Moving Forward with Digital Scientific Images: A Study of Infrastructure, Digitization Work, and Digital Research Practices

Olle Sköld1[0000-0002-0904-7222], Ulrika Kjellman2[0000-0002-3071-697X], Anna Orrghen3[0000-0002-2598-2608], and Jenny Beckman4

[1] Uppsala University, Department of ALM, Sweden
[2] Uppsala University, Department of ALM, Sweden
[3] Uppsala University, Department of Art History, Sweden
[4] Uppsala University, Department of History of Science and Ideas, Sweden

Abstract. Scientific images are important and complex objects of study in the field of digital humanities for two principal reasons. Firstly, scientific images are key components in the making and communication of science in the present day and constitute central source materials in scholarly projects seeking to elucidate the historical practices of research and the development of scientific disciplines. Secondly, the archives, libraries, and museums (ALM) sector invest significant resources into the digitization and mediation of scientific images and it is a crucial success factor for both ALM institutions and future research initiatives that the premises and consequences of such efforts are thoroughly explored. This paper seeks to map which avenues of study and work that are crucial to pursue if available modes of curation, access, search, and analysis in digital collections of scientific images are to be meaningfully improved. The paper is based on a literature review and an overview of the current state of digitization work, digital collections, and digital infrastructures for storage and mediation at Uppsala University Libraries. Methodologically the paper makes use of action research and an adaptable, pragmatic, and 'exploratory' approach to academic research. The study identifies five themes of study and work that, if competently pursued, promise to push the boundaries of what is known about scientific images forward in many areas of the digitization spectrum both in terms of best practices and theoretical understandings. The themes are: (1) method and infrastructure focus; (2) method focus; (3) digitization work focus; (4) epistemic and research-practice focus; (5) epistemic, methodological, and historiographical focus.

Keywords: Scientific Images, Digitization, Image Extraction and Indexing, Digital Research Practices, Digital Image Infrastructures

1 Introduction

Bohr's model for visualizing atoms; Linnaeus' sketches of plants; Rosling's computer-facilitated infographics—these are but three examples of how massively impactful images are in the production, organization, and dissemination of scientific knowledge. The roles that scientific images play in the venture of science are, both presently and historically, complex and continuously changing. Throughout the history of science, the purposes and compositions of scientific images have been tied to shifting research practices and available technological means of image creation and reproduction. Their intimate relationship with the workings of science have made scientific images an interdisciplinary focal point of a wide array of research interests (Rudwick, 1976; Knorr-Cetina, 1999; Latour, 1999; Lynch & Woolgar, 1990; Daston & Galison, 2007). The possibilities and potentials of research centered on scientific images stemming from pre-digital workflows have been enhanced by the advent of large-scale digitization initiatives and the development of easily-navigable platforms of access. In the US and in the EU, libraries, museums, and archives have taken a great interest in digitizing their collections. Examples abound: The Library of Congress and The British Library make large quantities of images available via image-hosting service Flickr (www.flickr.com); the EU-financed Europeana projects invites the user to "[e]xplore the natural world in 3,415,352 drawings, specimens, images and documents from European collections" (Europeana, n.d.). Scientific images are hence important and complex objects of study in the field of digital humanities for two principal and interrelated reasons. Firstly, scientific images are key components in the making and communication of science in the present day and constitute central source materials in scholarly projects seeking to elucidate the historical practices of research and the development of scientific disciplines. Secondly, the archives, libraries, and museums (ALM) sector invest significant resources into the digitization and mediation of scientific images and it is a crucial success factor for both ALM institutions and future research initiatives that the premises and consequences of such efforts are thoroughly explored.

The aim of this paper is to map which avenues of study and work that are crucial to pursue if available modes of curation, access, search, and analysis in digital collections of scientific images are to be meaningfully improved. The paper is based on a literature review and an overview of the current state of digitization work, digital collections, and digital infrastructures for storage and mediation at Uppsala University Libraries (UUL). Uppsala University Libraries is a large university library organization with significant collections that have been built up through donations, spoils of war, Swedish legal deposits, and purchases. The special collections—which include e.g., old prints, manuscripts, and images—hold scientific images that are both valuable and rare, and of great value to many scholars from different disciplines and research interests.

