Arch Sci
DOI 10.1007/s10502-014-9219-z

ORIGINAL PAPER

Digital transformations and the archival nature of surrogates

Paul Conway

Abstract Large-scale digitization is generating extraordinary collections of visual and textual surrogates, potentially endowed with transcendent long-term cultural and research values. Understanding the nature of digital surrogacy is a substantial intellectual opportunity for archival science and the digital humanities, because of the increasing independence of surrogate collections from their archival sources. The paper presents an argument that one of the most significant requirements for the long-term access to collections of digital surrogates is to treat digital surrogates as archival records that embody traces of their fluid lifecycles and therefore are worthy of management and preservation as archives. It advances a theory of the archival nature of surrogacy founded on longstanding notions of archival quality, the traces of their source and the conditions of their creation, and the functional "work of the archive." The paper presents evidence supporting a "secondary provenance" derived from re-digitization, re-ingestion of multiple versions, and de facto replacement of the original sources. The design of the underlying research that motivates the paper and summary findings are reported separately. The research has been supported generously by the US Institute of Museum and Library Services.

Keywords Large-scale digitization · Preservation repositories · Archival quality · Surrogacy · Digitization

Introduction

Digitization of the photographs, archives and manuscripts, maps, and books that comprise the contents of cultural heritage organizations is now a ubiquitous activity.

P. Conway (✉)
University of Michigan School of Information, 4427 North Quad, 105 S State Street, Ann Arbor, MI 48109, USA
e-mail: pconway@umich.edu

Published online: 20 April 2014

Archives are either actively digitizing their collections (as bitmaps and/or computer-readable text) or desiring to do so. It is becoming increasingly clear that if information from analog sources is not readily available in digital form, it simply does not exist from the perspective of the vast majority of potential users.

Digitization establishes the affordances of transformative access. If we define this type of access as "widespread, open, and universal availability through the Internet," then the underlying technologies that make possible the creation, dissemination, and management of digital surrogates also have the power to transform lives. The convenience and efficiency of access to digital surrogates create a lively and interactive communication between the evidence of our past and our present human condition, as well as with our hopes and aspirations for the future. We are beginning to experience digitization on a scale that may yet foster a revolutionary reversal of our collective disdain for the past that seems overwhelmed by the latest game, by the next big thing, and by the flood of bits and bytes.

Today, we have reached a new threshold where the preservation of digital surrogates is an increasingly urgent responsibility for archivists and librarians (Tibbo 2003). This article will argue—and provide some supporting evidence—that one of the most significant requirements for the long-term care of collections of digital surrogates is to respect these collections as archives in their own right, worthy of management, and maintenance as a record of their creation, organic existence, and use. The rhetorical argument consists of three steps: (1) contextualizing digitization in the debate over preservation and access; (2) articulating the components of archival theory that generate a hypothesis on the archival nature of surrogates; and (3) providing some evidence for testing these hypotheses. The reasoning here is a work in progress, only partially formed, but it is offered in an effort to stimulate debate within the archival science community on an issue that has been underexplored in spite of the increasing presence of transformed cultural heritage collections.

Digitization, preservation, and access

Digitization of books by third-party organizations such as Google and the Internet Archive is generating extraordinary collections of visual and textual surrogates. A surrogate is something that stands in for or takes the place of something else, in this case the original source. The preservation of the surrogates that result from large-scale digitization is premised partly upon their long-term cultural and research value, rather than on their distinctive qualities, which may fall short of the standardized ideal. For example, the HathiTrust Digital Library found its initial mission in the preservation of digital surrogates, rather than in providing access to them (York 2008). The currency of digital preservation in HathiTrust is the copy, rather than the born digital source.

The creation of digital surrogates from archival sources is fundamentally a process of representation, far more interesting and complex than merely copying from one medium to another. Theories of representation—and the vast literature derived from them—are at the heart of many disciplines' scholarship and of

particular relevance for scholars who work primarily or exclusively in the digital domain. Mitchell defines representation through signs or symbols as a mediated relationship between the maker and the viewer of one object that stands for another. "Representation is always of something or someone, by something or someone, to someone" (Mitchell 1990: 12). To Mitchell, representation is an intentional relationship between the maker and the viewer, fraught with the potential for communication problems ranging from misinterpretation and error to falsehood and forgery. Scholars from a wide variety of disciplines have explored how theories of visual representation bear upon the creation and use of digital collections whose origins are in photography (Scruton 1981).

Geoffrey Yeo adapts the considerable scholarship on representation in his innovative two-part work on the nature of the archival record. In a footnote, he cites 32 widely varied and compelling publications across philosophy, linguistics, art, and six other academic disciplines, including interdisciplinary studies itself (Yeo 2007: 334). According to Yeo, most definitions of "archival records" in the professional and scholarly literature have descriptive aims; they attempt to delineate the nature of records as archivists perceive it as a way of communicating ideas within the professions, not necessarily as a way to communicate with those who encounter archives in any form. Yeo postulates that information products constructed from archival sources carry with them their archival nature and exist as "persistent representations of an occurrent" (Yeo 2008: 136), by which he means *both* the archival nature of the source and the temporal activities that transform them into another manifestation of archive. Persistence, in Yeo's view, encompasses the intentional acts of preservation as well as the recording technologies that affect a representational transformation.

From this perspective, surrogates created through digitization processes carry their own forms of materiality across the "axis of representation" (Mitchell 1990). Digital humanities scholar Matthew Kirschenbaum (2003, p. 146) reminds us that the ofttimes eloquent transformation of published books to digital code and algorithm "are themselves always subject to the functional constraints imposed by the material variables of computation. Understood at this level," he writes, "digital surrogates are just as 'real' (and tangible) as their analog counterparts." The relationship between source and digital surrogate conforms to the "law of contact" proposed by the Australian anthropologist Michael Taussig (1993, p. 52), who writes in the context of the effigy that "things which have once been in contact with each other continue to act on each other at a distance after the physical contact has been severed." This dynamism of meaning making, embedded in the act of reproduction, extends outward through the production of the digital object at the precise moment of reader contact, the phenomenon that Drucker (2013, para. 11) has labeled "performative materiality." "This shift from an approach grounded in what something *is* to how something *works* [leads] us into the lifecycle of production, use, control, resource consumption, labor, cost, environmental impact and so on—so that an artifact's materiality is read as a snapshot moment within continuous interdependent systems."

