures and textual productions co-construct each other.

At a practical level, what do the stories and narratives mean for the profession, the discipline, and practice? The data curation literature abounds with urgency—standards, frameworks, education, and other pragmatic concerns necessarily pervade the work in this field. There is a great deal more to be said, some of which is quite troubling, about the myths and stories that are used to bolster the "global data economy" (Johns). This is not to suggest that there are conspiracy theories afoot, but what we need to articulate is an ethos of data curation that we hope can provide a critical voice to the breathless rhetoric of overwhelmingness in the face of endless potential. "Many of us," Dean suggests, "are overwhelmed and undermined by an all-pervasive uncertainty amidst 'seemingly bottomless vats of information' (Dean)." Access to information is in itself not empowering. It is hoped that this paper unveils what stories of NASA data crashes, lost or stolen lab files, or obsolete data formats may mean. As Dr. Jillian Wallis (personal communication), a well-published expert in data curation, noted, librarians and information professionals can be seen as shoe-cobbling elves, cleaning up the data after it is created and making sure that the leaky parts are repaired in good order. Nor are the researchers heroes, bravely putting their fingers in the dikes of data they themselves are responsible for creating.

7 Conclusion: A Call for the Study of Narratives

Data curation is itself predicated upon a central narrative—the life cycle of data (Wallis et al., among others). The curation life cycle model, which visually describes the steps and actions from the creation of data to its curation over time—has to some extent become the de facto central structuring narrative of the field, almost a grand narrative for data curation.

However, other narratives matter and in different ways to different groups and individuals in data curation. People express belonging and empathy, make explicit implicit values, impart cautions and warnings through the stories they tell. In data curation, stories function in the same ways: to raise enthusiasm among potentially bewildered content creators and sceptical administrators, to argue for its importance, and to convey cautionary tales. Such stories are told through PowerPoint slides and professional presentations, YouTube videos, parodies, and other kinds of contemporary information sharing as well as through "traditional" means (orally). Many of these stories have a horror component of unrecoverable loss to individuals but convey a sense of unrecoverable loss to the community or perhaps even to humanity. Enrolling commitment to the data curation cause requires, by definition, planning for it at the inception of creation of content, hence the exhortations to think early and often about planning for it. Words like "stewardship" imply that data curation is a calling that is distributed among the creators, the professional curators, and the stakeholders—who finance the endeavor.

> "For want of a nail the shoe was lost.
> For want of a shoe the horse was lost.
> For want of a horse the rider was lost.
> For want of a rider the message was lost.
> For want of a message the battle was lost.
> For want of a battle the kingdom was lost.
> And all for the want of a horseshoe nail."
> – Traditional Nursery Rhyme

It is appropriate to end an article on stories with another story. The author interviewed a public sector statistician who said he used this nursery rhyme in a PowerPoint presentation to public agency staff to illustrate linked open data. He said, "The nail isn't directly connected to the kingdom but through intervening steps, the two are linked. That's how open data works. People get it." As we tell others stories, we are constructing our own.

This article has wrestled with and distilled some stories and their roles and values in the practices of data curation from the literature, from curators, and from researchers. These narratives imply a past being recounted or an unspecified future good: increased reputation and uptake for the researcher, enhancement of institutional prestige, and better research, all bolstered with the "facts" of scholarly and professional articles that speak to such success stories through empirical analysis. It is worth remembering that increasingly, data curation presents not only economic, material, and cultural opportunities, but

occasionally professional crises, and potential clashes of the public and private good.

References

Anderson, Chris. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired* (23 June 2008); available at http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed 06/29/15).

Bassi, Alessandro, Spyros Denazis, and Pierpaolo Giacomin. "Towards a Noah's Ark for the Upcoming Data Deluge." *Concurrency and Computation: Practice and Experience* 20.17 (2008): 2061–74.

Beagrie, Neil, Julia Chruszcz, and Brian Lavoie. *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities. Final Report to JISC*. April 2008. Available at http://sitecore.jisc.ac.uk/publications/reports/2008/keepingresearchdatasafe.aspx (accessed 6/29/15).

Borgman, Christine. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: The MIT Press, 2014.

Bruner, Jerome S. *Making Stories: Law, Literature, Life*. Cambridge, MA: Harvard University Press, 2003.

Callon, Michel and John Law. "On Interests and Their Transformation: Enrolment and Counter-enrolment." *Social Studies of Science* 12:14 (1983): 615–25.

Coopmans, Catelijne, Janet Vertesi, Michael E. Lynch, and Steve Woolgar. *Representation in Scientific Practice Revisited*. Cambridge, MA: The MIT Press, 2014.

Corti, Louise. "Recent Developments in Archiving Social Research." *International Journal of Social Research Methodology* 15.4 (2012): 281–90.