Insights into the premises and possible directions of development regarding UUL's modes of work and infrastructures related to digital scientific images were attained in informal collaboration with the library's strategic development manager, digitization coordinator, and a system developer. Methodologically the paper makes use of action research (Checkland & Holwell, 2007) and the adaptable, pragmatic, and 'exploratory' approach to academic research put forward by Stebbins (2001). The principal impetus of action research is thus incorporated in the foundational premise of the present paper: to by the way of reflection, critical study, and rigorous research procedures benefit the processes of cultural-heritage digitization and data-driven research practices in the field of digital humanities.

2 Outline of the Problem Area

The importance of scientific images as knowledge-producing artifacts have been stressed by scholars from numerous disciplines, including Information Studies, Science and Technology studies (STS), and History of Art (Latour, 1990; Lynch & Woolgar, 1990; Kemp & Wallace, 2001; Daston & Galison, 2017). Research on scientific images are often hindered by matters of access and visibility: scientific images tend to be 'hidden' in publications and are often not efficiently findable through the traditional knowledge organization systems (KOS) of the ALM sector. Common explanations to this state of affairs include a lack of metadata (see e.g., Enser, 2008; Christensen, 2017) and indexing practices, search tools, and metadata systems that do not match the requirements of images and image research. Traditional KOS have often been developed with text documents in mind, and research-based insights into the special requirements that images put on the indexing process are limited (Kjellman, 2006). While nations, foundations, workgroups, researchers, and other actants in the ALM and wider cultural-heritage sectors devote significant time and resources to the technologies of digitization and infrastructures of storage, there is a considerable need for improving access, indexing, retrieval, and in-depth research seeking to explore the epistemic and methodological limitations and opportunities offered by image digitization (Enser, 2008; Christensen, 2017). It is also crucial to attain a better understanding of the effects of the digitizing process on the procedures of scholarly knowledge production. While the technical processes of digitization and metadata markup are well understood, and there are frameworks able to explain the general nature and effects of the work of knowledge organization and production which digitization is an instance of (e.g., Björk, 2015), it is important to better come to know how these elements interact. Otherwise, digitization initiatives and efforts of method development are at risk of becoming increasingly centered on and driven solely by technological considerations.

Significant and meaningful development of the methods and infrastructures of digitization and digitized corpuses of scientific images is hence dependent on the availability and cognizant application of research-based insight into how digitization work is affected and influenced by practical, administrative, technical and theoretical considerations in the digitization workplace, and of the epistemic influences of this work. The relevance of this line of argumentation is additionally heightened in the context of scientific images. Scientific images are a valuable yet under-studied and under-utilized material and the search tools and metadata systems providing access to digitized materials are, as previously pointed out, seldom adapted to the special requirements of images and image research. It is hence important to create corpuses, methods of digitization, and software tools focused on scientific images so as to encourage and facilitate high-quality research on digitized cultural heritage materials in the digital-humanities field.

3 Literature Review

This paper connects to two main veins of previous research: images as actants in science and image digitization, and methods for image indexing and retrieval. Although the focus on scientific images as actants in science is comparatively recent, several studies have demonstrated the wealth of different types of expressions and representational devices—among them images—that are used in scientific communication and documentation and thus must be seen as knowledge-carriers (Frohmann, 2004; Lund, 2010; Lynch and Woolgar, 1990; Latour, 1990). There is also previous research of how images play a part in scientific practice in a variety of fields including, but not limited to, anatomy and medicine (e.g., Kemp and Wallace, 2000), geology (Rudwick, 1985), botany and zoology (Törnvall, 2013, 2017; Dal, 1996; Secord, 2007; Blunt and Stearn, 1994). Other studies have inquired into how the technologies of printing and photography (Ivins, 1969; Wilder, 2009) and digital imaging technologies (Bredekamp et al., 2015; Coopmans et al., 2014; Dussauge, 2008) function as vehicles for ideals of scientific objectivity.

