

Proceedings from the Document Academy

Volume 6
Issue 1 *Proceedings from the 2019 Annual
Meeting of the Document Academy*

Article 8


2019

Documentary Provenance and Digitized Collections: Concepts and Problems

Mats Dahlström
University of Borås, mats.dahlstrom@hb.se

Joacim Hansson
Linnaeus University, joacim.hansson@lnu.se

Follow this and additional works at: <https://ideaexchange.uakron.edu/docam>

 Part of the [Cataloging and Metadata Commons](#), [Digital Humanities Commons](#), and the [Museum Studies Commons](#)

Please take a moment to share how this work helps you [through this survey](#). Your feedback will be important as we plan further development of our repository.

Recommended Citation

Dahlström, Mats and Hansson, Joacim (2019) "Documentary Provenance and Digitized Collections: Concepts and Problems," *Proceedings from the Document Academy*. Vol. 6 : Iss. 1 , Article 8.

DOI: <https://doi.org/10.35492/docam/6/1/8>

Available at: <https://ideaexchange.uakron.edu/docam/vol6/iss1/8>

This Conference Proceeding is brought to you for free and open access by University of Akron Press Managed at IdeaExchange@Uakron, the institutional repository of The University of Akron in Akron, Ohio, USA. It has been accepted for inclusion in Proceedings from the Document Academy by an authorized administrator of IdeaExchange@Uakron. For more information, please contact mjon@uakron.edu, uapress@uakron.edu.

Introduction

For decades, memory institutions such as libraries, archives and museums have been engaged in digitizing cultural heritage materials in their holdings (also in the form of large private-public partnerships such as Google Books). The projects usually result in image reproductions (scans, digital photographs), text transcriptions (OCR'd or manually keyed) and varying degrees and types of metadata (usually legal and/or rudimentary bibliographic metadata). Significant resources are being invested in digitization in the US as well in Europe, and a whole range of humanities research depends on having these digitized collections available. Further, in the digital humanities and in particular in digital scholarly editing, the reproductions produced by memory institutions are not only referred to, but incorporated as building blocks in the digital editions themselves. In such cases their role extends beyond mere illustrations accompanying a scholarly text transcription, to serve as research tools and as instruments for accountability and accessibility (Dahlström & Dillen, 2017). In working with institutionally digitized collections, it is necessary to keep track of what particular document is actually presented on the screen, and in what relation the digital reproduction stands to the physical source document. This is the problem of digital provenance.

In this article, we place the digitized and edited document and document collection as such at the centre of interest, by asking a number of basic questions, such as; if digitized reproductions are taken at face value in, for instance, historical or textual scholarship, does it matter? What are the consequences when the ties of provenance are difficult to identify or are broken? How might editing practices subsequent to image capture affect the authenticity, credibility and accuracy of the reproduction in relation to the source document, particularly if we have insufficient knowledge of the various stages of the editing process? These questions will be dealt with on a conceptual level in order to provide a basis for further empirical study on the problem of digital provenance.

The need for a critical approach

Digital reproductions are often uncritically taken at face value, as straightforward representations of source documents. In particular, this face value approach is sustained by mass digitization projects such as Google Books where there is little room for manual intervention during the image capture phase. As a result, image capture appears to be a fairly trivial and straightforward task that can be entirely subjected to automated procedures.¹

¹ The face value approach is apparent also in more high-quality and critical projects. In most Scandinavian digital scholarly editing projects for instance, the editing team is handed digital reproductions of source documents from e.g. a library without really questioning their

However, rather than providing an exact copy, the reproduction offers just a sample of the source document's characteristics. The selection may be the result of careful and deliberate consideration, but it might also be due to random or incidental factors, factors over which the digitizing staff does not have full control, or even understand. Further, there are in fact many types and levels of digitization at memory institutions, suggesting a map of variety with mass digitization in one corner and what has been termed critical (Dahlström, 2010) or slow (Prescott & Hughes, 2018) digitization in the other. The choices made in terms of preferred processes have been shown to matter to the way the memory institutions are conceived of (Dahlström, Hansson and Kjellman, 2012). They also affect the range of scholarly inquiry that can be made into these institutionally produced and, in terms of authenticity, sanctioned reproductions.

