Alex H. Poole

The Conceptual Landscape of Digital Curation

We swim in a sea of data…and the sea level is rising rapidly.
-Anderson & Rainie (2012)

Virtually no one in academia perceives that they have a professional responsibility or mandate for research data management functions.

Introduction

“Contemplating the digital universe is a little like contemplating Avogadro’s number,” claim Gantz and his colleagues (2008). “It’s big. Bigger than anything we can touch, feel, or see, and thus impossible to understand in context” (p. 3). IBM concludes, “Every day, we create 2.5 quintillion bytes of data—so much that 90% of the data in the world today has been created in the last two years alone.”¹ Researchers grapple with “a tsunami of information that paradoxically feeds the growing scientific output while simultaneously crushing researchers with its weight” (Haendel, et al., 2012, p. 1). The data deluge is truly upon us.

That data deluge presents unprecedented challenges in preserving digital assets across all sectors of society, whether organizational, technological, legal, cultural, or business. Rothenberg (1995) opts for an apothegm: “Digital objects last forever—or five years, whichever comes first” (p. 42). Other scholars worry about the potential loss of important data, identifying “A Public Trust at Risk” (Heritage Preservation, 2005), data’s “shameful neglect” (Nature, 2009), or the specter of a “digital dark age” (Bollacker, 2010). Technical obsolescence or fragility, lack of resources, ignorance of good practices, uncertainty over appropriate infrastructure—all constitute serious risks to data (Harvey, 2010). Digital curation tackles these risks.

This article concentrates on foundational Anglo-American digital curation research and practice. It first defines digital curation, discusses its activities, and sets forth the roles and responsibilities of digital curators. Second, it unpacks data and their role in the research enterprise. Along these lines, it argues for the importance of the data lifecycle and of metadata to digital curation. Third, sharing and reuse of data are discussed; digital curation facilitates both
practices. Fourth, this article addresses researcher behaviors and the ways in which disciplinarity, communities of practice, and collaboration shape the sharing and reuse of data. Fifth, digital curation’s impact upon scholarly communication, namely on cyberscholarship, publication and citation, and copyright, is diagrammed. Sixth, digital curation needs to be embedded in institutional and scholarly infrastructure: archives, centers, libraries, and institutional repositories are key parts of that infrastructure. Seventh, standards, governance and policy, planning and data management plans, risk management, evaluation, and metrics, sustainability, and outreach represent overarching concerns for digital curation stakeholders. Finally, directions for future research are adumbrated.

**Digital Curation**

Digital curation centers on “maintaining and adding value to a trusted body of digital information for current and future use.” First used in 2001, the term embraces digital preservation, data curation, and the management of assets over their lifecycle (Lee & Tibbo, 2007; Yakel, 2007).

Digital curation bridges research, practice, and training across nations, disciplines, institutions, repositories, and data formats (Gold, 2010; Ray, 2009). Given the diversity of its stakeholders and the environments in which it is conducted, digital curation potentially involves anyone who interacts with digital information during its lifecycle (Dallas, 2007; Winget, et al., 2009).

The National Research Council of the National Academies (2015) concludes, “The field has grown from practices hardly recognized as curation per se…to international consortia engaged in defining shared norms and standards for digital curation” (p. 17). Researchers, moreover, increasingly recognize its salience. One recent study found that 90% of doctoral supervisors, doctoral holders, and students view digital curation as moderately or extremely important (Abbott, 2015). In spite of the long term importance of digital curation, however, researchers often postpone it (Rusbridge, 2007). This is not necessarily surprising, as they face exigent challenges: a lack of standards, of a common vocabulary, of authority control(s), of appropriate hardware and software, and of storage space (Latham & Poe, 2012; Pryor, 2013).
Digital Curators

Digital curation implicates roles and responsibilities that meld Library and Information Science (LIS) and non-LIS domain skills (Vivarelli, et al., 2013). Notwithstanding technical skills, digital curators require so-called soft skills, namely project management, negotiation, team-building, and collaborative problem-solving (Harvey, 2010; Swan & Brown, 2008).

Nevertheless, the role of digital curation professional is not “a discrete profession labeled ‘digital curator’ with a defined set of tasks undertaken in a dedicated setting…it is…a series of activities undertaken by a range of personnel in a great variety of settings” (National Research Council of the National Academies, 2015, p. 12). Educational programs should therefore cultivate “professional allrounders like a Swiss army knife” (Osswald, 2013).

The Education Advisory Board (2014) pinpoints both a dearth of qualified professionals but an increased demand for such professionals in both the public and private spheres. National employer demand for digital curation professionals increased by more than 50% between H1 of 2010 and H1 of 2013 and by 10% more between H2 of 2013 and H1 of 2014 (Education Advisory Board, 2014).

Digital curation work may seem mundane or even invisible, “a service-oriented back office activity” (Osswald, 2013). Indeed, neither the demand nor the impact of curation activities can always easily be measured, much less communicated (Abram, 2014; National Research Council of the National Academies, 2015). But in advising creators as well as users, ensuring long-term access, facilitating discovery, retrieval, use, and reuse, promoting interoperability, and helping users maximize the usefulness of the curated content, digital curators perform essential tasks (Harvey, 2010).