Digital surrogates produced through high-volume copy-making contain traces of the circumstances of their creation. The notion that the "trace" is capable of

simultaneously providing evidence and highlighting absence or loss is a powerful metaphor that resonates with interdisciplinary scholars who grapple with the archive as a place of remembering and forgetting. German media theorist and historian Jens Ruchatz writes that because traces are generated unintentionally they are particularly authentic and trustworthy testimonies. Ruchatz argues that recognizing the trace is a form of decoding. "Making sense of a trace is to take it as evidence of what is shown on it and to reconstruct the situation of its origins" (Ruchatz 2008, p. 370). Digital images produce an exceptional class of traces: "They show—but do not explain—what has caused them," writes Ruchatz. Once read, however, traces may inevitably affect the trust that is essential to the acceptance of digital surrogates as sources of scholarship. Melissa Terras, a leading digitization theorist, draws out the important implication of ignoring the deeper meanings of production traces: "If we cannot trust our means of reproduction of images of texts, can we trust the readings from them?" (Terras 2011, p. 43).

York University professor of English Marcus Boon, however, argues persuasively that a philosophical and theoretical impasse has resulted from the efforts of post-modern critical theorists, ranging from Gilles Deleuze and Jacques Derrida to Michael Foucault and Jean Baudrillard, to deconstruct the distinctions between the original and copies in terms of traces or trust. Instead, he builds on anthropological Taussig's insights about contact and draws on Buddhist philosophical perspectives by focusing on bonding in the act of copying. Boon writes that "bonding indicates a set of intentions, practices, and structures that work to produce the experience of subjective and objective things, including copies" (2010, p. 33).

Other scholars argue for the fundamental difference in the digital copy/surrogate. Media theorist Wolfgang Ernst associates permanency with constant change in suggesting that "the digital archive itself has become an entity always already in flux, continuously in-formation, and its analysis requires new conceptual tools" (Ernst 2013, p. 42). University of Glasgow social scientist Andrew Hoskins establishes this dynamic archive as a new memory space, free from both spatial and institutional constraints. Reinforcing Kirschenbaum's digital materialism, Hoskins writes that "the traditional materiality associated with the artefactual archive has been challenged by the fluidity, reproducibility, and transferability of digital data." Hoskins shows how the networked digital archive has become "a key strata of our technological unconscious," by which he and other media scholars mean "the everyday habits initiated, regulated, and disciplined by multiple strata of technological devices and inventions" (Hoskins 2009, p. 97).

The distinction between digitizing for access and digitizing for preservation, so deeply embedded in the professional perspectives of archivists and librarians, is artificial and misleading. In the digital world, access is the natural and obvious outcome of digital transformation, even if access is fully realized only through functioning electronic networks and the legal frameworks that manage permissions. It is very important to separate this potential for widespread, open, and universal access from the barriers that are often imposed by the international intellectual property law. In a networked environment, access is a fixed state of digital transformation, whereas intellectual property regimes are malleable, subject to resistance, and ridden with loopholes and exceptions. Access is free as in "free

speech, not free beer," is not limited by time or space, and is never fully subverted by legal constraints (Raymond 1999).

Because access is, therefore, a given in digitization practice, preservation becomes the measure of the value that archivists place on the capital and labor of their digitization efforts. Such a choice to assign worth is what Ketelaar (2001) terms "archivalization." The processes of large-scale digitization, however, defy the decades long tendency of information technology to substitute capital equipment for skilled labor (Brynjolfsson 1993). Digitization is not now nor will it ever be a fully automated process. Indeed, today's large-scale digitization programs are relentlessly manual processes that engage a new class of "information workers," not unlike the factory operations that fueled the industrial revolution of the nineteenth and twentieth centuries and that continues today in the production of clothing in Bangladesh, toys and electronics in China, and most of the twenty-first century's consumer product superstructure. Nevertheless, these large human-driven digitization efforts have tremendous capital value and generate huge investments in new information products. Two examples may suffice. First, in 2012, a London-based private equity firm purchased for $1.6 billion the content together with the customer base of the world's largest genealogical service, Ancestry.com Inc. (Bloomberg 2012). Second, conservatively estimating digitization production costs at $50.00 per book (University of Michigan 2001), it is fair to assume that Google has invested well over $1 billion of its profits in the digitization of 20 million books from the world's best research libraries. Leetaru (2008) estimates Google's production costs at $10.00 per volume but provides no basis for this figure. Neither Google nor the Internet Archive has released information about their investments in digitization. Bia et al. (2010) have developed the richest digitization cost model designed to measure high production processes in a research library setting.

At any scale, large or small, a commitment to preserve collections of digital surrogates represents a decision that the value of the investment in their creation is not temporally bound. Access to digital surrogates generates the need for preservation because, over time, users shift their perspective and their inquiry tactics from original sources to digital surrogates (Conway 2010). Digital access creates new dependencies. When demand migrates to digital resources, users will rarely, if ever, return to the original source (Hirtle 2002). Large-scale digitization by commercial vendors has generated new demands and expectations that ALL archival material should be digitized and put online. Archival organizations that do not have the resources to meet this demand may find themselves locked in a three-way dilemma: marginalized because their collections are not accessible in digital form, impoverished through the reallocation to digitization of fixed or declining resources, or outsourced to sources and sites that can deliver acceptable digital content. The point is not to bemoan the costs and challenges of digitization for libraries and archives but rather to highlight the absolute value of transformative digitization and, by implication, the extraordinary and continuing investment value of the resulting products.

Surrogates and the nature of the archive

It is now clear that large-scale digitization of cultural heritage resources is a complex development with tremendous impact not just on those organizations that manage collections of digital surrogates but also on the underlying theories that govern the management of these resources. The key to examining digital surrogacy in the context of archival thought is acceptance of the proposition that archives are social constructs whose trace meanings change over time as they are described, transformed, and used. Nesmith (2002) provides an excellent introduction to this "post-modern" approach to archival theory, which recognizes the powers of remembrance and forgetting embedded in the procedures of archival management. Post-modern archival thinking is logical and compelling in the abstract and difficult to demonstrate in practice (Cook 2001).