Digitization is a complex set of processes, ranging from pre-processing (planning, project design, selection), processing (image and text capture), post-processing (metadata, editing, versioning) up to publishing, archiving and long-term maintenance. For this to work, a range of practices and skills are required. Digitization therefore involves a number of professions making crucial decisions not only in a purely technical sense, but in terms of initiated critical analysis as well. Image capture in particular, along with subsequent post-processing and editing of the digital images, has the potential to make digitizing staff recognize that virtually all parameters in the process (image size, colour, granularity, bleed-through, contrast, layers, resolution etc.) require critical intellectual choices, interpretation, and manipulation. And if one looks closely at high-quality digital imaging projects in a library or museum, it is clear that teams of conservators, technicians, and photographic experts constantly make series of decisions informed by critical and bibliographical analysis and by a highly specialized knowledge of the graphic, historical, and other research related aspects of the object they are digitizing. Work comprises activities such as

- critical discrimination or collation between varying source documents,
- image editing and emendation,
- critically matching the reproduction to the source with respect to exhaustiveness and faithfulness, and
- producing large amounts of metadata, descriptive encoding and bibliographical information.

In doing so, scholarly work is embedded in the objects already at very early stages of the digitization process. Results diverge in appearance; there are well-

authenticity and provenance, i.e. what has been done with the image during the process, if the proportions, colours have been tampered with, and what specific physical copy (or copies) the reproduction stems from.

known examples of how documents from different geographical holdings can be digitally brought together in a virtual collection, such as *The International Dunhuang Project* (British Library, n.d.), and of how fragments of a single but scattered document have been digitally reunited in a virtual representation, as in the *Codex Sinaiticus* (n.d.). One can even conceive of a virtual ‘ideal copy’ of a source by critically amalgamating ‘best’ fragments digitized from different extant copies of an edition, akin to how classical textual criticism establishes a text through an amalgam of readings from several different versions - an early printed example would be Inger Bom’s 1974 edition of a 16th-century *Hortulus synonymorum* (see Kondrup, 2011, p. 68). And finally, as part of what might be seen as a pre-scholarly analysis of the documents, photographers in digitization staff regularly produce a number of versions of the facsimile (varying in colour, light, resolution, size, formats) to serve different aims, both internally within the library and externally. There is even room for creative work, since digitizing staff occasionally retouch images or strengthen their contrasts to convey the appearance of an original in better shape and readability than it actually is.

So there seems to be room for an increased critical understanding of digital reproductions as interpretations based on scholarly informed deliberation, or a ‘document criticism’ for digital image reproductions in the manner of how textual criticism has been established since centuries to establish the history, relation and provenance of texts and their versions.

Digital provenance

When a copy of a printed book is chosen as a source document, digitized and made available on the web, the relationship between this source and digital reproductions (scans, photographs) derived from it might seem to imply a simple linear relation. Digital culture however dissolves this linearity in more ways than one, and suggests spiral, recursive processes in place of linearity (Hillesund, 2005). As the contents of a digitized book have been shipped into the plural streams of the web therefore, the questions of which text, document, and display that constitutes originals and which constitute copies largely depend on which streams one is looking at, and on where in the recursive processes one starts looking. What is it that we see on the screen, and *of what* is it a digitized version?

A digital document reproduction not only carries an implicit and interpretable history of production in the form of its graphical and textual display (as printed objects do), but also an explicit documentation of its unique production, usage, and version history, embedded in its technical layers, metadata and paratexts. This does, of course, not just apply to textual documents. During image capture and processing, the image can be edited at bit level without a human eye being able to discern that a change has been made from one instance to the next. Our concept of authenticity is thus challenged, and perceived of as different in the case of digital photographs compared to the case of analogue photos. One might even suggest that digitization means that

our entire ‘truth contract’ towards images needs to be renegotiated. Photographers are beginning to embed so called family trees into digital images. In other words, they include metadata about the history, versions, and updates of the object in order to provide transparency and strengthen authenticity. The user is thus better equipped to discern the steps in the production process and the degree to which the image has been edited. In institutionally governed digitization projects, related tools would be the calibrating stick, the ruler, and the colour chart, which all enable the user to check the reproduction.