Data

Defining Data

Unprecedentedly vast in scale and scope, data undergird the scientific record and enable the production of new knowledge (National Academy of Science, 2009; Pryor, 2012; Rusbridge, 2007). Despite this lofty station, the very notion of “data” remains vexatious, seemingly immune even to precise definition. For example, the National Science Board (2005) defines data as “any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc.” (p. 9). Data can be
framed in even more protean ways: they, “like beauty, exist in the eye of the beholder,” Borgman and her colleagues (2012) insist (p. 517). Borgman (2015) similarly expounds, “Data can be many things to many people, all at the same time” (p. xvii).

Grappling with this malleability, stakeholders may define data through example, for instance labeling facts, numbers, or symbols “data.” Similarly, they may resort to operational definitions such as those found in the Open Archival Information System (OAIS) model and the Data Documentation Initiative (Borgman, 2015). The National Science Board (2005) adds some clarity by adumbrating a taxonomy based on three facets: the nature of the data, including its format (images, software, or models, for example) or its origins (observational, computational, experimental, or derivative); the data’s degree of processing; and the data’s reproducibility. But overall, definitional quandaries persist. This inability to achieve consensus hamstrings efforts to optimize Data Management Plans, open data policies, and curation overall (Borgman, 2015).

The Data Lifecycle and Lifecycle Models

There is an unprecedented need to address the entire lifespan of digital content and to do so as early as possible in the research process (Ball, 2012; Tibbo, 2015). Digital content’s lifecycle embraces many components: appraisal and ingest, classification, indexing, cataloging, and authority management, enhancement, presentation, publication, and dissemination, use and user experience and modeling, preservation, and repository management (Constantopoulos & Dallas, 2010).

Lifecycle models describe the ways in which the preservation as well as the addition of value to data can be effected (Pryor, 2012). Models allow stakeholders to map their work progress, to discern vulnerabilities, to encourage documentation, to develop standards, and to identify tools and services; thus they help preserve data’s authenticity, reliability, and usability (Harvey, 2010; Higgins, 2008).

Eight lifecycle models merit consideration: the Digital Curation Center Curation Lifecycle Model (Higgins, 2008), the I2S2 Idealized Scientific Research Activity Lifecycle Model, the Data Documentation Initiative (DDI) Combined Life Cycle Model, ANDS Data Sharing Verbs, the DataONE Data Lifecycle, the Research360 Institutional Research Lifecycle, the Capability Maturity Model for Scientific Data Management, and the UK Data Archive Data Lifecycle. Finally, the Open Archival Information System (OAIS) model broke new ground in the late 1990s and became ISO standard 14721: 2003. It does not, however
constitute a full-fledged lifecycle model; it also neglects to specify guidelines for creating or (re)using data (Lee, 2005; Lee, 2009).

In developing or following an appropriate lifecycle model, designers should consider issues such as scope (what services will be offered and to what audiences), best practices or community standards or both, and model’s ability to represent work practices (Carlson, 2014). Lifecycle models constitute vital resources for digital curation work.

**Metadata**

Metadata deal with data attributes “that describe, provide context, indicate the quality, or document other object (or data) characteristics” (Greenberg, 2005, p. 20). Supporting almost all of the steps in the digital curation lifecycle, they are tantamount in importance to data themselves (Levine, 2014; Riley, 2014). But one recent study underscores researchers’ unfamiliarity with creating or documenting metadata; another discerns that researchers put little effort into metadata creation because they cannot predict the needs of future reusers (Akers & Doty, 2013; Wallis, et al., 2013). Edwards et al. (2011) underline researchers’ recalcitrance in recording even basic metadata.

The benefits of metadata include not only a controlled vocabulary, but also information on related objects, on intellectual property, on user information, on versioning, on integrity checks, and on preservation (Harvey, 2010). Stakeholders consider abundant structured metadata a “holy grail” with respect to sharing and reuse (Edwards, et al., 2011, p. 672).

**Sharing and Reuse**

**Sharing**

Sharing data allows scholars to detect fraud or to address disputes, to reproduce or to verify research findings, to make findings generated by publicly-funded research available, to enable meta-analysis, to increase citation, to reduce loss, to enrich teaching and learning, to generate new knowledge from familiar data and thus to leverage investments, to avoid the costs of re-collecting data, and to foster economic development (Borgman, 2012; Borgman, 2015; Faniel & Zimmerman, 2011; Heidorn, 2008; Lyon, 2009; McLure, et al., 2014; Organisation for Economic Co-Operation and Development, 2007; Parsons & Duerr, 2005; Ray, 2014; Tenopir, et al., 2011; Whitlock, 2011; Zimmerman, 2008). Researchers can share data through deposit in a data center, archive, bank, or Institutional Repository (IR), through submission to a journal as
part of an article, through discrete publication, through mounting them online, and through peer exchange (Akers & Doty, 2013; Van den Eynden, et al., 2010; Wallis, et al., 2013).

Researchers who consider sharing must know which data can be shared and why, by and with whom, under what conditions, and to what effect: discipline, age, research as opposed to teaching focus, and geographical region comprise important considerations (Borgman, 2012; Tenopir, et al., 2011). Prior to sharing, moreover, data must be gathered, structured, and documented in ways that accommodate particular domains and disciplines (Akers & Doty, 2013; Haendel, et al., 2012; Ray, 2014; Van den Eynden, et al., 2010). Yet even with considerable documentation, sharing remains difficult: accessibility is not commensurate with usability (Akmon, et al., 2011; Tenopir, et al., 2011; Wallis, et al., 2013).