Research on image digitization have delved into a range of topics like copyright issues (Harper, 2007), economical (Williams, 2003) and institutional (Dahlström et al., 2009) conditions, and digital image collections as pedagogical resources (Marmor, 2002). Other veins of research that links to this paper have explored the consequences of digitization for users, for instance how digital images and digital collections affect the work of museum professionals (Koo, 2006) and librarians (Gushee et al., 2005). The processes of selection that direct the creation of digitized archives and the limitations of digital sources (Ogilvie, 2016), the difference between 'mass digitization' and 'critical digitization' (Dahlström et al., 2012), and the consequences that standardized metadata carry for digitized cultural heritage material have also been inquired into (Kjellman, 2009).

Several studies of methods for image indexing and retrieval show that image collections commonly suffer from bad indexing (Enser, 1995; Kjellman, 2006) for two main reasons: the index and retrieval systems have been developed with text documents in mind and do not take into account the special requirements that images put on the indexing process, and a naïve trust has been placed in technical solutions. User studies have also pointed to the limited value that commonly chosen methods of image indexing has to large users groups (e.g., Jörgensen, 1998). Attempts to provide conceptual models for manual indexing have been criticized of not providing enough richness, complexity, and consistency (Christensen, 2017) and for not taking user expectations into account (Jörgensen, 2003).

4 Results and Discussion

The study identifies five avenues of study and work that, if competently pursued, promise to push the boundaries of what is known about scientific images forward in many areas of the digitization spectrum (selection, digitization, indexing, retrieval, use)—

both in terms of best practices and theoretical understandings. In order to best operationalize the findings of this paper, they are grouped below in five separate themes ('T1-5'; see Table 1 for an overview) strategically positioned in the process of research-based method development and digitization in the domain of scientific images.

Themes 3-5 interrogate the epistemic dimensions of digitization. T3 focuses on exploring digitization as an instance of knowledge work while T4 and T5 delve into the consequences that digitization and different means of organizing and describing digitized scientific images carry for the posing of research questions, the application of methods of study, and the drawing of conclusions in humanistic and social sciences. Themes 1-2 are oriented towards research-based evaluations of image-indexing methods and develop a digitization software infrastructure respectively. By design, the themes are interlinked and build on each other. The studies of digitization and image-indexing methods (themes 2, 3-5) are operationalized in the development of digitization software (T1). The discussion of each theme below will touch upon relevant empirics, methods, theoretical dimensions, and plausible outcomes.

Table 1. Overview of the themes for future work and research suggested in this paper.

Theme	Description	Output
1. Method and Infrastructure Focus	Development of specialized software infrastructures	Increased capacity for image-extraction, digitization, metadata enrichment, and image searches
2. Method Focus	Evaluation of OCR, CBIR, crowd sourcing, and their interoperabilities	Identification of the benefits the chosen image-indexing methods to various research communities and user groups
3. Digitization Work Focus	Ethnographic investigations of digitization as situated practice	Insights into how image-digitization work is enacted in relation to and effected by workplace practices, hardware, software, local policies, international standards
4. Epistemic and Research-Practice Focus	Investigations of the connection between digitized corpuses and research practices	Better understandings of the epistemological opportunities, challenges, and limitations offered by large-scale collections of digital images
5. Epistemic, Methodological and Historiographical Focus	Studies of the relationship between analogue images and their digital reproductions	The extent to which traditional art historical methods for image analysis are applicable in the study of digital imagery

4.1 Theme 1: Method and Infrastructure Focus

The object of T1 is to develop, test, and implement an image-extraction and image-indexing software infrastructure designed to provide UUL with capabilities for mass image-extraction and digitization, metadata enrichment, and tools for advanced image search of digitized scientific images. The software infrastructure should be designed enable efficient and large-scale digitization of images at UUL by providing the following functionalities: the efficient extraction of images from books and already digitized materials; the capability to make connections between digitized images and existing metadata; multiple methods of image-indexing including crowd-sourcing of users external to UUL (see T5); a graphical user interface and search tools to enable researchers to find, analyze, and compare digitized images; and integration with the existing Alvin—a cross-ALM platform focused on digitized cultural heritage collections, maintained and developed by a consortium consisting of UUL, Lund University Libraries, and University of Gothenburg Libraries (Alvin, n.d.)—infrastructure.