When defining the relation between a source document and its digital reproductions, whether as part of a collection or as a singular entity, the question of provenance comes to the fore, and in research it has been the subject of interest from a number of aspects. While the practical digitization process, as discussed above, contains a number of decisions critical to the quality of the reproduction, the present scholarly discussion progress a critical discussion of not just the production of facsimiles, but of how current processes and conventions determine our ability to use and understand both the digitized documents themselves, and the situations or conditions of which they bear witness. The range of these discussions is wide and stretches from the significance of physical location placements of works in relation to their location in digital collections in museums (Padfield *et al.*, 2019) to the question of what is lost in the digitization process through process standardization in collaborative digitization projects within the library and museum sectors (Kjellman, 2008). No matter how the discussion on the relation between the original source and its digital reproduction or the relation between various reproductions is dealt with empirically, recurring themes are authenticity, metadata and usability, all central to the problem of digital provenance. There is, to put it simply, a risk that the digitizing process as such, as described above, creates a distance between the original document and the digitized representation. This distance creates a problem of authenticity, and this in turn affect the reliability of the document. Duranti (1995) writes:

[A] record is authentic when it is the document it claims to be. Proving a record’s authenticity does not make it more reliable than it was when created. It only warrants that the record does not result from any manipulation, substitution, or falsification occurring after the completion of the procedure of creation, and that it is therefore what it purports to be. (Duranti, 1995, pp. 7–8)

In cultural heritage digitization however, manipulation and sometimes even substitution may be legitimized through the opportunities offered by specific tools and techniques to enhance aspects of the original, for instance through colour manipulation, multi-spectral imaging or 3D-scanning. This does, however, create a demand on an explicit and accessible account of the relation between the original and the manipulated reproduction in order to maintain, not only authenticity, but reliability as well. High-quality digital imaging in library

digitization and in digital scholarly editions should therefore ideally provide the user with links to the uncompressed raw files as they were prior to being manipulated and edited. It all comes down to a matter of trust. It is not always necessary to gain this from just links to raw files; in addition, a transparent account of the production history, versionality, technical parameters, and editing history of the image files is needed. These accounts also need to be complemented by the institutional authority of the digitizing organization. And in fact, for high-quality manuscript digitization performed in several steps with several derivatives along the way, what might be the desirable output is not just *a* digital facsimile but a whole *archive* of digital facsimiles of the source document (Prescott & Hughes, 2018).

Although demands and concerns such as these may seem peripheral or may rarely come up, they constitute important issues, particularly when we are addressing the relationship between originals and reproductions. When the relation is defined within the production process, and then connected to its institutional context, there needs to be consensus over how the connection between the original document, the digital reproduction, and the digitizing organization (a library, an archive, an EU funding body) should be defined. In the ideal case, this construct also creates the ability for the digital reproduction to connect to other documents and reproductions, perhaps even located in institutions of different kinds. A potential tool for this is metadata standards.

Metadata

Ascribing correct and sufficient metadata is a central part of the post-processing phase of digitizing. However, the role of established metadata standards such as FRBR or Dublin Core differ, as does the strength and type of bond between versions of a document defined by them through the document's institutional context. Archives and museums usually digitize single artefacts into coherent virtual collections that do not necessarily correspond exactly to its physical organization. Libraries, of course, do the same in the form of manuscript artefacts, but in terms of scale, their predominant digitized object is a printed document, the source of which is one copy among many, such as an edition. In FRBR terms, we are dealing with two levels: item (the single copy) and manifestation (the edition, of which the copy is a part). The metadata and description of the digital reproduction usually refers to the manifestation level, but what we see on the screen is the item level. Björk comments:

In most cases this conceptual gap is of no concern. But as the digitisation process produces representations that are used as information resources in their own right, while at the same time “referring back” to a source document, the resilience of this relation becomes increasingly important. (Björk, 2015, p. 163)

Tennis (2015) emphasizes the fact that there are different traditions, or metadata lineages, when describing documents in the archival, and the library

sector. This begs the question of how to migrate and harmonize these lineages to be able to construct cross-institutional projects such as the above-mentioned Dunhuang project, with a maintained provenance transparency. This calls for critical analysis of metadata practices and exposes a need for a separate analysis of metadata provenance, studying how index terms and other metadata elements live and change over time between different environments of standardization. Tomasi (2017) goes even further and suggests, by example of both structured collections and unstructured documents, that the building of metadata ontologies may be a way to capture the authoritativeness of a digitized document or edition in that it enables provenance preservation in a conceptual dimension.