Direct contact among researchers and other professionals remains the most important precondition for collaboration (Kroll & Forsman, 2010). Trust nurtured through personal relationships usually undergirds sharing (Akers & Doty, 2013; Akmon, et al., 2011; Cragin, et al., 2010; Duerr, et al., 2004; Faniel & Jacobsen, 2010; Kroll & Forsman, 2010; Pryor, 2009; Sayogo & Pardo, 2013; Tenopir, et al., 2015; Wallis, et al., 2013; Zimmerman, 2008). Private communication therefore plays an integral role: researchers not only can convey content, but also can provide documentation about the data’s attributes and their applicability (Borgman, 2015). It also renders tacit knowledge more transferrable (Birnholtz & Bietz, 2003; Polanyi, 1966; Wallis, et al., 2013). On the other hand, private communication fails to enhance discoverability, usability, or longevity (Borgman, 2015).

Disincentives to share data also persist. Edwards et al. (2011) insist, “Every moment of data across an interface comes at some cost in time, energy, and human attention”—i.e. “data friction” (p. 669). Six more granular concerns also apply. First, preparing data, for instance by writing documentation, is resource-intensive (Wallis, et al., 2013). Concomitantly, researchers may not believe their data can be useful to others (Steinhart, et al., 2012). Second, researchers fear they will not receive credit for sharing their intellectual capital (Akers & Doty, 2013; Tenopir, et al., 2011; Wallis, et al., 2013). Third, researchers want an embargo period that allows them to analyze or otherwise use their data (Wallis, et al., 2013). For example, Cragin et al. (2010) found that 40% of respondents resisted sharing prior to an embargo period unless that sharing was circumscribed to collaborators or associates. In another study, although 95% of respondents expressed willingness to share their data at some point in their research, more than
two thirds (68%) wanted to wait at least six months after analysis to do so (Steinhart, et al., 2012). Fourth, researchers worry about misuse, primarily about misinterpretation, but also about intellectual property (Akers & Doty, 2013; Cragin, et al., 2010). One study determined that if their data were misused, researchers expressed less willingness to share data subsequently (Cragin, et al., 2010). Fifth, ethical concerns such as confidentiality or privacy can discourage sharing (Akers & Doty, 2013; Borgman, 2012; Cragin, et al., 2010; Pryor, 2009; Steinhart, et al., 2012). Finally, legal and rights issues can militate against sharing (Steinhart, et al., 2012).

Discerning data’s potential reuse value and the resources needed to ensure fit-for-purpose remains the litmus test for sharing (Palmer, et al., 2011). Most problematic, purported willingness to share may not in fact lead to data release, much less reuse (Borgman, 2015) (Scaramozzino, et al., 2012; Wallis, et al., 2013). Small wonder sharing seems a concept “almost byzantine” (Carlson, et al., 2011, p. 647) and remains “maddeningly difficult” (Edwards, et al., 2011, p. 670).

Reuse

Reuse may include migrating, enhancing, aggregating, or reanalyzing data (McLeod, et al., 2013). But reusing shared data erects further hurdles. First, the provenance of the dataset may be ambiguous (Zimmerman, 2008). Second, optimal contextualization and documentation relies upon understanding a discipline’s history and the nuances of its research community (Carlson & Anderson, 2007). Indeed, researchers often defer to disciplinary norms or to conventional wisdom (however defined). Third, reusability necessitates not only additional policies, but also mechanisms for enforcement and compliance (Zimmerman, 2008). Fourth, aligning terms of use or agreements can prove a stumbling block (Smith, 2014). Overall, sharing will likely thrive in communities that realize direct benefits or in situations in which infrastructure is in place (Jones, 2012).

Relying upon current eclectic sharing practices contravenes best scholarly practices (Nicholson & Bennett, 2011). Wallis et al. (2013) report:

The originating investigator bears the cost of data preparation. Other entities such as data repositories, universities, libraries, and funding agencies are likely to bear the cost of curating those data for sustainable access. Unknown—and often nonexistent—reusers reap the benefits. This equation is not viable in economic or social terms (p. 15).
Exacerbating the economic challenges of reuse, the amount of data reuse overall is unknown; in the majority of fields, in fact, demand for reusable data seems close to nil (Borgman, 2015).

**Researchers and Researcher Behaviors**

*Disciplinarity*

“Within or across disciplines,” states Prior (2012), “members of that workforce will over time combine, disperse and recombine with seeming fluidity; the research undertaken will rarely follow an exclusive and linear path and as a community they will exhibit changing patterns of allegiance and interests” (p. 9). As a result, curation services must embrace a range of interdisciplinary, disciplinary, and domain practices (Akers & Doty, 2013; Cragin, et al., 2010; Karasti, et al., 2006; Lage, et al., 2011; Lyon, 2007; Lyon & Brenner, 2015; Lyon, et al., 2009; Myers, et al., 2005; Palmer & Cragin, 2008; Palmer, et al., 2009; Pryor, 2009; Shankar, 2007).

It appears that data are not—nor perhaps soon will be—fully translatable among fields (Borgman, 2007). The disciplinary culture of agronomy at one university impeded sharing data (Carlson & Stowell-Bracke, 2013). A smattering of studies aside, the ways in which data travel among scholarly fields remains an understudied area (Edwards, et al., 2011).

*Communities of Practice and Collaboration*

A community of practice “denotes the level of the social world at which a particular practice is common and coordinated, at which generic understandings are created and shared, and negotiation is conducted” (Davenport & Hall, 2002, p. 172). Communities of practice reconcile an organization’s traditional perspective and its evolving priorities for future activities (Brown & Duguid, 1991).