Dawson, Patrick, and Peter McLean. "Miners' Tales Stories and the Storying Process for Understanding the Collective Sensemaking of Employees during Contested Change." *Group & Organization Management* 38.2 (2013): 198–229.

Dean, Jodi. "Theorizing Conspiracy Theory." *Theory & Event* 4:3 (2000). Available at http:// muse.jhu.edu/journals/theory_&_event/toc/tae4.3.html (accessed 6/29/15).

Doolin, Bill. "Narratives of Change: Discourse, Technology and Organization." *Organization* 10.4 (2003): 751–70.

"The Data Deluge." *The Economist*. (2010); available at http://www.economist.com/node/15579717 (accessed 6/29/15).

Emerson, Robert M., Fretz Rachel I., and Linda L. Shaw. *Writing Ethnographic Fieldnotes*. Chicago, IL: University of Chicago Press, 1995.

Erwin, Tracey, and Julie Sweetkind-Singer. "The National Geospatial Digital Archive: A Collaborative Project to Archive Geospatial Data." *Journal of Map & Geography Libraries* 6.1 (2009): 6–25.

Fenton, Christopher, and Ann Langley. "Strategy as Practice and the Narrative Turn." *Organization Studies* 32.9 (2011): 1171–96.

Gabridge, Tracy. "The Last Mile: Liaison Roles in Curating Science and Engineering Research Data." *Research Library Issues: A Bimonthly Report from ARL, CNI, and SPARC* 265 (2009): 15. Available at http://old.arl.org/bm~doc/rli-265-gabridge.pdf (accessed 6/29/15).

Gitelman, Lisa, ed. *Raw Data is an Oxymoron*. MIT Press: Cambridge, MA: MIT Press, 2013.

Green, Ann G., and Myron P. Gutmann. "Building Partnerships among Social Science Researchers, Institution-Based Repositories and Domain Specific Data Archives." *OCLC Systems & Services: International Digital Library Perspectives* 23.1 (2007): 35–53; available at http://deepblue.lib.umich.edu/handle/2027.42/41214 (accessed 6/29/15).

Heidorn, P. Bryan. "The Emerging Role of Libraries in Data Curation and E-Science." *Journal of Library Administration* 51.7–8 (2011): 662–72.

Jahnke, Lori M., Andrew Asher, and Spencer D.C. Keralis. *The Problem of Data: Data Management and Curation Practices among University Researchers*. Washington, DC: Council on Library and Information Resources, 2012. Available at; http://www.a51.nl/storage/pdf/pub154.pdf (accessed 6/29/15).

Johns, Fleur. "The Deluge." *London Review of International Law* 1.1 (2013): 9-34.

Kouper, Inna. "CLIR/DLF Digital Curation Postdoctoral Fellowship– the Hybrid Role of Data Curator." *Bulletin of the American Society for Information Science and Technology* 39.2 (2013): 46–47. Available at http://www.asis.org/Bulletin/Dec-12/DecJan13_RDAP_Kouper.html (accessed 6/29/15).

Kunda, Sue, and Mark Anderson-Wilk. "Community Stories and Institutional Stewardship: Digital Curation's Dual Roles of Story Creation and Resource Preservation." *portal: Libraries and the Academy* 11.4 (2011): 895–914; available at http://ir.library.oregonstate.edu/xmlui/bitstream/handle/1957/51322/Kunda_AndersonWilk_CommunityStories.pdf?sequence=1 (accessed 6/29/15).

Labov, William, and Joshua Waletzky. "Narrative Analysis: Oral Versions of Personal Experience." *Journal of Narrative & Life History* 7.1–4 (1997): 3–38.

Lampland, Martha, and Susan Leigh Star, eds. *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*. Ithaca, NY: Cornell University Press 2010.

Linde, Charlotte. "Narrative and Social Tacit Knowledge." *Journal of Knowledge Management* 5.2 (2001): 160–71.

Loseke, Donileen R. "The Study of Identity as Cultural, Institutional, Organizational, and Personal Narratives: Theoretical and Empirical Integrations." *The Sociological Quarterly* 48.4 (2007): 661–88.

May, Carl, and Christine Fleming. "The Professional Imagination: Narrative and the Symbolic Boundaries between Medicine and Nursing." *Journal of Advanced Nursing* 25.5 (1997): 1094–1100.

Meyer, Eric T., and Ralph Schroeder. *Knowledge Machines: Digital Transformation of the Sciences and Humanities*. Cambridge, MA: MIT Press, 2015.

Morgan, Gareth. "More On Metaphor: Why We Cannot Control Tropes in Administrative Science." *Administrative Science Quarterly* (1983): 601–07.