As for recording the provenance of the physical manuscripts, there are several initiatives to come up with international standards. The TEI Header for digitised and encoded textual material offers the possibility to record provenance information, albeit in loose form, about the physical manuscript and its history, primarily in the <msDesc> (Manuscript Description) element and its sub-elements <acquisition> and <provenance>. Efforts are also being made to use the CIDOC-CRM model to map manuscript history or to use the Nodegoat data model to record and visualise the history of manuscripts (Burrows, 2017). But whereas efforts such as those aim to track the trajectory of the physical artefacts across history, geography and collections, what we are calling for are similar efforts to document the forward trajectories, from source documents, through possible intermediaries (such as microfilm) and up to the digital reproductions produced from them.²

Paradata

In parallel to the importance of relevant and transparent metadata, an interesting form of metadata is known as paradata. Paradata are data documenting the processes of how digital sets of data were collected and curated (Stieger & Reips, 2010), and are particularly important in computer game studies. Paradata can also document the context of not only the dataset's creation and development, but also of the decisions made during the process, the dataset's maintenance, life cycle and use. Their applicability could thus be far wider, not least in terms of data about states, production history, and digital provenance. More to the point, paradata can be used to document the process of digitizing and curating artefacts and documents, and the decisions made during a digitization project (Bentkowska-Kafel, Denard & Baker, 2012).

² A significant number of digitization projects are digitizing not the physical source documents, but intermediaries in the form of microfilm reels made decades ago, a second-hand transformation. Such microfilm reproductions are sometimes of quite poor quality, in black/white rather than colour, and inevitably with some degree of distortions or other technical artefacts stemming from the sequential historical conversions – distortions which are then inherited by the new digital collection (cf Kichuk 2007 or Mak 2014).

For some of the problems discussed elsewhere in the paper, for instance that the version history of the reproduction might be lacking, or that layers of bibliographic metadata, contextual information or text encoding have been changed or even lost due to generations of technical conversion of the digital collections, paradata might be a valuable key when ascertaining authenticity and usability. So for instance, a researcher working with a digitized manuscript might find it difficult to decide if two sections on a page with slightly varying colour are written in different ink, or if the varying colour nuances she sees on the screen is a technical artifact of the particular settings of light during the image capture. Likewise, warped letters in the margins might be a distortional effect from managing the object during image capture, or they might have occurred during earlier conversion processes - in cases such as this, a paradata documentation of the production and conversion processes that took place prior to the digital image the researcher sees on the screen, can help in providing some of the answers. And if paradata is used at the collection level, to document the processual changes that the collection has undergone subsequent to the original image capture and the decisions made during conversion and maintenance, we stand a better chance of further strengthening the transparency and authenticity of the reproductions and to provide keys for understanding the potentials and limits of their usability. And as mentioned earlier, knowledge and information accumulated during a digitization project tend to get lost, when the digitized collection is subjected to standardization and/or merging with other projects and collections. Documenting these changes through paradata increases our chances of rescuing such knowledge or at least being better able to ascertain the value and usability of the collection we are facing.

Nevertheless, although some of the more elaborate metadata schemes for encoded textual documents, such as TEI Headers, already include elements of paradata (e.g revision history), most metadata schemas, particularly for image material, do not.³ In such cases, this kind of paradata information might be documented and preserved elsewhere, as in project reports, about-pages, internal wikis or project evaluations, but far from all digitization projects engage in this sort of documentation and even fewer make it available online to end-users. A further circumstance impeding transparency into the creation and process history of a digitized collection is outsourcing, which makes it difficult or even impossible to provide or acquire a full account of what has been done to the materials during the historical cycles of processes. Jarlbrink and Snickars draw a grim conclusion:

³ IIIF (the International Image Interoperability Framework; <https://iiif.io>) is an emerging framework that might be used to support this form of paradata, if it succeeds at becoming fully embraced by the international memory institution community engaged in producing digital facsimiles.