Communities of practice revolve around problem-solving, knowledge mapping, fulfilling information requests, reusing assets, coordination, documentation, and conversations and visits (Wenger, 2006). The Web offers opportunities for expanding existing and creating new communities of practice (Wenger, 2006).

Communities of practice also depend upon collaboration (Wood & Gray, 1991; Yarmey & Baker, 2013). Ideally, collaborations rest upon openness, patience, and tolerance of alternative research philosophies (Gunawardena, et al., 2010). Perhaps most important, communities of practice and the collaborative relationships they nourish ground new forms of scholarship.
Scholarly Communication

Cyberscholarship

High performance computing, visualization, databases and datasets, and networking enable a new form of scholarly communication: cyberscholarship (Lynch, 2014). Cyberscholarship rests upon collaborative resource discovery, discussion, and analysis of texts, images (both moving and static), sound recordings, and maps and geographical information systems (Green & Roy, 2008). It allows the pursuit of new forms of research, developing tools for collection-building and analysis, and creating new intellectual products (American Council of Learned Societies, 2006; Arms, 2008). Waters (2007) comments,

Although the systematic exploration of large quantities of information is not a new scholarly practice, what does seem new is the formalization of the very traditional interpretive activities of data-mining, pattern-matching, and simulation in powerful algorithms that represent large and complex sets of data in terms of multiple features and variables that can be analyzed, tested, replicated, and changed at the scale and speed afforded by advanced computation (p. 8).

The development of infrastructure for cyberscholarship demands national, international, and interdisciplinary coordination (Arms & Larsen, 2007). For stakeholders, it requires domain and computational knowledge and a lingua franca as well (Bowker & Star, 2009). Last, cyberscholarship requires strategies to deal with publication and citation as well as with copyright and legal issues. Lynch (2014) concludes, “Changes in the practice of scholarship need to go hand-in-hand with changes in the communication and documentation of scholarship” (p. 15). Data publication and citation represents a crucial aspect of such communication and documentation.

Data Publication and Citation

The contention that datasets should be part of the scholarly record and treated as first-class research products continues to attract acolytes (Kratz & Strasser, 2014). Parsons and Fox (2013) characterize data publication as “a metaphor of choice to describe the desired, rigorous, data stewardship approach that creates and curates data as first class objects” (p. WDS32). Consensus obtains regarding data publication: they should be available publicly and ideally, in
perpetuity; they should be sufficiently documented; and they should be citable (Kratz & Strasser, 2014). But Parsons and Fox (2013) rightly wonder: “Is data publication the right metaphor?”

Citations once bolstered scholarly arguments; now, too, they are implicated in provenance, discovery, quality, and attribution (Brase, et al., 2014). Citations allow identification, retrieval, and attribution of data; they incentivize sharing and reuse, thereby promoting scholarly productivity (Mooney & Newton, 2012).

Data citation remains exceptional, however: inertia hampers many research communities (Mayernik, 2012). Researchers may not know that they should cite data. Even should they know, they might not know how to cite them; journals seldom provide instructions. One content analysis found data citation practices desultory (Mooney & Newton, 2012). Even the notion that data citation will impel researchers to share their data remains unproven (Borgman, 2015).

Data citations should provide a unique and persistent identifier; receive the same credit as traditional citations; give intellectual and legal attribution; enable access to data, metadata, and documentation; accommodate provenance, documentation, standards, and domain practices; and be usable by tools as well as persons (Ball & Duke, 2012; Brase, et al., 2014). Outstanding challenges include granularity, microattribution, identifiers for contributions, and citation placement (Ball & Duke, 2012).

Copyright

Like data publication and citation, copyright and intellectual property are bedeviling issues in digital curation. For example, copyright law seems both “ubiquitous and yet not at all intuitive” (Levine, 2014, p. 140). But copyright may be less knotty than other intellectual property issues. Smith (2014) sums up, “Copyright law, while complex and nuanced, is largely harmonized (in the sense that it is structured, interpreted, and enforced in the same way) worldwide, unlike other types of intellectual property law (e.g., sui generis database rights or patent rights)” (p. 46).

Legal issues interpenetrate digital curation work: technological change always outstrips legal change, for instance, and both countries’ copyright laws and the disciplinary practices of researchers vary (Levine, 2014). Legal tools for protecting or sharing data include contracts, public licenses, and waivers (Smith, 2014). Fundamental impediments may be more nebulous than copyright, however, namely issues surrounding credit and attribution and promotion and tenure (Levine, 2014).
As issues related to copyright and intellectual property, data publication and citation, and cyberscholarship suggest, scholarly communication’s “evolution, revolution, or crisis” continues (Borgman, 2007, p. 9). Infrastructure looms large in this regard.

Infrastructure

As Poole (2015) suggests, four types of institutions show much potential with respect to digital curation: archives, centers, libraries, and institutional repositories. Yet researchers remain underinformed about the strengths and weaknesses of each one (Steinhart, et al., 2012).

Archives

Archivists’ areas of expertise include ownership (and its transference), donor relations, policy development, cyberinfrastructure, repository management, intellectual property, selection and appraisal, creation, use, and reuse contexts, authenticity, trustworthiness, provenance, access and use restrictions, permanence, and metadata (Akmon, et al., 2011; Dooley, 2015; Gilliland, 2014; Prom, 2011). Wallis and her colleagues (2008) stump for implementing archival practices as early as possible in the digital curation lifecycle.