Generations of archivists, beginning with Sir Hilary Jenkinson (Ellis, p. 197), have rejected the archival nature of surrogates, considering them "artificial collections" at least one step removed from the original source and therefore subject to even stricter tests of authenticity and reliability (Smith 1999). In defining the still-prevalent perspective on the nature of archives, Eastwood (2012, p. 7) notes that "the qualities of naturalness, interrelatedness, and uniqueness together constitute the core of the traditional organic concept of archives." By creating boundaries around the concept of the archive, traditional archivists must, almost by definition, exclude artificial collections, which Eastwood notes "have their own coherence dictated by the purposes for which or the circumstances in which they were formed; these determine the cast they have." My key theme is that in the environment of large, mutable digital collections, the boundaries between the "natural" archive and the "man-made" collection have blurred, such that the "organic whole" of a large collection of surrogates, built "in accordance with fixed rules," may indeed hold "archival qualities." What are the grounds for such an idea?

An answer starts with the foundations of archival science. In his seminal essay on the foundations of archival science, Dutch archival theorist Thomassen (2001, p. 383) asserts that the aims of archival science are "the establishment and maintenance of archival quality; that is to say: of the optimal visibility and durability of the records, the generating work processes, and their mutual bond." In archival science, archival bonds are not a fixed or static asserts Thomassen, but are always subject to breaks between form, structure and context, and the contents: "reliable becomes unreliable, high quality become low quality, archives to documentary collections, evidence turns to documentation, documents to loose data." To prevent the occurrence of such processes, writes Thomassen, "one has to maintain the relationship between content data on the one hand and the form, the structure and the context of the creation of these data on the other, or carefully document the changes that are made to it." Such an articulation of archival science theory places archival quality at the center of a comprehensive theoretical framework that demands accountability through documentation of process and procedure. Thomassen frees the analysis of the archives to encompass information resources that are not or never have been a part of a formal archive, including digital surrogates.

Although archival thinking had worked its way into thinking about digital preservation a decade earlier (Waters and Garrett 1996), Seamus Ross was the first scholar to make an explicit argument for applying archival practices to the management of digital library content. In his keynote to the European Conference on Digital Libraries, Ross (2007, p. 13) states that "if we think more carefully about digital libraries we easily observe that they may be libraries by name, but they are archives by nature." He uses the rich science of diplomatics to demonstrate how archivists might provide assurances regarding the authenticity and even the reliability of digital documents. Ross ends his exposition on the relevance of archival practices by calling for research that is at least as rigorous as research on archival appraisal. "Quality is a property of digital objects that needs attention alongside authenticity and reliability."

Canadian archivist Lori Podolsky Nordland (2004, p. 154) constructs a case study of how a particular record assumes new identities and new meanings as it is "interpreted, reinterpreted, and represented at different points in time." Drawing on and reflecting Hugh Taylor's (1987) insights on technologically driven change in the meaning of archives, Nordland writes that "with each '*transmedia shift*' [Taylor's term], new meanings or layers are added to the record's context and structure, in a continual evolution of the history of the record, even after it is 'fixed' in archival custody. In turn, the record is reinterpreted to suit the new media and that author's wishes." The "author" in this quote is the archivist/digitizer. Nordland explains how the application of digitization technologies augments the original provenance of the item with a *secondary provenance* [Nordland's term]. Another Canadian archivist and educator, Emily Monks-Leeson (2011, p. 56) also embraces "secondary provenance" for digitized collections presented on the Web. "As records take on new meanings and new contexts, understandings of provenance can shift to encompass not only the original contexts of creation … but also those new contexts to which records come to belong." University College London scholar Geoffrey Yeo (2009, p. 59) pinpoints the source of "secondary provenance" in the material–custodial history of archival collections. He argues that the interpretation of records is affected by the "previous selection and aggregation decisions" taken by both creators and custodians.

Utilizing a separate terminology, but one that has striking similarities to Kirschenbaum's and Drucker's theories of digital materiality, Dutch archivist Eric Ketelaar argues (2001, p.) that "every interaction, intervention, interrogation, and interpretation by creator, user, and archivist is an *activation* [emphasis added] of the record." Ketelaar asserts (2012, p. 29) that "each activation leaves fingerprints that are attributes to the archive's infinite meaning. The archive is therefore not static, but a dynamic open-ended process." In contemporary thinking, the idea of the "trace" is fundamental to understanding the archive as a malleable, unfixed, changeable artifact. Terry Eastwood asserts (2012) that "archival documents are traces of the past bearing witness to their creators and to the society they inhabited, the preservation and appreciation of these representations of the past constitute the goals of archival science, and the archivist is a participant in the construction of an historical discourse." Marshaling a wealth of evidence, Brothman (2002, p. 337) goes further in claiming that the evidence in archives is not of some immutable

truth. Instead, "evidence appears as traces that record creators unknowingly leave. Record creators cannot set out to knowingly produce traces. Evidence is discovered by knowing agents and is a matter of post hoc interpretation to serve specific interests and purposes."

If archives preserve "traces of thought, expression, and activity," as Brothman and other post-modernist theorists suggest, rather than fixed, immutable sources of truth, then this theory demands to be tested and made concrete. What follows are three sketches demonstrating significant differences that transformational traces render in versions of digital surrogates over time. The first involves evidence of production fixed within the page image. The second involves wholesale re-digitization of archival source materials. The third involves iterative re-ingest of digital surrogates. Each of these three examples deserves full treatment as a case study; this article simply teases out the essence of the story to support the larger argument that collections of digital surrogates are archival in nature because of the actual or potential evidence that they generate about the secondary provenance embedded in their custodial histories.

Re-production: traces of process

The first case of archival traces involves the evidence left in the surrogate of the processes of its production. Large-scale digitization is an intense combination of manual and machine-assisted processes. Leetaru (2008) provides the most detailed, but not particularly well-documented description of scanning operations managed by Google and the Internet Archive. In search of maximum efficiency (and hence low per-item production costs), third-party digitization practices take place in "information factories" staffed by people who are trained and managed to carry out their work in assembly line fashion. Digital images created in this way (exceeding 30,000 volumes per week) are then fed to high-capacity server and storage systems, where they are batch-processed in a complex, multistage workflow to extract computer-readable text and to create deliverables for the Web. The amount and detail of the information available about scanning and post-scan enhancements vary depending on the vendor. The Internet Archive, for example, tends to be open about its processes, while Google's procedures are more tightly held.