4.2 Theme 2: Method Focus

T2 seeks to investigate and discuss in what way different image-extraction and image-indexing approaches can be used and synthesized to improve the access to digitized scientific images. It addresses the problems of accessing and retrieving images from digitized image collections, which are well-known and described in several studies (Jörgensen, 2003; Enser 2008; Christensen, 2017). During the last decades, different solutions have been presented on the market—both automatic and manual methods. None has proven to be a panacea; some might be appropriate in one instances or context but not in another. In T2 the following retrieval methods should be investigated and evaluated: Optical Character Recognition (OCR), Content Based Image Retrieval (CBIR), metadata, and crowd sourcing. Apart from identifying the benefits the chosen indexing methods to various research communities and user groups, T2 will also explore how multiple methods can be integrated in the same platform. Since previous research on image retrieval has been very specialized and focused on one specific indexing-method, the combinatory approach of T2 offers new opportunities to both compare different retrieval methods and to identify possible synergies. By shedding light on the interoperability of image-indexing methods, T2 also strives to encourage increased communication between the research community of metadata/manual indexing and the community of CBIR/automatic indexing research which rarely interact (Enser, 2008).

4.3 Theme 3: Digitization Work Focus

T3 sets out to explore how situated digitization work in the ALM and cultural heritage sectors— enacted in relation to workplace practices, the affordances operation of machinery, and local-global orders of work (local policies and work processes–international standards)—affects digitization outcomes, and to provide insights into the epistemic implications of digitization for scholarly knowledge production. T3 approaches the 'un-black boxing' of digitization of cultural heritage artifacts in the ALM sector by the way of an ethnographic study into digitization as a mode of situated knowledge work; digitization is here approached to be not solely a matter of media transfer (see e.g., Bolter and Grusin, 1999), but a refashioning bound to processes of knowledge production and organization. T3 will be based on fieldwork, participant observation, document studies, and interviews geared towards examining digitization work along the empirical trajectories of digitization practices, documents and standards, machine use, and organizational efforts (project planning, the workplace context of the digitization work). The theoretical framing will draw on practice-theoretical writings on knowledge (Gherardi, 2012, Nicolini, 2012, Orlikowski, 2002), sociological studies of science (Knorr-Cetina, 1999, Latour and Woolgar, 1979) and workplaces (Blackler, 1995, Harper, 1998, Luff et al., 2000, Orr, 1996).

4.4 Theme 4: Epistemic and Research-Practice Focus

T4 seeks to identify what kind of new research questions arise when a larger corpus of digitized images is presented to the researcher, and to elucidate if the number of images available bring new insights on the nature and function of scientific images. Studies in the history of scientific images have so far focused on a specific era, discipline, or technique. A larger set of image data may offer a possibility to identify more general patterns and, e.g., follow how motives travel, how technique develop and change, and how illustration practices and rhetoric vary between different disciplines. T4 will also investigate how different knowledge organizational tools and indexing practices serve different research agendas and interests. Previous studies (Hjørland, 2002, Ørom 2003, Kjellman, 2008) have put forward the necessity of paying attention to domain specific interests and demands when developing knowledge organizational tools. Accordingly, T4 will inquire into how different image retrieval methods (OCR, CBIR, metadata) meet the demands of different scholarly communities and disciplines. In this sense, T4 connects to T2, which aims at investigating different image retrieval methods in relation to the digitized material.

4.5 Theme 5: Epistemic, Methodological and Historiographical Focus

T5 is based on the critical examination of the relationship between source materials and their digital reproduction with a particular focus on methods for image analysis and what kind of knowledge the digitized reproduction generates. Scientific images are a recurrent source material in art historical studies (e.g., Bredekamp et al., 2015; Voss, 2010). Given the close relationship between the histories of art, science and technology, T5 will thus pay particular attention to the implications of digitization for art history: what is the relation between the source material and its digital reproduction?; to what extent are traditional art historical methods for image analysis applicable in relation to the digitized material?; and, finally, what kind of art history is made possible by using the digitized material? Earlier research on the use of reproductions in art history has paid attention to the lack of a critical approach in relation to digital reproductions (Christensen, 2010). By describing and analyzing the digitized corpus in relation to earlier art historical research on scientific images that are not digitized, research in T5 takes a historiographical approach to the examination of the epistemological and methodological limitations and opportunities of digitization.