Despite the applicability of these skills, archivists’ involvement in digital curation remains exceptional—even as archival principles increasingly permeate sundry academic fields (Dooley, 2015; Manoff, 2004; Manoff, 2010; Poole, 2014). As Redwine et al. (2013) point out, archivists often acquire digital materials incidentally. Archivists already face substantial existing backlogs of paper materials (Prom, 2011). More than a third (34%) of repositories in one study reported that more than half of their holdings remained unprocessed (Greene & Meissner, 2005). Further, archivists may see data as peripheral to historical research; they may prefer to deposit them in discipline-specific repositories (Akmon, et al., 2011).

Archivists can serve as digital curation participants or consultants or both. A repository’s mission statement and its collection development policy should steer its involvement (Noonan & Chute, 2014). A recent survey of college and university archivists found well over half (57%) its respondents involved in institutional deliberations regarding digital curation. What was more, nearly half of respondents collected institutional or research data. Nonetheless, size mattered: the largest institutions demonstrated the most archivist involvement. Most important, the vast majority of participants (86%) believed archivists should be involved with digital curation on
some level, but only 54% felt capable of fulfilling their perceived roles (Noonan & Chute, 2014). Archivists are increasingly well-placed to address pressing digital curation needs (Prom, 2011).

Centers

Centers such as the Digital Curation Centre can create or store data, can enable data discovery, can improve research efficiency, can increase return on investment, and can raise awareness of data’s disciplinary importance (Beagrie, 2004; Collins, 2012; Donnelly, 2013; Hockx-Yu, 2007; Pryor, 2013; Rusbridge, et al., 2005). Centers advocate for scholarship and pedagogy, promote interdisciplinarity, attract new audiences, engage professional and scholarly communities, and stimulate collaborations (Zorich, 2009). But still lagging is the ability to embed these centers’ expertise institutionally (Lyon, 2007). Indeed, centers struggle to sustain themselves; funders or administrators may overlook research centers’ merits (Collins, 2012).

Libraries


Librarians understand metadata, information literacy, scholarly communication, open access, and both collection and repository management (Corrall, et al., 2013). Subject specialists, metadata librarians, institutional repository coordinators, systems and IT librarians, copyright specialists, collection managers, and acquisitions librarians—all can be involved.

Libraries and librarians can increase awareness of digital curation’s importance, can provide preservation services, and can cultivate new professional practices (Swan & Brown, 2008). Perhaps most important, digital curation facilitates librarians’ further embedding themselves in research processes (Nicholls, et al., 2014). Partnering with disciplinary data repositories such as Dryad is one strategy for doing so (Akers & Green, 2014).
As yet, though, the commitment of libraries to digital curation is inconsistent (Council on Library and Information Resources, 2013). Librarians who work with data often do so more or less in isolation (Steinhart, 2014). Moreover, determining data’s place in a library’s prosaic operations is hardly intuitive (Steinhart, 2014). Finally, the scholarly literature evinces little consensus regarding the application of librarians’ skills—much less what additional skills they need to cultivate (Carlson, 2013). That said, librarians need certain competencies: not only knowledge of institutional, disciplinary, and domain cultures, but also of research methods, workflows, lifecycles, technology, and relationship-building (Corrall, 2012). Localized studies can help determine the appropriate mix of skills librarians need (McLure, et al., 2014).

Complicating matters, researchers lament their inability to create sharable metadata and to manage their data efficiently, but show little awareness that libraries and librarians can help (Kroll & Forsman, 2010). The library seems a mere “dispenser of goods” (Jahnke & Asher, 2012, p. 4). Similarly, researchers may think of librarians as support services, not as true collaborators (Wright, et al., 2014). Concern persists that researchers will neglect to work with librarians; Data Management Plans (DMPs) may suffer as a result (Hswe & Holt, 2011). As with archivists, librarians’ potential with respect to digital curation has only just been tapped (Carlson, et al., 2013) (Fox, 2013).

Digital Curation Profiles (DCPs)

Developed at Purdue University, Digital Curation Profiles constitute a key resource for librarians (Witt, et al., 2009). DCPs provide information about data creation or use or both, about data management, and about the curation needs of the researcher(s). Comprehensible to non-specialists, they enable comparison across domains and disciplines (Carlson, 2013).

A recent study of twelve DCP workshops determined that 237 of 259 participants (all librarians), had data responsibilities. Participants encountered familiar obstacles, however, primarily organizational support, time, staff, and resources. Although workshops boosted attendees’ confidence levels, their levels of engagement changed little (Carlson, 2013).

The DCP initiative also developed a Digital Curation Profiles Toolkit (DCPT). After users critiqued the DCP Toolkit as time-consuming, insufficiently modular, and overly science- and engineering-inflected, Purdue personnel set to developing a new version (Brandt & Kim, 2014; Zhang, et al., 2015). Other institutions might profit from Purdue’s innovative example.
Like DCPs building on libraries’ potential for digital curation work, the National Science Foundation’s 2008 DataNet initiative funded Data Conservancy and DataONE (Lee, et al., 2009; Sandusky, et al., 2009). First, though defunded in 2013 Data Conservancy established a template for developing a distributed human and technical infrastructure (Treloar, et al., 2012). Second, the DataONE project traverses disciplines, institutions, and nations; it focuses on the biological, ecological, and environmental sciences lifecycles (Allard, 2014; Preservation and Metadata Working Group, n.d.; Tenopir, et al., 2011).