Orr, Julian E. *Talking about Machines: An Ethnography of a Modern Job*. Ithaca, NY: Cornell University Press, 1996.

Pennock, Maureen. "Digital Curation: A Life-Cycle Approach to Managing and Preserving Usable Digital Information." *Library & Archives,* January 1 (2007): 34–45.

Piwowar, Heather A., Roger S. Day, and Douglas B. Fridsma. "Sharing Detailed Research Data is Associated with Increased Citation Rate." *PloS One* 2.3 (2007): e308; available at http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000308 (accessed 6/26/15).

Sawyer, Steve. "Data Wealth, Data Poverty, Science and Cyberinfrastructure." *Prometheus* 26.4 (2008): 355–371.

Shanahan, Elizabeth A., Michael D. Jones, Mark K. McBeth, and Ross R. Lane. "An Angel on the Wind: How Heroic Policy Narratives Shape Policy Realities." *Policy Studies Journal* 41.3 (2013): 453–83.

Shankar, Kalpana, Kristin Eschenfelder, Greg Downey, Peter Fleming, Cynthia Engerson, Rebecca Lin, and Jennifer Nygren McBurney. "Social Science Data Archives: Case Studies in Data Sustainability." *iConference 2014 Proceedings* (Berlin, March 4–7, 2014). Available at https://www.ideals.illinois.edu/handle/2142/47417 (accessed 10/18/15).

Shilton, Katie. "Anticipatory Ethics for a Future Internet: Analyzing Values during the Design of an Internet Infrastructure." *Science and Engineering Ethics* 21.1 (2015): 1–18.

Squire, Corinne. "Reading Narratives." *Group Analysis* 38.1 (2005): 91–107.

Tidline, Tonyia J. "The Mythology of Information Overload." *Library Trends* 47.3 (1999): 485–506.

Van House, Nancy A., Mark H. Butler, and Lisa R. Schiff. "Cooperative Knowledge Work and Practices of Trust: Sharing Environmental Planning Data Sets." In *Proceedings of the 1998 ACM conference on Computer Supported Cooperative Work*, pp. 335–343. ACM, 1998.

Van House, Nancy A. "Digital Libraries and Practices of Trust: Networked Biodiversity Information." *Social Epistemology* 16.1 (2002): 99–114.

Wallis, Jillian C. and Christine L. Borgman. "Who is Responsible for Data? An Exploratory Study of Data Authorship, Ownership, and Responsibility." In *Proceedings of the American Society for Information Science and Technology* 48. 1 (2011): 1–10; available at http://works.bepress.com/borgman/251/ (accessed 6/29/15).

Wallis, Jullian C., Christine L. Borgman, Matthew S. Mayernik, and Alberto Pepe. "Moving Archival Practices Upstream: An Exploration of the Life Cycle of Ecological Sensing Data in Collaborative Field Research." *International Journal of Digital Curation* 3.1 (2008): 114–26; available at http://www.ijdc.net/index.php/ijdc/article/view/67 (accessed 6/29/15).

Willis, Craig, Jane Greenberg, and Hollie White. "Analysis and Synthesis of Metadata Goals for Scientific Data." *Journal of the American Society for Information Science and Technology* 63.8 (2012):1505–20.

Wilson, James A. J., Luis Martinez-Uribe, Michael A. Fraser, and Paul Jeffreys. "An Institutional Approach to Developing Research Data Management Infrastructure." *International Journal of Digital Curation* 6.2 (2011): 274–87; available at http://ijdc.net/index.php/ijdc/article/view/198 (accessed 6/29/15).

Witt, Michael. "Co-designing, Co-developing, and Co-implementing an Institutional Data Repository Service." *Journal of Library Administration* 52.2 (2012): 172–88.

World Economic Forum. *Personal Data: Emergence of a New Asset Class.* January 2011. available at: http://www3weforumorg/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011pdf (accessed 6/29/15).

Zimmerman, Ann. "Not by Metadata Alone: The Use of Diverse Forms of Knowledge to Locate Data for Reuse." *International Journal on Digital Libraries* 7.1–2 (2007): 5–16; available at: http://scientificdatasharing.com/wp-content/uploads/2011/02/Zimmerman_Not-by-metadata-alone_2007.pdf (accessed 6/29/15).

Bionotes

Kalpana Shankar

Kalpana Shankar is a Lecturer in the School of Information and Communication Studies at University College Dublin, Ireland. She was previously on the faculty of Indiana University-Bloomington, and completed post-doctoral work at the National Institutes of Health (USA) and the Center for Embedded Networked Sensing, University of California, Los Angeles (UCLA). She received her Ph.D. in Library and Information Studies from UCLA. Her research has been funded by the National Science Foundation (USA), the Irish Research Council, and the Alfred P. Sloan Foundation.