All factory-based manufacturing processes are imperfect by nature. The extent and effectiveness of a factory's quality control processes turn on three complex and interrelated factors: first, the inclination of the factory to recognize error; second, the feasibility of statistically oriented process controls; and third, the ability of the factory to do anything about the flaws in its product that they find at any point in the manufacturing process (Oakland 2008). High-volume digitization leaves in its wake traces of production processes. These traces take the form of visible anomalies, or error. Depending on its location, frequency, and severity, digitization error may have huge negative consequences on the viability of the product or may merely be an annoying feature that may deter intensive use (Conway 2011).

Traces of production processes, in the form of undetected or uncorrected error, are a common feature of large-scale digitization programs. In summary of research

on the distribution of error reported separately (Conway 2013), rigorous analysis of over 4,000 volumes chosen randomly from a population of 8 million volumes shows the very common prevalence of low-level error that does not affect readability but that may affect the acceptance of digitized books. More severe error that affects the readability, understandability, and usefulness of the digital surrogate occurs rarely (well under 1 % of examined page images) and usually randomly throughout a given volume, with three exceptions: foldouts, illustrations, and hands. Regarding foldouts, it is well known and frequently commented upon (McEathron 2011; Duguid 2007) that Google does not digitize folded maps, charts, and other materials in books, some of which is essential to the usefulness of the volume. Less studied, but easily perceived, is the generally poor reproduction of printed illustrations, which suffer from varying degrees and types of moiré patterns, false colorization, file format anomalies, and post-scan processing distortions.

Perhaps of greatest interest from an archival perspective is the prevalence in digital surrogates produced by Google (and likely by other large-scale digitization efforts) of transparent evidence of the human touch of scanning books that have nearly infinite complexity and Google's algorithmic efforts to remove this evidence. The liminal evidence of hands and fingers lingering in digitized books, as well as scanning and post-scan distortions, is the subject of a community Tumblr site, The Art of Google Books (2013), that presently (April 2014) contains over 2,500 categorized but decontextualized examples. An analysis of 511 page images where all or a large portion of a page is obstructed shows that 86.3 % of these images are of pages located either at the front or rear of a volume or on a book page with no text present. These locations are in areas of volumes that are not likely to be surveilled by Google's automated text-processing procedures and may represent a subtle form of resistance to the relentless flow of pages under the scanner's eye. Figures 1 and 2 are examples of visual evidence of the manual processes of scanning: one a full hand on book end papers and the other a pink-tipped finger securing an illustration. The presence of hands may indeed be a form of digital "signature" by information factory workers, much as artists, crafts persons, carpenters, and steel workers leave their identities on their products in unobtrusive places. One avenue of future research is the dynamic relationship between the people who work in digitization factories and the machines and materials of their labor.

Re-digitizing images traces of change

The second case of potentially valuable archival traces in digital surrogacy involves re-digitization of source materials digitized at some point in the past.

The US Library of Congress holds in its Prints and Photographs Reading Room extraordinary and powerful visual collections that document the American experience. Approximately 95 % of the Division's 14 million items are described individually or collectively in an online catalog, which includes more than 1.2 million digitized images.[1] The online catalog itself (PPOC) is emerging as a

[1] Library of Congress. Prints and Photographs Online Catalog. http://www.loc.gov/pictures/.

Fig. 1 Scanner's hand on end paper (Van Denburg 1895, p. 380)

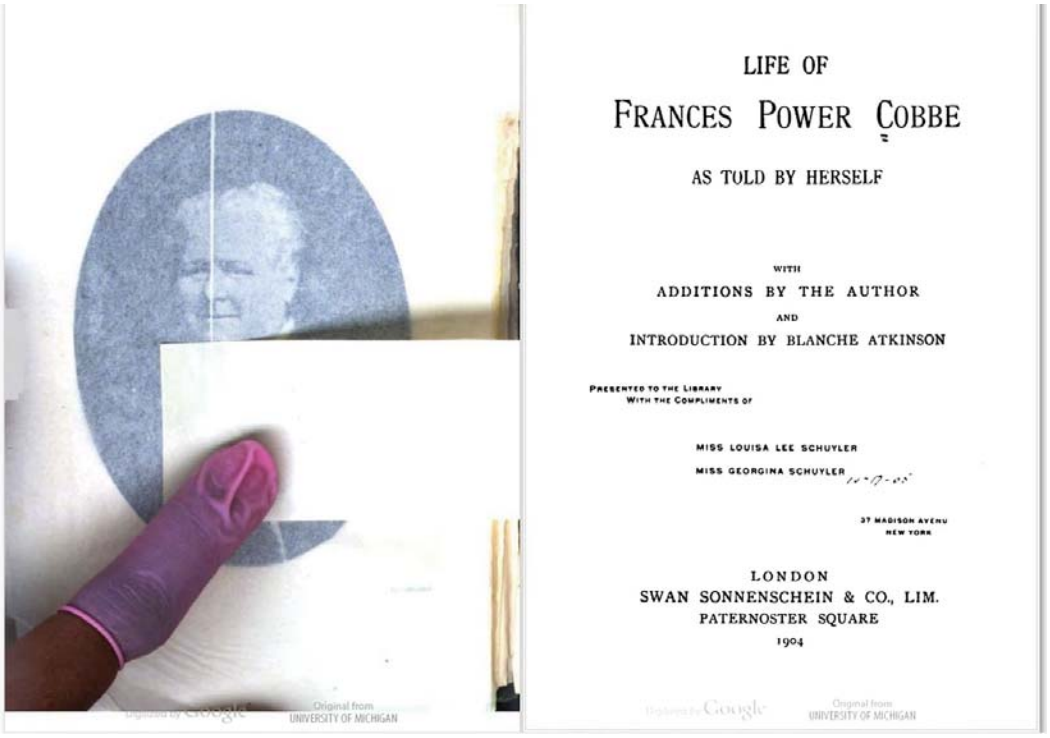

Fig. 2 Finger, tab, tissue, book illustration (Cobbe and Atkinson 1904, p. ii)

national treasure in its own right, providing widespread, open, and universal access, while preserving the digital surrogates in the Library's repository system.