Three DataNet projects (2011-2016) succeeded Data Conservancy and DataONE. First, Sustainable Environment through Actionable Data (SEAD) effects “sophisticated management of heterogeneous data while dramatically lowering the cost and effort required to curate and preserve data for long-term community use.” Second, the DataNet Federation Consortium employs the integrated Rule-Oriented Data System (iRODS) to offer nationally federated data grid infrastructure and to encourage collaborative research among scientists and engineers. Third, Terra Populus provides tools for data integration across the domains of social and environmental science. In hosting DataNet projects and developing DCPs, among other activities, libraries seem poised to play a key role in digital curation work. Institutional repositories may prove similarly critical. Here, too, librarians play a central if still unsettled role (Allard, et al., 2005).

**Institutional Repositories (IRs)**

Often associated with libraries, IRs fuse human intervention and institutional investment (Kunda & Anderson-Wilk, 2011). IRs may support project conception, proposal development, scheduling, documenting, embargoing, communicating within and among research groups, data exchange, and storage (Walters, 2014). Not only can they manage scholarship and data, but also software, tools, and code (Cragin, et al., 2010; Walters, 2014).

The Purdue University Research Repository (PURR) and the Merritt Repository hosted by the University of California exemplify cutting-edge IR work. PURR seeks to offer a “‘cradle-to-grave’ service” that focuses on lifecycle services, consultations, publishing functionality, and dataset discovery and preservation (Brandt, 2014, p. 333). The Merritt offers storage, integrity checking, versioning, a metadata catalog, access control rules, use agreements, preservation services, and asynchronous delivery (Abrams, et al., 2014).
As PURR and the Merritt suggest, institutional repositories may bridge personal and university servers and national and international repositories (Akers & Green, 2014). Macdonald & Martinez-Uribe (2010) call for further intra-institutional, inter-institutional, and cross-facility services.

Given the overall scale and complexity of the digital curation mandate, there are surely roles for IRs, libraries, centers, and archives. But institutions and organizations need to reconcile variegated disciplinary, domain, and research community perspectives and practices and to determine the role of funding agencies (Lynch, 2014).

**Overarching Concerns**

Overarching concerns in digital curation include standards, governance and policy, planning, risk management, evaluation, and metrics, sustainability, and outreach.

*Standards*

Standards are “mechanisms of coordination that facilitate the exchange and comparability of information and practices in the global community” (Yarmey & Baker, 2013, p. 157). Key in transforming local into public knowledge, they shore up authority, authenticity, reliability, and usability (Higgins, 2009). Standards smooth the way as well for discoverability, accessibility, interoperability, collaboration, and preservation (Higgins, 2009; Lyon, 2007; Zimmerman, 2008).

Lack of coordination hinders the development and adoption of standards and best practices, as do broader political, cultural, scientific, and technical issues (Griffiths, 2009; National Research Council of the National Academies, 2015; Zimmerman, 2008). Many disciplines and domains lack standards and remain ill-informed about standards’ affordances (Dallmeier-Tiessen, et al., 2014; Lyon, 2007; Tenopir, et al., 2011). Choosing among sometimes competing standards adds yet another wrinkle (Dallmeier-Tiessen, et al., 2014). One study discerned that the absence of standards hampered graduate students’ efforts to manage their agronomic data (Carlson & Stowell-Bracke, 2013).

Without incentives, researchers may well resist (Baker & Yarmey, 2009; Edwards, 2004). Despite the potential cost financially and temporally of doing so, Yarmey and Baker (2013) lobby for standard-making as an evolutionary process. Standards consistently justify their costs;
they encourage the development of local best practices (Lynch, 2008; Yarmey & Baker, 2013). Still, they may cause as many problems as they solve (Borgman, 2015).

**Governance and Policy**

“The system of decision rights and responsibilities covering who can take what actions with what data, when, under what circumstances and using what methods,” data governance ensures that data can be trusted and that stakeholders remain accountable (Smith, 2014, pp. 45-46). Governance centers on risk management, legal and policy issues, attribution and citation, archiving and preservation, discovery and provenance, data schema and ontologies, and infrastructure (Smith, 2014).

In digital curation, policy vacuums abound with respect to copyright and other intellectual property issues (Smith, 2014). Researchers negotiate Institutional Review Board and funders’ requirements, professional ethical codes, and cultural differences (Council on Library and Information Resources, 2013). But policies often gloss over or ignore inter-domain differences in research practices (Borgman, 2015). They tend also to emphasize supplying more data, probing neither scholarly motivations nor expected reuse demand nor necessary infrastructural investment (Borgman, 2015). In some cases, policies are at loggerheads.

Two recent studies illuminate the inchoate state of policy development. First, only 9% of respondents to the DataRes Project’s survey said their institution had a data management policy. Nearly three-quarters, by contrast (72%), said their institutions did not have such a policy and fully 19% said they did not know (Council on Library and Information Resources, 2013). Second, Dietrich et al. (2012) found that funders’ policies often lacked specific language and emphasized access to more than preservation of data—even though they provided little guidance about enabling access. Policies also neglected by and large to address open access.