5 Conclusions

Digitized corpuses of cultural heritage artifacts play increasingly important roles in scholarly inquiry, both as source materials and as the focal points of method developments. The work by which such digital corpuses come into being is however poorly understood, thus making it difficult to competently grasp the conditions of present-day humanistic and social-science research. This paper outlines how to push the boundaries of what is known about the work that underpins collections of digitized cultural heritage artifacts in the ALM-sector, and investigate its relation to standards, technology, local orders of work, and workplace processes. The paper also suggests that the current debates about access, indexing, metadata, and searchability should be engaged in on a theoretical as well as a practical level. Examples of fruitful explorations include how these issues affect the digitization process as well as the management of digitized material and the construction of collections, and how the results affect scholarship on scientific communication.

The questions that this paper turns towards to are of particular significance for three target groups: archives, libraries, museums and other institutions and actors administrating collections containing images, researchers within the numerous disciplines present in the digital humanities field, and, finally software engineers developing digital methods for indexing and retrieval.

References

1. Björk L (2015) How Reproductive is a Reproduction? Digital Transmission of Textbased Documents. University of Borås, Borås
2. Blackler F (1995) Knowledge, Knowledge Work and Organizations: An Overview and Interpretation. Organization Studies, 16(6):1021–1046
3. Blunt W, Stearn WT (1994) The Art of Botanical Illustration. Antique Collectors' Club, Woodbridge
4. Bolter JD, Grusin R (1999) Remediation: Understanding New Media. MIT Press, Cambridge, MA
5. Bredekamp H, Dünkel V, Schneider B, editors (2015) The Technical Image: A History of Styles in Scientific Imagery. The University of Chicago Press, Chicago, IL
6. Checkland P, Holwell S (2007) Action Research. In: Kock N (ed) Information Systems Action Research. Integrated Series in Information Systems, volume 13. Springer, Boston, MA, pp 3–17
7. Christensen HD (2010) The Repressive Logic of a Profession? On the Use of Reproductions in Art History. Journal of Art History, 79(4): 200–215
8. Christensen HD (2017) Rethinking Image Indexing? Journal of the Association for Information Science and Technology, 68(7): 1782–1785
9. Coopmans C, Vertesi J, Lynch M, Woolgar S (2014) Representation in Scientific Practice Revisited. MIT Press, Cambridge, MA
10. Dahlström M, Hansson J, Kjellman U (2009) Documents Reconstructed: Digitization and Institutional Practice as Mediation. In: Proceedings from DOCAM '09, March 28–29, 2009, Madison, Wisconsin
11. Dahlström M, Hansson J, Kjellman U (2012) As We May Digitize' — Institutions and Documents Reconfigured. Liber Quarterly: The Journal of European Research Libraries, 21(3/4)
12. Dal B (1996) Sveriges zoologiska litteratur: En berättande översikt om svenska zoologer och deras tryckta verk 1483–1920. Orbis pictus, Fjälkinge
13. Daston L, Galison, P (2007) Objectivity. Zone Books, New York, NY
14. Dussauge I (2008) Technomedical Visions: Magnetic Resonance Imaging in 1980s Sweden. Kungliga Tekniska högskolan, Stockholm
15. Enser P (1995) Progress in Documentation Pictorial Information Retrieval. Journal of Documentation, 51(2):126–170
16. Enser P (2008) The Evolution of Visual Information Retrieval. Journal of Information Science, 34(4): 531–546
17. Europeana (n.d.) Europeana Natural History. https://www.europeana.eu/portal/fr/collections/natural-history
18. Frohmann B (2004) Deflating Information: From Science to Documentation. University of Toronto Press, Toronto
19. Gherardi S (2012) How to Conduct a Practice-Based Study: Problems and Methods. Edward Elgar, Cheltenham