In fact, being on location and studying the digitization process, it swiftly became apparent to us that sticking to any idea of a provenance chain is impossible, since so many steps have been outsourced to unknown factors affecting the source document. On the one hand, the sacrosanct software conditions output in concealed ways and on the other hand institutional factors do the same [...] They all affect the digitization process – in different and increasingly vague ways. In short, the further up these actors are in the digitization chain, the less they seem to know of the processes that turn papers into digital data. (Jarlbrink & Snickars, 2017)

Conclusions and further research

As suggested above, these kinds of problems affect not only our conception of the validity and authenticity of a digital reproduction, but they also affect the informative capacity of reproductions (for instance to serve in place of the source documents, or as enhancements of them, cf. Björk, 2015), their usability in a more general sense, and in prolongation, their re-usability, for instance as building blocks in subsequent projects such as digital scholarly editions or as source materials in their own right for historical studies.

The digital image reproduction invokes the virtual presence of the source, so the bond between reproduction and source is not only graphical and material but is also defined by a retrospective relationship between two points in history, the then and the now. What seems to be missing for many current reproductions in digitized collections is the historical-bibliographical link between, on the one hand, what we see on the screen and, on the other, a particular identified artefact in a physical collection. In other words, which document was actually used when producing a given digital reproduction?⁴

To learn more about this, further empirical research is needed. A relatively small amount of research has successfully explored the accuracy, usability and reusability of digital representations to humanities scholars, the kind of research questions they open up for, and what degree of authenticity and trust we are able to ascribe to them. Conway (2013) has for instance conducted valuable studies on rates of errors in large-scale digitization projects. A more specific set of inquiry concerns what it is that scholars are presented with on screen when using the digitized collections, and how their production history affects (and manifests itself visibly in) the appearance and quality of the collections. An early example is Kichuk's study (2007) on the layers of remediation in a large digitized collection such as the Early English Books

⁴ As a very simple example: a frequent option in digitized collections of printed books is to have access to both a digital facsimile of the source document and an OCR'd (and possible XML encoded) transcription of its text, sometimes even displayed synoptically side by side on the screen. Occasionally however, the text transcription and the digital facsimile stem from two different copies of the edition (or even from different editions).

Online (EEBO), which was followed up by Mak's (2014) extensive archaeological uncovering of historical layers and remediated 'noise' in EEBO. A more recent study has similarly explored the presence of errors and noise, resulting from the history and forms of digitization, in a large ongoing project to digitize Swedish newspapers (Jarlbrink & Snickars, 2017). But the user of such digitized collections stands a better chance of understanding and managing such errors and noise if the collection is transparent about its production history and, specifically, provides information about – and keys to – the historical *bond* between the digital reproductions and the objects they reproduce.

A potentially rewarding research avenue might thus be in the form of case-studying a variety of digitized collections to see to what degree - and in what form - the digitizing agents provide such keys for users to ascertain the provenance of the digital reproductions, and to explore the user needs and potentials of such keys and instruments. We believe that exhaustive paradata and metadata for the digital images might, as suggested above, be of paramount importance, providing information about states, production history, and digital provenance. Other keys to map the historical bond between sources and reproductions and to ascertain the authenticity (and thus also its usability) of the reproductions can be extant written documentation (plans, reports and evaluations of digitization projects), paratextual material (about-pages, wikis, FAQ's, etc.) and of course archival access to the information-rich, uncompressed master files from which the presented primary reproduction derives.

A heightened awareness of this on the basis of a dedicated image criticism could also serve as an incentive for digitizing institutions to increase the transparency of the production history of such images and to subject their degree of authenticity and (un)certainty to better scrutiny.