Policies should address accountability, legitimacy, best practices, and monitoring and review (Harvey, 2010). They should accommodate various types of institutions and data products (Lyon, 2007). Though institutional policies should align with funding agency mandates, both institutional inertia and the still-ambiguous status of data qua research product undercut such efforts (Council on Library and Information Resources, 2013). Ultimately, judicious planning periodically reviewed can mold governance and policy.
Planning and Data Management Plans (DMPs)

Data Management Plan requirements reflect funders’ efforts to show salutary economic and societal impact (Lynch, 2014). Following the lead of the National Institutes of Health (2003), in 2010 the National Science Foundation mandated that all grant applications beginning in 2011 include a Data Management Plan (DMP). This represented the first step in ensuring that federally funded research would become available to the general public (Mervis, 2010). In implementing its own Data Management Plan requirements, the White House’s Office of Science and Technology Policy (2013) impelled researchers to make the results of federally-funded scientific research available to maximize the impact of the federal government’s investment and to ensure the accountability of those funded.

Three studies indicate the ambiguous situation of DMPs in researchers’ work. The DataRes project’s survey found that the vast majority (87%) of respondents agreed or strongly agreed that a data management plan was important (Council on Library and Information Resources, 2013). But Steinhart et al.’s (2012) study at Cornell University found great uncertainty among Principal Investigators about National Science Foundation requirements. Nearly two-thirds (62%) of respondents wanted help writing DMPs (only 13% expressed no such interest). Another study found many researchers at Colorado State University ignorant of what a DMP was as well as varied perspectives as to what a DMP mandated in practice (McLure, et al., 2014). Librarians can assist by conducting workshops to disseminate best practices, by fostering communication, and by promoting tools such as DMPTool 2 (Nicholls, et al., 2014).

Despite researchers’ ostensible commitment to the process, planning may be postponed or even dismissed (Donnelly, 2012). Many researchers neglect to develop DMPs unless prodded by funding agencies (Jahnke & Asher, 2012). Though researchers can scarcely ignore funder requirements, funders struggle to track compliance (Hswe & Holt, 2011). Barring reliable vehicles for enforcement, audits may prevail—hardly an ideal option (Lynch, 2014).

Key considerations regarding a DMP include the data to be generated and shared, funder, publisher, and internal requirements, disciplinary and domain standards and best practices, deposit location and provisions for backup, legal and ethical obligations, and costing (Donnelly, 2012; Ray, 2014; Van den Eynden, et al., 2010). A DMP should address stakeholders ranging from researchers, librarians, and center or repository managers, to support staff such as grants officers, specialists such as archivists, and technical and laboratory staff (Donnelly, 2012).
Planning Tools

Researchers’ current practices appear fragmented largely because funding agencies propagate broad requirements and provide few resources (Sallans & Lake, 2014). A vicious circle may prevail: most disciplines lack standardized procedures for managing data and most researchers have yet to skill up (Sallans & Lake, 2014, p. 88). Tools such as DMP Online and DMPTool can help.

Building upon librarians’ expertise, DMP Online and DMPTool both provide structured environments for data management planning and link to funder requirements (Sallans & Donnelly, 2012). “While the development paths of the tools have diverged,” Sallans and Donnelly (2012) conclude, “both groups retain a broader vision of a joined-up tool…that serves as a coordinating hub for the management of data across many disciplines, many funding agencies, many institutions and many countries, with shared good practice as a common goal” (p. 128).

Recently the six original partner institutions involved in DMPTool developed DMPTool 2. Goals for DMPTool 2 included settling upon best practices, allowing local adaptation, cultivating an open source community, promoting the sharing of institutional resources, and supporting the data lifecycle overall (Strasser, et al., 2014).

Risk Management, Evaluation, and Metrics

In digital curation, risk management constitutes “a continuously developing arena whose ultimate goal is to define preservation and control mechanisms to address the risk attached to specific activities and valuable assets” (Barateiro, et al., 2010, p. 6). Risk management procedures include discerning potential risks, assessing the degrees of risk and parsing them into acceptable and unacceptable, gathering information, settling upon controls, reporting results, and developing action plans (Price & Smith, 2000).

Minimizing risks to data involves providing for persistent access, multiple copies, periodic backup, a delimited number of file formats, secure storage, disaster recovery, and community monitoring (Harvey, 2010). Further, in choosing a backup strategy stakeholders should assess local conditions, the value of the data, the systems that store them, and acceptable levels of risk (Van den Eynden, et al., 2010).

Risk management funnels into evaluation. Evaluation stimulates a culture of quality and responsibility (Lakos & Phipps, 2004). Audits allow stakeholders to identify previously
unnoticed risks and to refine their definitions of those designated communities whom they serve (Reilly & Waltz, 2014). Four tools lend themselves to this process.

First, the *Trustworthy Repositories Audit & Certification: Criteria and Checklist* (TRAC) helps repositories transparently tackle risks. Although it does not stipulate standards and specifications, on one hand, it insists that the designated community’s priorities are non-negotiable, on the other (Center for Research Libraries, 2007; Reilly & Waltz, 2014). Second, the Planning Tool for Trusted Electronic Repositories (PLATTER) helps repositories establish trusted status by assisting in the development of performance objectives (DigitalPreservationEurope, 2008). Third, the Digital Repository Audit Method Based on Risk Assessment (DRAMORA) allows repositories to self-assess and thus to quantify their risks (McHugh, et al., 2008). Finally, the Consultative Committee for Space Data Systems’ (CCSDS)’s 2011 *Audit and Certification of Trustworthy Digital Repositories* premises continuous improvement. (It became ISO 16363 in 2012.)