20. Gushee E, Parker J, Whiteside AB (2005) From Analog to Digital at the University of Virginia – One Institution's Journey. VRA Bulletin, 31(3): 35–39
21. Harper, RHR (1998) Inside the IMF: An Ethnography of Documents, Technology, and Organisational Action. Academic Press, San Diego, CA
22. Hjørland B (2002) Domain Analysis in Information Science: Eleven Approaches – Traditional as well as Innovative. Journal of Documentation, 58(4): 422–462
23. Ivins WM (1969) Prints and Visual Communication. MIT Press, Cambridge, MA
24. Jörgensen C (2003) Image Retrieval: Theory and Research. Scarecrow Press, Lanham, MD
25. Jörgensen C (1998) Attributes of Images in Describing Tasks. Information Processing & Management, 34(2):161–174
26. Kemp M, Wallace M (2000) Spectacular Bodies: The Art and Science of the Human Body. University of California Press, Berkeley, MA
27. Kjellman U (2006) Från kungaporträtt till läsketikett: en domänanalytisk studie över Kungl.bibliotekets bildsamling med särskild inriktning mot katalogiserings- och indexeringsfrågor. Department of ALM, Uppsala University, Uppsala
28. Kjellman U (2008) Visual Knowledge Organization: Towards an International Standard or Local Institutional Practice. In: Proceedings of the International Society for Knowledge Organization 12: Advances in knowledge organization, pp 289–294
29. Kjellman U (2009) Digitalisering som standardisering: Verktyg i konciperingen av kulturarv. In: Lund ND (ed) Digital formidling af kulturarv: Fra samling til sampling. Multivers, Köpenhamn, pp 219–237
30. Knorr-Cetina K (1999) Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press, Cambridge, MA
31. Koo B (2006) The Use of Digital Images by Art Museum Professionals: Preferences, Perceptions, and Implications for Museum Practice. The Florida State University, Tallahassee, FL
32. Latour B (1990) Drawing Things Together. In: Lynch M, Woolgar S (eds) Representation in Scientific Practice. MIT Press, Cambridge, MA, pp 19–68
33. Latour B (1999) Circulating Reference: Sampling the Soil in the Amazon Forest. In: Pandora's Hope: Essays on the Reality of Science Studies. Harvard University Press, Cambridge, MA, pp 25–79
34. Luff P, Hindmarsh J, Heath C, editors (2000) Workplace Studies: Recovering Work Practice and Informing System Design. Cambridge University Press, Cambridge, UK
35. Lynch M, Woolgar S (1990) Sociological Orientations to Representational Practice in Science. In: Lynch M, Woolgar S (eds) Representation in Scientific Practice. MIT Press, Cambridge, MA, pp 1–18
36. Lund NW (2010) Document, Text and Medium: Concepts, Theories and Disciplines. Journal of Documentation, 66 (5): 734–749
37. Marmor M (2002) ArtSTOR: A Digital Library for the History of Art. Art Libraries Journal, 27(3): 26–29
38. Nicolini D (2012) Practice Theory, Work, and Organization: An Introduction. Oxford University Press, Oxford
39. Ogilvie B (2016) Scientific Archives in the Age of Digitization. Isis, 107(1): 77–85
40. Orr JE (1996) Talking About Machines: An Ethnography of a Modern Job. ILR Press, Ithaca, NY
41. Orlikowski WJ (2002) Knowing in Practice: Enacting a Collective Capability in Distributed Organizing. Organization Science, 13(3): 249–273
42. Rudwick MJS (1976) The Emergence of a Visual Language for Geological Science 1760–1840. History of Science, 14(3):149–195

43. Secord A (2002) Botany on a Plate. Isis, 93:28–57
44. Stebbins RA (2001) Exploratory Research in the Social Sciences. Sage, Thousand Oaks, CA
45. Törnvall G (2013) Botaniska bilder till allmänheten: om utgivningen av Carl Lindmans Bilder ur Nordens flora. Atlantis, Stockholm
46. Wilder KE (2009) Photography and Science. Reaktion books, London
47. Williams SJ (2003) Technology: Per Unit (Image) Cost and Labor for Digitizing an Image. VRA Bulletin, 30(1): 44
48. Voss J (2010) Darwin's Pictures: Views of Evolutionary Theory, 1837–1874. Yale University Press, Yale, CT
49. Ørom A (2003) Knowledge Organization in the Domain of Art Studies – History, Transition and Conceptual Changes. Knowledge Organization, 30 (3/4): 128–143