One of the most powerful and extensively used collections in the online catalog is the digitized negatives of the Farm Security Administration and Office of War Information (FSA/OWI), forming an extensive pictorial record of American life between 1935 and 1944. The collection consists of about 175,000 black and white negatives and 1,600 color negatives (FSA/OWA n.d.b). The collection is available through an online catalog, as well as through the Library's American Memory Web site, and partially through the Library's Flickr Commons site and harvested by OAIster, using the OAI-PMH protocols.[2] In its original form as study prints and camera negatives, and now even more so as a comprehensive digital collection, the FSA collection is perhaps the most heavily used and most frequently cited resource in the Library of Congress (FSA/OWA n.d.c).

The currently available digital images are the product of successive efforts to apply digital imaging technologies with the goal of minimizing the handling of fragile and deteriorating nitrate and di-acetate negatives. In the early 1990s, working with 35 and 70 mm film intermediates, a contractor first produced a reference service videodisc (now obsolete). In the mid-1990s, a separate contractor reprocessed the videodisc images to create discrete digital images suitable for delivery through the Internet. The JPEG compressed images have a tonal resolution of 8 bits per pixel, and the color images have a tonal resolution of 24 bits per pixel.

In 2009, the Library developed a plan to rescan the collection to fully capture the subject content of each photograph, including the finest details and the full range of tones. The Library is creating these new digital images of the entire collection using recommendations of the Federal Agencies Digitization Guidelines Initiative (FADGI 2010). In 2010, re-digitization began for 90,000 nitrate negatives, starting with the 45,000 35-mm film frames. Each 35-mm frame is being digitized at a sampling frequency of 2,800 pixels-per-inch, 14 bits-per-pixel tonal range capture (available as 16-bit), and in uncompressed TIFF format (FSA/OWA n.d.a) (Fig. 3).

When the re-digitization project is complete, the new files will replace the old files, which may or may not be retained. Technical information about the re-digitization project, designed for the nonspecialist, is available as a separate, summary document associated with the overall collection. But technical metadata regarding re-digitization and replacement is not associated with individual images within the collection. The Library's re-digitization program sets a very important precedent for re-digitization of resources in response to user expectations, new technical specifications, and the marshaling of resources for activities once thought either impossible or unwise. The re-digitization project also highlights the value of digitization at the level of "full information capture" as a strategy to limit use of physical originals.

Wholesale re-digitization of photographic archives (or any other documentary source, for that matter) poses important issues for which archival science provides

[2] Library of Congress, Flickr Photostream. http://www.flickr.com/photos/library_of_congress/; Library of Congress, America from the Great Depression to World War II: Black-and White Photographs from the FSA-OWI, 1935-45. http://memory.loc.gov/ammem/fsahtml/fahome.html; OCLC, The OAIster Database. http://www.oclc.org/oaister.en.html.

Fig. 3 Library of congress display of image download options (Delano 1941)

some insight. First, the provenance—or chain of custody—of any given manifestation of a digital surrogate encompasses all of the previous versions that were publically accessible. In the case of the FSA collection, the truthfulness and trustworthiness of the interpretations made from digital surrogates turn on two factors: first, being able to see the version that was used to make the initial interpretation; and second, knowing the visual evidence revealed through re-digitization and contemporary image file manipulation. Bigger files may not necessarily be better files if delivering the new comes at the expense of undermining interpretations of previous versions. The simplest but most costly solution is to keep multiple versions, document the technical processes that produced the new version, and track the changes from one version to the next. For example, the PREMIS 2.0 metadata standard is capable of capturing and organizing documentation on change events, but mechanisms for exposing PREMIS data to the end user must be fully developed (Caplan 2009). This type of provenance documentation helps endow the surrogate collection distinctive archival properties.

Re-ingest: traces of content

The third case of variability and mutability in digital surrogacy involves the repeated re-ingesting of alternative versions of the same digital object. In large-scale digital preservation repositories that have established ongoing and long-term relationships with content providers, the content of the repository is less likely to be fixed in one-time deposits, but rather remain malleable and subject to changes over time that may or may not be immediately transparent to the end user. This is a significant break from past practices, where the preservation community exercised a form of vertical integration of digitization practice though the development and promulgation of best practices and a strong preference to keep scanning activities close at hand and under curatorial control (Kenney and Rieger 2000).

The HathiTrust Digital Library is a preservation repository that exemplifies this new style.[3] Its nearly eighty research library members have joined their resources, built a robust and sustainable digital storage and delivery platform, and established a governance structure with the stated mission "to contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge." This enormous commitment to a longstanding mandate of research libraries masks a simple reality: HathiTrust is now and is likely to be for the foreseeable future primarily a repository for one of the products of Google Books, Google's foray into the large-scale digitization of books and serials. HathiTrust now (Oct. 2013) contains well over 10 million digitized volumes, 97 % of which have been digitized by Google from the contents of at least 18 library collections (York 2010). The digital surrogates in HathiTrust encompass at least 429 languages across the spectrum of library classification and the history of books and printing since Gutenberg. In terms of size, the HathiTrust collection now ranks approximately 12th among the 126 members of the US Association of Research Libraries and ranks in the top 25 of the world's research libraries.

Google continuously adjusts the algorithms it uses to process the raw images it captures from library volumes and modifies the processing workflow in search of improved text capture from digitized page images. When new "improved" versions of volumes from Google are available, HathiTrust re-ingests them in batches of multi-hundred thousand volumes, after they pass a quality threshold test that largely focuses on the proper rendering of the image files and the completeness of the technical metadata. Ingest triggers a series of preservation events (e.g., fixity check, digest calculation, validation, and ingest) that are recorded as metadata conforming to the PREMIS metadata standard. All previous events are retained in the metadata record as traces of past activity, and thereby providing a means to determine whether and how many times an object has been ingested (York 2012). At ingest, the new version of the digital volume overwrites the existing version. The original version is not saved. More important, information about the differences between the overwritten version and the newly displayed version is not recorded in HathiTrust because such technical metadata is not supplied by Google in its ingest package.

[3] HathiTrust Digital Library. http://www.hathitrust.org/about.