Such tools cannot inoculate an organization against risk, however; measurement to discern levels of risk is crucial. Measurement facilitates learning, decision-making, and awareness-raising (Whyte, et al., 2014). It should focus both on developing institutions’ research support services and infrastructure and on ultimately descrying their economic impact (Whyte, et al., 2014). As benefits can resist measurement, a dynamic mix of strategies that incorporates qualitative as well as quantitative analysis is pivotal (Fry, et al., 2008). Metrics can be leveraged to provide for sustainability.

*Sustainability*

Inadequate financial resources loom large (National Research Council of the National Academies, 2015). Counterintuitively, quantifying benefits may be *more* difficult than quantifying costs (Fry, et al., 2008). Though Lavoie (2012) argues that sustainability is inextricable from economics, the Repository Task Force (2009) insists that sustainability involves organizational commitment and collaboration. At any rate, optimal sustainability plans feature value propositions, varied revenue streams, and commitments to accountability (Maron, et al., 2009).

Projects may assume two approaches to costing. First, the project can price all activities and resources for the entire lifecycle. This allows the calculation of the total cost of data
generation, sharing, and preservation. Second, the project can determine the additional expenses that will be incurred to allow the data to be sharable (Van den Eynden, et al., 2010).

The LIFE Project, for example, improved planning, comparison, and evaluation of digital lifecycles (LIFE Project Team, 2008). The Keeping Research Data Safe (KRDS) project constructed an activity model and a TRAC-based resource template (Beagrie, et al., 2008). Keeping Research Data Safe 2 reviewed and extended KRDS’s activity model and crafted a benefits framework based on case studies (Beagrie, et al., 2010). Finally, the European Union’s 4C: Collaborating to Clarify the Cost of Curation project developed the Digital Curation Sustainability Model (DCSM) and the Curation Cost Concept Model and built a Curation Costs Exchange (CCEx) (Kilbride & Norris, 2014; Middleton, 2014).

Ultimately, the sustainability of digital curation work should focus on making decisions that are optimal for current needs and yet will not preclude future stakeholders from changing course (Kunda & Anderson-Wilk, 2011). A dearth of funding imperils data and saps the development of a competent workforce to ensure their proper curation (National Research Council of the National Academies, 2015). Lyon (2007) insists, “We need to construct new economic models to provide a robust foundation to funding plans, models which link research strategy and program development, to operational support and infrastructure provision.” In addition to financial concerns, sustainability incorporates raising awareness and outreach.

**Outreach**

Both cognizance and comprehension of digital curation best practices remains low (Lyon, 2007). Researchers tend to satisfice, as many feel that funding agencies, publishers, and professional societies neglect to extend sufficient support (Carlson, et al., 2013; Kroll & Forsman, 2010). Researchers who participated in a study at Emory University emphasized their interest in attending digital curation workshops and in receiving help preparing their Data Management Plans (Akers & Doty, 2013).

Incentives are needed to propel cultural change in research communities (Jones, 2012; Mooney & Newton, 2012; Whitlock, 2011). “Researchers,” conclude Patrick and Wilson (2013) astutely, “are not rebellious schoolchildren who need to be bullied into working harder; they are generally highly motivated and highly skilled individuals who take a great deal of pride in what they do, and thus are more likely to embrace digital curation as a worthy goal if persuaded of its merits.” Outreach and engagement with actual user behaviors is imperative to improve services
(Reilly & Waltz, 2014). Outreach, after all, necessitates little if any further financial investment—merely a receptive audience (Shorish, 2012).

***

“We are at the early stages of a genuine systemic and systematic response to the data stewardship challenges framed by the emergence of e-research, and to seizing the opportunities promised by more effective, broadscale sharing and reuse,” Lynch (2014) contends (p. 406). In addressing this transition, scholars might look to five questions. First, at what point of the lifecycle can digital curators’ input and support be most valuable to researchers? Second, what can further examination of researcher practices suggest about common digital curation needs across communities of practice, domains, disciplines, and institutions? Third, how can archivists best expand their role in digital curation work? Fourth, what strategies can be used to secure funding beyond one-shot grants and what strategies can be employed to secure funding for ongoing as well as for new digital curation initiatives? Fifth, what balance of funder carrots and sticks will influence researchers and other project personnel to carry out activities that maximize return on investment?

“A revolution in digital information is occurring across all realms of human endeavor,” notes the National Research Council of the National Academies (2015, p. 7). Digital curation can help put this revolution to work.

References


Bowker, G. C. & Star, S. L., 2009. Cyberscholarship; or, 'A Rose is a Rose is a...'. *EDUCAUSE Review*, 44(3), pp. 6-7.


Wenger, E., 2006. *Communities of Practice: A Brief Introduction*. [Online] Available at: https://scholarsbank.uoregon.edu/xmlui/bitstream/handle/1794/11736/A%20brief%20introduction%20to%20CoP.pdf?sequence=1


[2] “A person, company, etc., with a concern or (esp. financial) interest in ensuring the success of an organization, business, system, etc.” (*OED*).
[8] [http://www2.le.ac.uk/services/research-data/images/new_institution.PNG/view](http://www2.le.ac.uk/services/research-data/images/new_institution.PNG/view)
[10] [http://www.data-archive.ac.uk/create-manage/life-cycle](http://www.data-archive.ac.uk/create-manage/life-cycle)
[12] [https://dataconservancy.org/](https://dataconservancy.org/)
[13] [https://www.dataone.org/what-dataone](https://www.dataone.org/what-dataone)
[14] [http://sead-data.net/](http://sead-data.net/)
[15] [http://datafed.org/about/](http://datafed.org/about/)
[16] [http://www.terrapop.org/](http://www.terrapop.org/)