References

- Bentkowska-Kafel Anna, Hugh Denard and Drew Baker (2012). *Paradata and Transparency in Virtual Heritage*. London: Routledge.
- Björk, Lars (2015). *How Reproductive is Reproduction? Digital Transmission of Text-Based Documents*. Borås: University of Borås. <http://hb.diva-portal.org/smash/get/diva2:860844/INSIDE01.pdf>
- British Library (n.d.). *The International Dunhuang Project: The Silk Road Online*. <http://idp.bl.uk/>
- Burrows, Toby (2017). The History and Provenance of Manuscripts in the Collection of Sir Thomas Phillipps: New Approaches to Digital Representation. *Speculum: A Journal of Medieval Studies* 92.1: 39–64. <https://www.journals.uchicago.edu/doi/full/10.1086/693438>
- Codex Sinaiticus (n.d.). <http://www.codexsinaiticus.org/en/>
- Conway, Paul (2013). Preserving Imperfection: Assessing the Incidence of Digitization Error in HathiTrust. *Preservation, Digital Technology & Culture* 42.1: 17–30.

- https://www.si.umich.edu/sites/default/files/conway_shera_award.pdf
- Dahlström, Mats (2010). Critical Editing and Critical Digitization. In: E. Thoutenhoofd, A. van der Weel & W. Th. van Peursen (Eds.), *Text Comparison and Digital Creativity: The Production of Presence and Meaning in Digital Text Scholarship*. Amsterdam: Brill. 79–97.
- Dahlström, Mats, Joacim Hansson and Ulrika Kjellman (2012). ‘As we may digitize’: Institutions and Documents Reconfigured. *Liber Quarterly* 21.3/4: 455–474.
<https://www.liberquarterly.eu/articles/10.18352/lq.8036/>
- Dahlström, Mats and Wout Dillen (2017). The Swedish Literary Bank. *RIDE* 6.doi: 10.18716/ride.a.6.2.
<https://ride.i-d-e.de/issues/issue-6/litteraturbanken-the-swedish-literature-bank/>
- Duranti, Luciana (1995) Reliability and authenticity: the concepts and their implications. *Archivaria*, 39 (Spring 1995), 5–10.
- Hillesund, Terje (2005). Digital text cycles: From medieval manuscripts to modern markup. *Journal of Digital Information* 6.1.
<http://journals.tdl.org/jodi/article/view/62/65>
- Jarlbrink, Johan and Pelle Snickars (2017). Cultural Heritage as Digital Noise: Nineteenth Century Newspapers in the Digital Archive. *Journal of Documentation* 73.6: 1228–1243.
- Kichuk, Diana (2007). Metamorphosis: Remediation in Early English Books Online (EEBO). *Literary and Linguistic Computing* 22.3: 291–303.
- Kjellman, Ulrika (2008) Visual Knowledge Organization: towards an international standard or local institutional practice. In: C. Arsenault and J. T. Tennis (eds.), *Culture and identity in knowledge organization: proceedings of the tenth International ISKO conference*. (Advances in Knowledge Organization, No. 11). Würzburg: Ergon, 289–294.
- Kondrup, Johnny (2011). *Editionsfilologi*. Copenhagen: Museum Tusulanum.
- Mak, Bonnie (2014). Archaeology of a Digitization. *Journal of the Association for Information Science and Technology* 65.8: 1515–26.
<http://doi.wiley.com/10.1002/asi.23061>
- Padfield, James, Kalliopi Kontiza, Antonis Bikakis, and Andreas Vlachidis (2019). Semantic representation and location provenance of cultural heritage information: the National Gallery collection in London. *Heritage*, 6, 648–665.
- Prescott, Andrew and Lorna Hughes (2018). Why Do We Digitize? The Case for Slow Digitization. *Archive Journal*, September 2018.
<http://www.archivejournal.net/essays/why-do-we-digitize-the-case-for-slow-digitization/>
- Stieger, Stefan and Ulf-Dietrich Reips (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior*, 26(6), 1488–1495. doi:10.1016/j.chb.2010.05.013

- Tennis, Joseph (2015). Archival metadata and digital cultural heritage: conceptual provenance, contextual forensics and the authority of the found digital object. *Digital Heritage*, 6, 399–402.
<https://ieeexplore.ieee.org/document/7419533>
- Tomasi, Francesca (2017). La preservazione del contenuto degli oggetti culturali: formalizzare la provenance. *Bibliothecae.it*, 6 (2), 17–40.
<https://bibliothecae.unibo.it/article/view/7531>