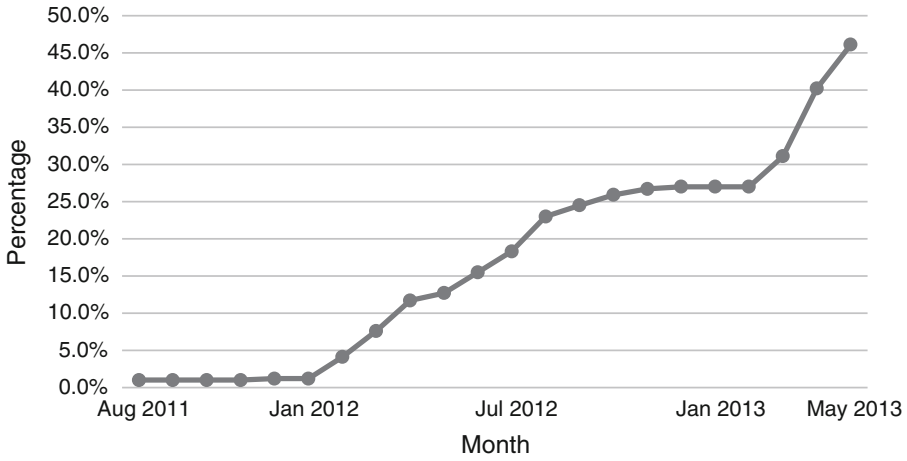

Fig. 4 Cumulative proportion of 1,000 volume sample re-ingested by HathiTrust, Aug 2011 to May 2013

HathiTrust uses this practice primarily because of the immense scale of the repository and the frequency with which volumes from Google are reprocessed and made newly available. Figure 4 is a display of the extent and pace of re-ingest of a random sample of 1,000 volumes from the HathiTrust collection. The chart shows that between August 2011, when the sample was drawn for quality analysis, and May 2013, nearly half (46.1 %) of the original volumes in the analysis sample had been replaced with updated versions, through a nearly continuous and ongoing series of re-ingest events. Three of the 1,000 volumes in the sample were re-ingested twice between the point of quality analysis and May 2013.

The cumulative effect of re-ingest on this scale is the removal of certainty regarding the appearance of any given volume's content. Such uncertainty threatens to undermine the authenticity that is a fundamental property of a protected archive. A digital book viewed at one point in time may not be the same representation weeks, months, or years later. In lieu of stability at the level of visible content, HathiTrust provides a persistent link to a digital object (book or page) whose bibliographic characteristics are relatively fixed, even if the content behind the link is mutable. Additionally, HathiTrust preserves (but does not display) the traces of change events to the digital object in the form of preservation metadata. Such traces indicate the time and place of change without documenting the specific nature of the change. Although the primary-stated purpose of re-ingest is to capture improvements in the re-processing of digital image files to improve quality, neither the specific aspects of algorithmic improvement nor the impact on the visual properties of the new product is known. Indeed, it is possible that re-processing and re-ingest remove one set of problems and leave new errors in their place. Without manual inspection of complete volumes or automated tracking of image processing modifications, it is impossible to certify the absolute quality or specify the adequacy of a given volume in a population that now approaches 11 million volumes. Proposals to mark preserved digital objects with a seal of quality will have to contend with the fluidity of a repository's content (Jacobs and Jacobs 2013).

Concluding thoughts

The focus of the argument in this article has been on the implications of the decision to capture and preserve the products of the large-scale digitization of books and photographs. The scope of the argument presented here is largely limited, for the time being, to archival surrogacy that emerges when there is a distinctive, transparent, and documentable connection between source documents and digital copies that themselves are subjected to the mediating forces of imaging technologies, routinized physical manipulation by human agents, and image file processing algorithms. Although some or all of these mediating forces may also exert themselves in the context of born digital archival records, the case for the archival nature of surrogacy turns on the unintentional traces left, accumulated, and internalized as a part of analog-to-digital transformational processes.

The case for the archival nature of surrogates produced through large-scale digitization is a special case of archives formation that may extend just as well, perhaps even more so, to the enormous collections of digitized archival records, photographs, maps, manuscripts, and other cultural heritage resources that are beginning to swell digital libraries today. Repositories of digitized archival records may not, on the surface, experience the levels and complexity of re-digitization, re-ingest, and re-production that are clearly associated with large-scale digitization of books. But behind the scenes, in server rooms and on the desktops of systems administrators, "artificial" digital collections are organic entities that grow and change their shape as new materials are added, new contextual relationships are established among objects, and new procedures are brought to bear on the organization and management of these large collections.

The evidence is accumulating that digital preservation is a worldwide problem that can be solved for the vast majority of the most common data formats, bitmaps among them. And yet, the evolution of digital preservation repositories over the past fifteen years has instilled in archivists and librarians a belief that to preserve is to fix digital content in a state where managed change is possible (Levy 1994). Metadata schemes, models of process control, and procedures for establishing and certifying the trustworthiness of digital repositories each in their own way contribute to an emerging confidence that the long-term management of digital content at scale is desirable and technically feasible. In particular, when information is born digital, no longer needed for its original purposes, appraised for continuing value, and transferred physically to its new home on archival servers, then the traditional concept of archival records will serve as an appropriate guide for managing this digital information over time. Archival science and records management conventions provide the theoretical muscle and practical guidance required to triage digital collections to identify those with enduring value, short- and long-term use, and the technical and economic feasibility.

In contrast, however, the emergence of very large collections of surrogates produced through digitizing the cultural heritage challenges the assumption of controlled and managed stability. At the level of the digital document, change is the norm. This change takes the form of re-digitization of sources or the reprocessing of

archival masters in response to new user expectations and improvements in the technologies of digital transformation. This change is also cultural, for with the ubiquitous access and overwhelming volume of digital content comes new ways of reading (distant and close), new interfaces with the past, and a kind of instantaneous knowledge that challenges our ability to pause for critical interpretation.

The argument in this article asserts that the management of large-scale collections of digital surrogates requires new archival thinking applied to resources whose value derives firstly from their association with original source materials, such as books and photographs. The technical quality of the digital content and its general association with a valuable source, however, are not sufficient grounds to ensure the survival of extraordinarily useful collections of digital surrogates. It is also necessary to recognize and protect the distinctive archival traces that derive from transformation processes and active management in preservation repositories. The more that archival science scholars can understand and articulate the terms of creation and the extent of error, changes, and stability over time of large, third-party digital collections, the more likely it is that these collections will support innovative uses. The fundamental principles that govern the building, managing, and use of archival resources are the same principles that endow the products of scholarship from these resources with trust, authenticity, and reliability. Trust through rigorous process control and the validation of authentic resources have been and continue to be the currency of archivally based scholarship.

In procedural terms, the question remains about what it means, in practice, to respect collections of digital surrogates as archives in their own right, worthy of management and maintenance as archives. Answering this question may have at least five components. First, archivists should recognize the agency that they exercise in each and every process they undertake to digitize and deliver archival resources. Second, archivists must record the "change events" that have even the potential to affect how end users interpret or re-interpret what they see and understand in the digital surrogate. Third, archivists should expose the metadata that contains this record of "secondary provenance" in understandable and intuitive ways. Fourth, archivists owe it to their constituencies to increase communication about the fluidity of digital preservation repositories, even at the risk of revising the terms of trust certification. Fifth, archivists should engage in ongoing conversations with the communities of users that have the largest stake in the trustworthiness of digital preservation repositories.

As large collections of digital surrogates begin to fulfill their societal promise through reliable access and the confidence that flows from the commitment to preservation, they assume a new and transformative value that is separate from the original source materials. As these digital archives live and age, they acquire a form of organic naturalness, including errors and anomalies, that endows them with a valuable and distinctive secondary provenance whose authenticity, reliability, and integrity must be protected and communicated to present and future generations of users.

Acknowledgments The ideas in this article were presented initially at the 5th Conference on Archival Databases about Archival Information in Rio de Janeiro, Brazil, June 4, 2013. A substantially revised

version was delivered as the 2013 Hilary Jenkinson Memorial Lecture at University College London, September 25, 2013. The Institute for Museum and Library Services provided support for the underlying research represented in this article.

References

Art of Google Books (2013) http://theartofgooglebooks.tumblr.com/. Accessed 20 Mar 2014

Bia A, Muñoz R, Gómez J (2010) DiCoMo: the digitization cost model. Int J Digit Libr 11(2):141–153

Bloomberg (2012) Permira agrees to buy ancestry com. 22 Oct 2012. http://www.bloomberg.com/news/2012-10-22/permira-agrees-to-buy-ancestry-com-for-about-1-6-billion.html. Accessed 20 Mar 2014

Boon M (2010) In praise of copying. Harvard University Press, Cambridge

Brothman B (2002) Afterglow: conceptions of record and evidence in archival discourse. Arch Sci 2:337–338

Brynjolfsson E (1993) The productivity paradox of information technology. Commun ACM 36(12):66–77. doi:10.1145/163298.163309

Caplan P (2009) Understanding PREMIS. Library of Congress Network Development and MARC Standards Office, Washington, DC http://www.loc.gov/standards/premis/understanding-premis.pdf. Accessed 20 Mar 2013

Cobbe FP, Atkinson B (1904) Life of Frances Power Cobbe as told by herself. Swan Sonnenschein & Co, London. Original from University of Michigan. HathiTrust handle: http://hdl.handle.net/2027/mdp.39015005338747?urlappend=%3Bseq=10

Conway P (2010) Modes of seeing: digitized photographic archives and the experienced user. Am Arch 73(2):425–462

Conway P (2011) Archival quality and long-term preservation: a research framework for validating the usefulness of digital surrogates. Arch Sci 11(3):293–309. doi:10.1007/s10502-011-9155-0

Conway P (2013) Preserving imperfection: assessing the incidence of digitization error in HathiTrust. Preserv Digit Technol Cult 42 (1): 17–30. http://hdl.handle.net/2027.42/99522

Cook T (2001) Fashionable nonsense or professional rebirth: postmodernism and the practice of archives. Archivaria 51:14–35

Delano J (1941) A prisoner dancing while another plays the guitar at a prison camp, Greene County, Georgia. Library of Congress, U.S. Farm Security Administration/Office of War Information Black & White Negatives http://hdl.loc.gov/loc.pnp/fsa.8c29073

Drucker J (2013) Performative materiality and theoretical approaches to interface. Digital Humanit Q 7(1). http://www.digitalhumanities.org/dhq/vol/7/1/000143/000143.html. Accessed 20 Mar 2014

Duguid P (2007) Inheritance and loss? A brief survey of Google Books. First Monday 12(8). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/1972/1847. Accessed 20 Mar 2014

Eastwood T (2012) A contested realm: the nature of archives and the orientation of archival science. In: Eastwood T, MacNeil H (eds) Currents of archival thinking. Libraries Unlimited, Santa Barbara, pp 3–21

Ellis R, Walne P (eds) (2010) Selected writings of Sir Hilary Jenkinson. Society of American Archivists, Chicago

Ernst W (2013) Digital memory and the archive. Univ of Minnesota Press, Minneapolis

FADGI (2010) Technical guidelines for digitizing cultural heritage materials. Federal Agencies Digitization Guidelines Initiative, Still Image Working Group. http://www.digitizationguidelines.gov/guidelines/digitize-technical.html. Accessed 20 Mar 2014

FSA/OWA (n.d.a) Digitizing the collection. Library of Congress, Farm Security Administration/Office of War Information black-and-white negatives. http://www.loc.gov/pictures/collection/fsa/digitizing.html Accessed 20 Mar 2014

FSA/OWA (n.d.b) Background and scope. Library of Congress, Farm Security Administration/Office of War Information black-and-white negatives. http://www.loc.gov/pictures/collection/fsa/background.html. Accessed 20 Mar 2014

FSA/OWA (n.d.c) Selected bibliography and related resources. Library of Congress, Farm Security Administration/Office of War Information black-and-white negatives. http://www.loc.gov/pictures/collection/fsa/bibliography.html. Accessed 20 Mar 2014

Hirtle PB (2002) The impact of digitization on special collections in libraries. Libr Cult 37(1):42–52

Hoskins A (2009) Digital network memory. In: Erll A, Rigney A (eds) Mediation, remediation, and the dynamics of cultural memory. de Gruyter, Berlin, pp 91–106
Jacobs JA, Jacobs JR (2013) The digital-surrogate seal of approval: a consumer-oriented standard. D-Lib Mag 19(3/4). http://www.dlib.org/dlib/march13/jacobs/03jacobs.html. Accessed 20 Mar 2014
Kenney AR, Rieger O (2000) Moving theory into practice: digital imaging for libraries and archives. Cornell University, Ithaca
Ketelaar E (2001) Tacit narratives: the meanings of archives. Arch Sci 1:143–155
Ketelaar E (2012) Cultivating archives: meanings and identities. Arch Sci 12(1):19–33
Kirschenbaum M (2003) The word as image in an age of digital reproduction. In: Hocks M, Kendrick M (eds) Eloquent images: word and image in the age of new media. MIT Press, Cambridge, pp 137–156
Leetaru K (2008) Mass book digitization: the deeper story of Google Books and the Open Content Alliance, First Monday 13(10), 6 October 2008 http://www.firstmonday.org/ojs/index.php/fm/article/view/2101/2037. Accessed 20 Mar 2014
Levy DM (1994) Fixed or fluid? Document stability and new media. ECHT '94 proceedings of the 1994 ACM European conference on hypermedia technology, pp 24–31
McEathron S (2011) An assessment of image quality in geology works from the HathiTrust Digital Library. Proc Geosci Inform Soc 41 http://hdl.handle.net/1808/8301
Mitchell WJT (1990) Representation. In: Lentricchia F, McLaughlin T (eds) Critical terms for literary study. University of Chicago Press, Chicago, pp 11–22
Monks-Leeson E (2011) Archives on the internet: representing contexts and provenance from repository to website. Am Arch 74(1):38–57
Nesmith T (2002) Seeing archives: postmodernism and the changing intellectual place of archives. Am Arch 65(1):24–41
Nordland LP (2004) The concept of 'Secondary Provenance': reinterpreting Ac ko mok ki's map as evolving text. Archivaria 58:147–159
Oakland JS (2008) Statistical process control, 6th edn. Elsevier Butterworth-Heinemann, London
Raymond R (1999) The cathedral and the bazaar. O'Reilly Media, Sebastopol
Ross S (2007) Digital preservation, archival science and methodological foundations for digital libraries. Keynote address at the 11th European conference on digital libraries (ECDL), 17 Sep 2007, Budapest, p 13
Ruchatz J (2008) The photograph as externalization and trace. In: Erll A, Nunning A (eds) Cultural memory studies: an international and interdisciplinary handbook. Walter de Gruyter, Berlin, pp 367–378
Scruton R (1981) Photography and representation. Crit Inq 7(3):577–603
Smith A (1999) Why digitize? Council on Library and Information Resources, Washington, DC
Taussig M (1993) Mimesis and alterity: a particular history of the senses. Routledge, London
Taylor H (1987) Transformation in the archives: technological adjustment or paradigm shift? Archivaria 25:12–28
Terras M (2011) Artefacts and errors: acknowledging issues of representation in the digital imaging of ancient texts. In: Fischer F, Fritze C, Vogeler, G (eds) Kodikologie und paläographie im digitalen zeitalter 2/Codicology and paleography in the digital age 2. Books on Demand, Norderstedt, Germany, pp 43–61. http://discovery.ucl.ac.uk/171362/. Accessed 20 Mar 2014
Thomassen T (2001) A first introduction to archival science. Arch Sci 1(2):373–385
Tibbo H (2003) On the nature and importance of archiving in the digital age. Adv Comp 57:1–67
University of Michigan (2001) Assessing the costs of conversion: making of America IV: the American voice 1850–1876. University of Michigan Library, Digital Library Production Service, Ann Arbor, MI http://www.lib.umich.edu/files/services/dlps/moa4_costs.pdf. Accessed 20 Mar 2014
Van Denburg MW (1895) A homœopathic materia medica on a new and original plan. Pub. by the author, Fort Edward, NY. HathiTrust handle. http://hdl.handle.net/2027/mdp.39015020206036?urlappend=%3Bseq=380
Waters DJ, Garrett J (1996) Preserving digital information: report of the task force on archiving digital information. Commission on Preservation and Access, Washington, DC
Yeo G (2007) Concepts of record (1): evidence, information, and persistent representations. Am Arch 70(2):315–343
Yeo G (2008) Concepts of record (2): prototypes and boundary objects. Am Arch 71(1):118–143
Yeo G (2009) Custodial history, provenance, and the description of personal records. Libr Cult Rec 44:59–60

York J (2008) This library never forgets: preservation, cooperation, and the making of the HathiTrust Digital Library. In: Proceedings of archiving 2008, 24–27 June 2008, Society for Imaging Science & Technology, Bern, Switzerland, pp 5–10

York J (2010) Building a future by preserving our past: the preservation infrastructure of the HathiTrust Digital Library. 76th IFLA general congress and assembly, 10–15 August 2010, Gothenburg, Sweden

York J (2012) A preservation infrastructure built to last: preservation, community, and HathiTrust. In: Proceedings of UNESCO memory of the world: digitization and preservation, 24–26 September 2012, Vancouver, BC, Canada

Paul Conway is associate professor at the University of Michigan School of Information. He teaches courses on digitization, preservation, archives, and the ethics of new technologies. His research encompasses the digitization of cultural heritage resources, particularly photographic archives, the use of digitized resources by experts in a variety of humanities contexts, and the measurement of image and text quality in large-scale digitization programs. He is a pioneer in charting the challenges and opportunities that digital information technologies present to preservation and archival science. He has extensive administrative experience in the cultural heritage sector and has made major contributions over the past 30 years to the literature on archival users and use, preservation management, and digital imaging technologies. He has held positions at the National Archives and Records Administration (1977–1987; 1989–1992), the Society of American Archivists (1988–1989), Yale University (1992–2001), and Duke University (2001–2006). He holds a Ph.D. from the University of Michigan. In 2011, he was awarded the Provost's Teaching Innovation Prize for his use of social media to teach undergraduate writing on ethics and technology. In 2005, he received the American Library Association's Paul Banks and Carolyn Harris Preservation Award for his contributions to the preservation field. He is a Fellow of the Society of American Archivists.